Beat gestures and speech processing: When prosody extends to the speaker's hands.

Emmanuel Biau

TESI DOCTORAL UPF 2015

DIRECTOR DE LA TESI

Dr. Salvador Soto-Faraco

Departament de Tecnologies de la Informació i les Comunicacions

Universitat Pompeu Fabra Barcelona

Acknowledgement

The present dissertation represents the completion of a long time endeavor, during which I have been unconditionally supported by my two families. My foremost thanks go to my parents and my sister in France. Here in Barcelona, my thanks go to my other family with my love Francesca, Federico, Fabrizio, Marcello, and Filippo who tried hard to ruin my PhD with all the last rounds, because “we all work tomorrow”, Camilla and Michele. I owe it all to them.

Then, thanks go to my supervisor, Salvador Soto-Faraco. Over these years he has taught me how to deal with scientific practice, critical thinking, support, care, and most of all, bore my moods.

Special thanks go to Henning Holle who supervised my stay in UK. His collaboration has been crucial for the fMRI work presented in this dissertation and I learnt a lot from him. Thanks to Ruth and Lluis as well, for their patience and kindness. It was a real pleasure to collaborate with them. I hope to see all of them soon on my career path again.

I am grateful to all my lab mates at CBC, those of the past and of the present. Many thanks go to Manuela, Martí, Luis, Mireia, Nara, Daria, Joan and the others from the MRG. I also want to thank Ruggero, Andrea, Nicolò, Alice, Marco and all the others for having shared with me these years in the bad and in the good times, in and outside from the university.

I am also grateful to Nuria Sebastián Gallés, Luca Bonatti, Albert Costa and Gustavo Deco that in these years have contributed to build a rich scientific environment from which I fully profited.

Then, I would like to thank Cristina, Xavi and Sylvia for their help on so many technical and bureaucratic issues but most of all, their daily good mood (literally, I could not make it without their help).

Finally, many thanks to the ones I forget here.

Abstract

Speakers naturally accompany their speech with hand gestures. In particular, they spontaneously extend the auditory prosody to visual modality through rapid and biphasic beat gestures, helping them to structure their narrative and emphasize relevant information. The present thesis aimed to increment the relatively less documented beat gestures and their neural correlates on the listener's side. We developed a naturalistic approach combining political discourse presentations with neuroimaging techniques (ERPs, EEG and fMRI) to investigate beats correlates in both temporal and spatial dimensions. We also set experimental procedures to determine behavioral measures indexing the influence of beat gestures on audiovisual speech processing. The main findings of the thesis first revealed that beat-speech processing engaged language-related areas, suggesting that gestures and auditory speech are part of the same language system. Second, the time course analyses revealed that the presence of beats modulated the auditory processing of affiliated words around their onsets and later at phonological stages. We concluded that listeners perceive beats as visual prosody and rely on their predictive value to anticipate relevant acoustic cues of their corresponding words, engaging local attentional processes. The present dissertation confirmed that, even if simple, spontaneous beats presented in continuous audiovisual speeches are a good alternative to investigate the neural correlates of gesture-speech processing.

Resumen

Los gestos de las manos acompañan de manera natural el discurso de los hablantes. El objetivo principal de esta tesis fue la investigación de la percepción de los gestos rítmicos y la actividad neuronal relacionada con estos, un área todavía relativamente inexplorada. Esta tesis se desarrolló con un enfoque naturalístico combinando la presentación de discursos políticos con técnicas de neuroimagen (ERPs, EEG y fMRI) para investigar la influencia de estos gestos, desde un punto de visto espacial y temporal, en la actividad neuronal. Se llevaron a cabo experimentos comportamental para medir la influencia de los gestos rítmicos en el procesamiento del lenguaje. Sus principales hallazgos fueron, primero, que el procesado conjunto del habla y gestos rítmicos involucraron áreas relacionadas con el lenguaje, esto sugiere que los gestos y el habla forman parte de un único sistema del lenguaje. Segundo, que los gestos rítmicos modulan el procesamiento de las palabras a las que acompañan tanto en el momento de su pronunciación como en etapas posteriores. Concluimos por tanto que los oyentes perciben los gestos rítmicos como parte de la prosodia visual y utilizan su valor predictivo para anticipar la señal acústica de la palabra a la que preceden a través de procesos locales de atención. Esta tesis también confirma que el estudio de la actividad neuronal relacionada con el procesamiento del lenguaje acompañado de gestos es posible utilizando gestos rítmicos espontáneos incluidos en un discurso audiovisual, incluso a pesar de la simpleza de los gestos rítmicos.

Preface:

Life in society implies that people interact with each other to work, ask for information, comment topics of common interest, or simply share feelings. Along with technology progresses, the format of human interactions evolved as well, and made possible to communicate without any visual contact from one part of the world to another by phone or simply by email. In the daily life however, conversations between two or more persons remained the most frequent way to communicate and obtain a solution to a problem. During these direct interactions, the protagonists have access to a large amount of congruent information conveyed through different parallel modalities. Obviously, speech is the more predominant channel as it allows the speaker to consciously express his thoughts and make them clear for any listener speaking the same language. As a perfect communicative tool, the speaker can manipulate the verbal utterance content to decide to which extend of honesty he wants to inform the listener, by partially hiding his thought, or else, misleading him with liars.

But face-to-face conversations are multisensory experiences and listeners have access to additional visual information from the speaker. As two normal persons look to each other while speaking, the listeners can also observe facial information. Non-verbal information can affect directly speech processing by conveying redundant information. For example, when two people try to have a conversation in a noisy bar, looking at the speaker's mouth helps to puzzle out the degraded speech with lips' movements and improve

comprehension. But in other cases, visual information can impact other aspects which are not expressly conveyed by speech. For instance, the eyebrows' movements allow inferring speaker's emotional states, as wrinkling them generally means anger or frustration. The shape of the mouth also gives some clues on the speaker's mood (smiling comes with a good mood or irony). Finally, head movements can bring complementary information as well. For example, speakers make rapid and short beats with the head to accompany the word "yes" and demonstrate that they agree with the interlocutor.

In addition to facial mimics or head movements, listeners have access to another type of prominent visual linguistic information with the speaker's hand gestures. Speakers often accompany their discourse with spontaneous hand gestures, even if they do not have always an explicit purpose to facilitate speech comprehension for the listener. These gestures can be categorized based on their shape, their semantic content or their relationship respect to speech (for example if they convey redundant or additional information which is not described in the verbal utterance), but are all part of a continuum of hand movements. As part of the visual linguistic channel, gestures may convey information on speaker's emotional state and intention. From a postulate stating that hand gestures may affect how listeners perceive the speaker's discourse, one can assume that different categories of gestures may impact this processing at different levels.

In the present thesis, we investigated the impact of one of the most frequent category of gestures during speech perception, at behavioral and neural levels. We established two main objectives: First, we developed a naturalistic approach to study gestures when they are spontaneously produced during a continuous speech. We designed new experimental procedures of audiovisual speech presentation using entire or segments of public addressees in which the speaker naturally accompanied his speech with gestures. Second, we investigated the neural correlates of gestures and their effects on speech processing in both temporal (using electroencephalogram recording set up) and spatial (using functional magnetic resonance imaging) dimensions.

In a first Introduction section, I will report relevant literature and relate it to the purpose of the thesis, to give the reader the necessary background to contextualize and understand our motivations. The second experimental section will describe three different studies addressing the impact of gestures at neural levels of speech processing. Then in a third section, I will discuss our findings and their potential impact in the field of research of cospeech gestures. Finally, I will conclude with general comments on possible further investigations.

TABLE OF CONTENT

1. INTRODUCTION

1.1 General overview: Multisensory experiences in life and communication

Humans experience multisensory situations in their environment, whereby sensory stimuli about events are captured through different sensory modalities but integrated as unitary percepts. When we walk through the door to go to work, the sound and sight of our neighbor's dog barking is perceived as a unitary whole. In fact, almost all events in our lives can be described as multisensory perceptions (Stein & Meredith, 1993; Driver & Spence, 2000; Spence & Driver, 2004; Calvert, Spence, & Stein, 2004; Calvert & Thesen, 2004).

Communication illustrates a paramount example of multisensory perception. Due to life in society, people constantly interact with each other and experience multisensory integration of audiovisual (AV) speech signals during conversations. In natural face-to-face conversations, conversation partners see each other and when listening, have access to visual information accompanying the speaker's verbal utterance. At first glance, the auditory modality appears to be the most prominent channel to convey the spoken message in normal hearing conditions, and the accessory visual information provided by the speaker may seem secondary. An illustration of this point of view is a phone conversation, in which two persons can perfectly communicate without seeing each other.

More recently, conversations via Internet, although ridden with audio-visual desynchronization and other kinds of interference, have rapidly become a common tool to communicate, suggesting that even if speech would be enough, seeing the speaker remained essential to make video conferences popular.

But appearances might be misleading as many other situations give weight to visual information. For example, when it becomes difficult to follow a conversation in noisy conditions such as crowded bars, listeners have to resort to additional cues to compensate acoustic degradation. Usually, listeners tend to focus on the speaker's face and particularly on his mouth trying to puzzle lip movements out and retrieve sounds. Soon in the 50's, Sumby and Pollack (1954) already demonstrated that the loss in correct word identification when the verbal utterance was presented alone at difficult signal-to-noise ratio, was compensated for when participants could see the lip movements of the speaker. From that finding, it was hypothesized that this benefit was possible because articulatory movements of the speaker (lips aperture) are closely related to the speech envelop modulations, facilitating phoneme perception (Vatikiotis-Bateson & Yehia, 1996; Yehia, Rubin, & Vatikiotis-Bateson, 1998; Grant & Seitz, 2000; Chandrasekaran et al., 2009). Another evidence of the importance of visual speech came from a multisensory illusion in AV speech perception. In 1976, McGurk and MacDonald accidentally discovered that when participants listened to the spoken syllable /ba/ presented simultaneously with the video of lip movements corresponding to

the syllable /ga/, they perceived the syllable /da/ (see Massaro & Stork, 1998). As subjects were not previously aware of such an illusion, results suggested a stronger than suspected influence of visual information on auditory speech perception. From a motor view of speech perception, one may argue that as speech sounds come from the same articulatory apparatus as the lip movements, it is fair to think that both reciprocally influence each other as perceptual cues to retrieve the speech message. Further, in noisy conditions, it has been shown that the sight of the speaker's head movements can improve intelligibility of speech when head beats are congruent with pitch accent (Munhall et al., 2004). More surprising, this benefit was found even when only the upper part of the face was visible (Davis & Kim, 2006). Eyebrow movements correlated with prosodic cues of speech were found to influence speech perception as well. Krahmer and Swers (2007) showed that in short sentences, the prominence of words accompanied by congruent eyebrows movements was increased (eyebrows moving up with the accentuated syllable of the affiliate word).

In real life however, listeners generally have access to all visual cues at once, including the whole upper part of the speaker's body and his hand gestures as well. The omnipresence of hands in conversations has been largely exploited in cartoons where nervous characters are often depicted executing large hand/arms movements, or in movies representing the stereotypic Italian people with frequent hand movements (or for French people, the famous actor Louis de Funès). A good and straightforward definition of hand

gestures has been given by David McNeill (1992): "*The gestures […] are the movements of the hands and arms that we see when people talk*". As hands' shapes and trajectories can describe actions, objects or feelings, one can assume that they impact both speech production on the speaker's side, and perception on the listener's side. A recent new strand of research focusing on the role of gestures in audiovisual speech has emerged in the last twenty years in parallel with neurophysiology and neuroimaging techniques. Thus, the role of gestures in speech production has been now relatively well established and different models have been proposed to describe the interactions between gestures and verbalization modalities (I will present briefly these models in the next part). Although there are a multitude of gesture types, meaningful gestures (those whom the hand shape describes a clear object or action) were the most studied, maybe because their impact on production seemed more obvious or for methodological issues. Consequently, the role of less elaborated gestures (like simple flicks of the hand or pointing) is still uncertain, even if they are the most frequent in narrative and public addressees. The purpose of the present thesis was actually to focus on less elaborated gestures (called "beats") to propose an alternative manner to investigate the neural correlates of spontaneous gestures during continuous audiovisual speech perception. The starting point of the present thesis was to find a manner to conserve the natural frequency of spontaneous gestures during continuous speech production and, at the same time, control for the gesture type (as many speech contexts make use of different types of gestures). Then, we had to think in a

new speech format satisfying both exigencies. It appeared to us that continuous public speeches (e.g. political discourses) in which the speakers produce almost all the time the same type of simple gestures (e.g. beats) provided a new way to investigate the neural correlates of gesture-speech processing in more naturalistic conditions of perception.

In the Introduction, I will first introduce the fact that gestures come with speech in various kinds of speech situations and that complex reciprocity suggests a common origin from the early lifetime. Then, I will describe the general structure of a gesture and the different categories depending on their relationship with utterance, commenting with different models that attempted to localize the role of gestures during speech production. From that, I will jump on the listener's side to report behavioral and neuroimaging evidences of gestures' impact on speech perception as well. Through that, I will raise some methodological issues and introduce why there is a need to find new alternative to study gesture-speech integration. Finally, I will develop on the type of gestures and speech contexts we chose to conduct the thesis, and the hypothesis that we rose to investigate neural correlates of gestures.

1.2 Gestures during speech production: Some cases and a common origin from early lifetime

Everyone gestures when speaking. This has been found independently from ages or cultures (Feyereisen & de Lannoy,

1991). Although being conveyed in two distinct modalities, gestures (visual) and utterance (audio) appear to be part of a single language system (McNeill, 1992; Goldin-meadow et al., 1993). Going back to the example of a phone conversation, the speaker often produces speech-related gestures despite the listener can obviously not see him/her. Even more striking, Iverson and Goldin-Meadow (1998) showed that congenitally blind people gesture when they speak just like sighted speakers do. Interestingly, blind speakers produced gestures at the same frequency regardless of whether the listeners were sighted or blind. The authors suggested then that gestures required neither a model nor an observant listener (Bavelas, Chovil, Lawrie & Wade, 1992; Iverson & Goldin-Meadow, 1998, 2001). In contrast, when speakers are prevented from gesturing when they speak, it seems to make speech production much more difficult (Cook, Yip & Goldin-Meadow, 2012; Ping & Goldin-Meadow, 2010; Goldin-Meadow et al., 2001). Thus, having such a generalized bimodal communication allows also humans to operate short-term multimodal shifts when environmental conditions change, in order to always maintain an optimal transmission and perception of the message (Partan, 2013). In noisy urbanized zones for instance, construction workers are used to switch to hand gestures to communicate with each other when a colleague is drilling with the jackhammer. This involves knowledge of communicative intentions in both gestures and speech modalities.

The infinite situations of speech production in which gestures accompany verbalization suppose a large variety of gestures and overall, an implicit knowledge on matching gestures with speech

content/context. This implies reciprocity in the relation between gesture and utterance that originates early in lifetime.

In fact, it is thought that spoken language probably developed from manual language. This idea arose in part from the observations of babies at the first stages of communication during the first months after birth. Indeed, babies generally begin to gesture before they pronounce their first word. At eight months, they use pointing gestures (deictic gestures) to refer to objects in their environment although they cannot verbalize their intention yet (Carpenter et al., 1983; Iverson & Goldin-Meadow, 2005). When children pronounce their first isolated words (i.e. “one-word period”), they begin to combine them with a gesture (for example they point at a spoon, saying “spoon”), before they start to combine one word with another. At first, children produce gestures with meaningful utterance or not, and often in an asynchronous manner. But step by step, they begin to produce congruent and synchronous gestures with meaningful word, suggesting a convergence period in which they acquire additional motor skills (hand and mouth) that allow them to combine speech and gesture in a single communicative act. Other studies suggested that the coordination between gesture and speech occurs even before the one-word period. In a recent study, Esteve-Gilbert and Prieto (2014) showed that before the first 11 months, babies already produce synchronous pointing gestures at the babbling stage with the prominence in gesture (i.e. the maximum extension point of the arm when the baby is pointing to an object) corresponding with the prominence in the utterance (i.e. the pitch peak accent of the word). The emergence of the

multimodal communication with an explicit purpose is crucial, and predicts the correct lexical and grammatical development (Iverson & Goldin-Meadow, 2005; Murillo & Belinchón, 2012; Wu & Gros-Louis, 2014; Igualada, Bosh & Prieto, 2014). Further, the simultaneous production of gestures with speech reflects the communicative intentions of babies to their interlocutors, and that their hands may serve to convey them. Also, it suggests that soon in their early months, humans learn to use multimodal communication to modulate listeners' attention and minimize their communicative efforts to convey a message in joint attention contexts. Igualada, Bosh and Prieto (2014) investigated the ability to combine gesture with speech according to the social context at 11 months (if the experimenter visually responded or not to the child when he pointed at a stimulus for example) predicted the subsequent language acquisition at 18 months. They showed that children that used more frequently multimodal communication in socially demanding conditions were also those who demonstrated a better vocabulary acquisition seven months later. Later on, adults maintain predominant multimodal communication to convey information even if lexical, vocabulary and grammatical acquisitions are fully achieved. As language acquisition goes on, gestures diversify as well, according to their function and relationship with speech content, leading to a variety of gestures that can be classified in restricted main categories McNeill (1992).

1.3 Different categories of gestures and their alignment with verbal utterance

1.3.1. General structure of gestures

Although there are different categories of gestures in human communication, a basic common structure of the gestural movements is always found with a certain number of sequential gestural phases (Wagner et al., 2014; Kendon, 2004):

(1) the *resting* phase, which is the immobile position from where the gesture is initialized.
(2) the *preparation* phase, which is the movement initiated from the rest position to reach the communicative moment of the gesture.
(3) the *stroke* phase, which ends at the meaningful moment, conveying the communicative function of the gesture. During this phase, the hand shape describes the semantic content.
(4) the *hold phase*, which is an immobile phase occurring after the peak of effort of the stroke.
(5) the *retraction* phase is the phase in which hands are retracted back to the resting position.

According to their functions, gestures will vary particularly during the stroke phase. Indeed, some gestures get to a culminant peak in which the hand shape becomes fully meaningful; others will never describe a clear semantic content because it is more their movement synchrony with speech modulations that is functional. However, there is a moment at the end of the stroke that is commonly found across the principal gesture categories, and which is the point of maximum hand extension in space (Wagner et al.,

2014; McNeill, 1992). This moment called *apex*, reflects the maximal muscular effort of the movement in speaker's space, and marks the end of an acceleration phase, as a hit or a change of direction (Leonard & Cummins, 2010; Kita et al. 1998).

1.3.2. Different categories of gestures

As previously suggested, gestures may play different roles in communication both on the speaker and the listener sides. Here I present the main categories of gestures, according to their possible semantic function (McNeill, 1992; Wagner, 2014).

(1) Iconic gestures: the shape of the hand conveys the physical aspects of an object or an action that are described in the accompanying speech. For example, the stroke phase describes a round shape evoking a ball when the speaker is speaking about playing basketball. Even if describing concrete entities, iconic gestures are dependent from speech as they are difficult to precisely interpret without accompanying utterance (i.e., the round shape in the example above could be difficult to pinpoint to a ball if seen outside the context of the conversation about basketball).

(2) Emblems: These gestures are highly cultural dependent as they convey conventionalized meaning that can be understood even without speech (for example, the "thumb up" meaning "all is good").

(3) Metaphoric gestures: theses gestures are iconic gestures but the pictorial content describes an abstract idea rather than a concrete object or action. For example, the speaker can touch the fingertips

of both hands to illustrate a deep relationship between twins (Nagels et al., 2013).

(4) Deictic gestures: theses gestures are classically a pointing during the narrative, serving to point out localization in abstract conceptual space. The speaker is not interested in the abstract location itself, but from the previous context of narration, he uses it to refer to a concept. For example the speaker allocates this space by pointing to the right to refer to a house where the story began. Then, he points out to the right any times he goes back to the house in the narrative.

(5) Beats: These are very simple gestures without semantic content in their shape. Rather, beats are rapid biphasic hand movements that tend to have the same shape independently from speech content. For example, beats can be up and down flicks of the hand. Beats index affiliated words as being relevant for their pragmatic content. In other words, beats contribute to the perceived prominence of accompanying speech segments and refer directly to the speaker, rather than content. When beats are produced in succession to emphasize the continuity of different points belonging to a common concept, they are called cohesive. In the present thesis, I always refer to these McNeill's classification.

It is worth noting that McNeill's classification is not the only possible. If we consider the degree of dependency between gestures and verbalization for example, one can generate a different continuum (Kendon, 1988): Gesticulation→ Language-like Gestures→ Pantomimes→ Emblems→ Sign Languages. Here, gesticulations refer to all the gestures that we never produce out of

speech (utterance obligatory). The language-like gestures refer to gestures that are grammatically integrated into speech (high dependency). For example, the hand shape can replace an adjective that would normally be uttered at the end of a sentence. Pantomimes are gestures depicting actions that are understood without accompanying speech (Willems, Özyürek & Hagoort, 2009). For example, the hand movements that we produce when we play to make people discover a job's name or else without speaking (low dependency). Emblems are those previously described in McNeill's categorization (highly conventionalized). Sign Languages constitute a special category as they are an entire language system with segmentation, lexicon, syntax and all the language-like rules.

1.3.3. Two main functions of co-speech gestures

From these categories, it appeared that gestures may play two main functions in accompanying discourse (Kendon, 2004; McNeill, 1992): First, the *substantive* gestures that contribute to the speech content, conveying redundant of additional semantic information which is not present in the verbal utterance (emblems or iconic gestures for instance). For example, when speaking about a party, the speaker moves his hand describing a U-shape, as if bringing an imaginary glass to his mouth (i.e. iconic gestures). Substantive gestures can also describe some dimensions of object/action by mean of the hand trajectory, motion or speed. Second, the *pragmatic* gestures that do not convey clear semantic information in their hand shape. They bring additional information about speaker's attitudes, emotions or agreement between the speaker and the

listener (deictic gestures for example). Also, they may play a role in attention by highlighting relevant information in the verbal utterance (i.e. beats). The pragmatic gestures can also serve to package speech units, linking for example various successive points of a discourse to a common main idea (i.e. cohesive beat gestures).

1.3.4. Three rules of synchrony between speech and gestures

Although gestures and verbalization convey information in different format, both modalities maintain a particularly precise temporal coordination during speech production. McNeill (1992) established three rules of synchronization between gestures and utterance, which are common to the different categories:

(1) The *phonological synchrony rule* states that the stroke phase of the gesture precedes or ends at the phonological peak syllable of the accompanied utterance to ensure the stroke to be integrated into the phonology of the corresponding word. The phonological synchrony is illustrated when a speaker misses his words. Even if the speaker gestured an object before finding the corresponding word, he holds the hand with the meaningful shape until it comes (also called post-hold stroke). The only condition to respect the phonological synchrony rule is to maintain the natural order of gestures initiation preceding peaks onset.

(2) The *semantic synchrony rule* states that gesture and speech describe the same meaning (i.e. idea unit) at the same time. Gesture can convey redundant or complementary semantic content to speech, but it never has an incongruent meaning (even if

theoretically a speaker could produce an unrelated gesture with accompanying speech).

(3) *The pragmatic synchrony rule* predicts that gesture and speech have the same pragmatic purpose. Verbal utterance conveys pragmatic details which help to describe the embedding context of a story (for example the characters of the story). At the same time, the gesture describes a bounded object to represent the story as a whole. Here, gesture and speech come together on a common pragmatic level to introduce respectively the entire aspect of the story and the main characters (McNeill, 1992).

The different types of gestures can be more substantive or pragmatic and it is not always easy to distinguish which synchrony rule applies more, or which utterance component (prominent syllable or word) are engaged when speaking about gesture and speech synchrony.

1.4 Gestures influence speech production at different possible stages

Gestures facilitate speech production and, importantly, speakers experience difficulty when they have to speak without gesturing. McNeill (1992) described this close relationship between gestures and speech production, underlining a certain number of common characteristics. Perhaps, the most relevant are the fact that people usually gesture only during speech production; gesture and verbal utterance are highly synchronous; gesture and speech break down together in aphasia. Based on that, gestures and speech may

form a synergy in which gestures help speech production, by conveying additional information that does not need to be verbally described. This suggests that the speaker has to conceptualize speech both in gesture and verbal modalities. As gestures always start before the affiliated utterance, some models posited from the perspective that gesture modulates verbalization in different manners. Here I present three main models that attempted to describe where and how gesture and verbalization interact during speech production. The first one (LRH) describes a local effect of gestures that may facilitate the lexical access of corresponding verbal information. The second one (VWM) explains how gestures may decrease the working memory load during speech production. Finally, the third model (IPH) addressees how gestures may help speech production by facilitating the organization and conceptual planning of the discourse.

1.4.1. The Lexical Retrieval Hypothesis (LRH)

Producing accompanying gestures may facilitate speaker's lexical access during speech by facilitating the associated word activation (Beattie & Coughlan, 1999, 1998; Rauscher, Krauss & Chen, 1996). The Lexical Retrieval Hypothesis (LRH) states that gestures representing semantic content in their shape facilitate lexical access by cross-modal priming. As gestures are generally initiated before the articulation of lexical affiliates, the motor representation of the concept described by the gesture primes the phonological representation of the words associated to its verbal description in speech (Rauscher, Krauss & Chen, 1996; Krauss, 1998; Gillespie et

al., 2014). Concretely, when speakers were not permitted to gesture to describe spatial content, speech fluency was affected by an increase of non-juncture filled pauses (associated with lexical retrieval difficulty, like “uh” or “hum”), and a decrease of velocity (word per second), indexing difficulties to access to their mental lexicon speech (Rauscher, Krauss & Chen, 1996).

1.4.2. The Verbal Working Memory (VWM) Hypothesis

Alternatively, the meaningfulness of the gestures and their temporal synchrony with corresponding speech may lighten the load of Verbal Working Memory (VWM) during production (Gillepsie et al., 2014; Cook, Yip & Goldin-Meadow, 2012; Baddeley, 1992). How gestures may reduce the working memory demand is still unclear but different hypotheses have been advanced. As gestures convey visual information, they may provide a previous sketch then facilitating speech production in a discrete format following complex linguistic rules. Also, gestures convey information in the visual modality, in contrast to speech that conveys it mostly through the auditory modality. The overlap of redundant audio-visual information may decrease the working memory load respect to maintaining content in a single modality (Cook, Yip & Goldin-Meadow, 2012; Goolkasian & Foos, 2005). Gesturing may also help to lighten the VWM by helping the speaker to remain focused on speech content by decreasing mental distractions. That is, gestures may constrain the speaker to remain concentrated on the initial idea he/she wants to express by speech, and would act as a filter against

distractions (Cook, Yip & Goldin-Meadow, 2012; Engle, 2002; Cowan et al., 2002). Different redundant models have been proposed to attempt to establish the relationship between gestures and speech, in relationship to working memory (Krauss & Hadar, 1999; de Ruiter, 2000 and Kita & Özyürek, 2003). According to Krauss and Hadar's model (1999, see Fig. 1), gestures originate from the spatial-dynamic representations in working memory that activate the feature-selector system to select elementary specifications of the movement (velocity, direction…). A motor planner translates the set of abstract movement features in a motor program that contains the instructions to execute the lexical gesture. Then, the motor system executes the instructions in the form of a gestural movement reflecting the lexical features (for example, if the abstract feature was "round", the gestural movement will depict a U-shape hand at the hand of the motor system execution). Finally, gestures are monitored to ensure congruent kinetics with speech. The gesture system production may affect speech production at the formulator level (Baddeley, 1992) where the lexical retrieval takes place (Fig. 1). The lexical facilitation in the formulator might rely on cross-modal priming in which the features of the concept selected in working memory and formulated by the gestural representation, precedes the verbal formulation of those features.

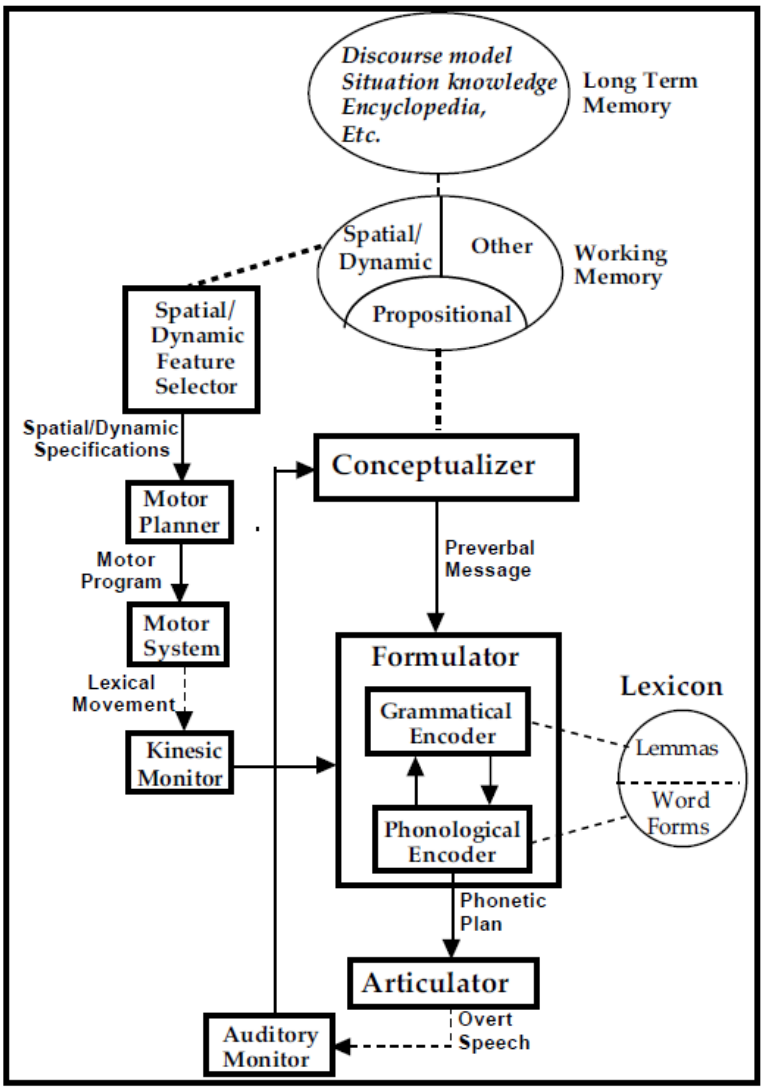

Figure 1. Interaction of the speech and gesture production systems and working memory (from Krauss & Hadar, 1992).

1.4.3. The Information Packaging Hypothesis (IPH)

Finally, accompanying gestures may help speakers to organize their narrative discourse (Alibali, Kita & Young, 2000; McNeill, 1992). This hypothesis has been exposed through the Information

Packaging Hypothesis (IPH). The IPH holds that gestures may facilitate the speaker's conceptual planning of the message. Basically, for a given lexical field, according to what the speaker wants to verbalize, he will produce qualitatively different gestures, even if the global vocabulary is the same. To test for the IPH, Alibali, Kita and Young (2000) investigated in children the production of gestures in two different conditions based on a Piagietan conservation task. In one case, children had to *explain* a situation after a change (i.e. *why* two items look different now?), while in the other one, they only had to *describe* it (i.e. *how* do they look different?). The conceptualization in the explanation condition was more complex and constraining than in the simple description, as speakers had first to decide if the two items were different and second, identify the dimensions relevant to the comparison. Alibali et al.'s results showed that children produced more gestures conveying dimensions of the objects (width…) by means of hand shape, motion, or placement (i.e. substantive gestures) in the explanation condition than in the description one. Additionally, the gestures contained less redundant information respect to the accompanying utterance because gestures had to bring very specific features that were more difficult to verbalize than in the simple description. The authors concluded that gestures helped speakers to conceptualize the message, depending on the conditional planning (explanation or description) to facilitate verbalization.

From a production perspective, speakers naturally gesture in temporal and semantic congruence with speech. Undeniably,

gestures promote language acquisition and later facilitate its production and transmission. However, a different and central question is to know what is the impact of these gestures on the listener (if they have an impact at all)? If so, which levels of speech perception are affected by co-occurring gestures when someone listens to a gesturing speaker? With the assessment of experimental procedures combined with neuroimaging techniques, a growing number of studies have recently evidenced the modulation of speech processing by gestures at neural levels.

In the next section, I report relevant studies that investigated the influence of gestures on speech processing on the listener's side. From my viewpoint, this will shed light on why new approaches are needed to investigate gestures neural correlates during speech perception. Indeed, most of the reported studies, although very relevant, focused on meaningful gestures (i.e. iconic or pantomimes) presented in very restricted speech contexts (isolated sentences or gestures for example). Thus, it will appear clear that using less elaborated gestures (i.e. beats) and change procedures of presentations (for example using continuous speeches) may constitute a more naturalistic alternative to investigate gesture-speech processing and their neural correlates.

1.5 Gestures and speech processing on the listener's side

From the last two decades, an increasing number of studies attempted to isolate the time course and the neural correlates of gestures during speech processing by combining behavioral procedures with ERP and fMRI recording. They have reported that gestures modulate different stages of speech processing, and their processing relies on a restricted neural network including language related brain areas.

1.5.1. The time course of gesture and speech processing

Kelly, Kravitz and Hopkins (2004) reported an ERP experiment in which participants were presented with short AV clips and had to attend to speech content only. Kelly at al. demonstrated that iconic gestures affected auditory processing at an early phonological integration stage of processing. When the gesture conveyed incongruent as compared to redundant information of verbal utterance, the ERP signal was modulated from 100 ms to around 200 ms after the corresponding word onset, corresponding to the moment of phonological processing. This time window corresponds to the N100/P200 classic ERP component (also called “N1-P2”), which has been described to reflect also multisensory processing in audiovisual speech (Stekelenburg & Vroomen, 2007; van Wassenhove, Grant & Poeppel, 2005; Näätänen, 2001; Rugg & Coles, 1995). In another experiment, listeners attended to audio-visual clips in which the speaker described a critical word by means of speech and spontaneous gestures (Wu and Coulson, 2010). The authors also reported less negative ERPs from 200 ms after the

onset of critical word when it was accompanied by spontaneous gestures as compared to when it was pronounced without gestures.

Both in Kelly, Kravitz and Hopkins (2004) and Wu and Coulson (2010) studies, later-occurring semantic stages of speech processing were modulated by the presence of gesture as well. Indeed, in Kelly Kravitz and Hopkins (2004), gestures that were semantically incongruent with speech content elicited more negative ERPs in a temporal window corresponding to the N400 of the targeted words, respect to words alone. The N400 is a negative ongoing component that reflects semantic integration, increasing when the integration of a word in context (i.e., a sentence) becomes difficult (for a complete review about the N400, see Hinojosa, Martin-Loeches & Rubia, 2001). More generally, semantic processing stages have been largely used to index the influence of gestures on audiovisual speech processing. Holle and Gunter (2007) presented participants with audiovisual sentences containing an ambiguous homonym (for example “mouse” can mean the animal or the computer tool) in their initial part that was disambiguated by a subsequent target word. The speaker also produced an iconic gesture with the homonym that semantically supported either one meaning or the other. The N400 was significantly smaller when gesture and target word were congruent both for dominant and subordinate meanings of the homonym. These results suggest that listeners can implicitly use the content of an accompanying gesture to facilitate the semantic processing of ambiguous sentences. More recently, other ERP studies investigated the influence of gestures at a syntactic parsing level of ambiguous sentences. Holle et al. (2012) used

German sentences that were structurally ambiguous with respect to their subject and object. German sentences in active form have the first noun as the subject and the second as the object (preferred structure; SOV), but in a passive form, the roles are inverted without changing the meaning of the sentence (complex structure; OSV). In Holle's experimental materials, the structure interpretation depended on a final-sentence critical word. In the audio-visual clips, the speaker produced a co-occurring beat gesture either with the first noun or the second noun to facilitate the syntactic analysis of the sentence before the critical word. When the gesture emphasized the second noun in the complex structure, the syntactic parsing was facilitated, as a decrease of the well-established P600 ERP component at critical word was observed. The P600 is a positive going wave reflecting some aspects of the syntactic analysis during sentence processing and it increases with ambiguity (van de Meerendonk et al., 2010; Haupt et al., 2008; Friederici, 2002; Frisch et al., 2002).

1.5.2. Localization of the neural correlates of gestures

Some fMRI studies investigated the localization of the neural correlates of gestures during AV speech processing. Holle et al. (2008) adapted the paradigm they used for their ERP study described above (Holle & Gunter, 2007) to fMRI. They compared the processing of an iconic gesture that could be congruent with the dominant or subordinate meaning of an accompanying homonym word, with simple grooming gestures that do not convey any communicative information (i.e. scratching). They hypothesized

that the brain areas engaged in the processing of meaningful gestures accompanying speech may show greater activations than simple grooming. Indeed, the processing of iconic gestures with corresponding speech elicited greater activations in the left posterior Superior Temporal Sulcus (left post STS), as compared to simple grooming meaningless gestures. The STS is known to be an important multisensory site and respond to audiovisual speech (Nath and Beauchamp, 2012; Calvert et al., 2000; Callan et al., 2004; Macaluso et al., 2004; Meyer et al., 2004; Campbell, 2008). For example, the left STS has been shown to be involved in the integration of lip movements with speech (Sekiyama et al., 2003; Calvert et al., 2000). As iconic gestures interpretation depends on the semantic context provided by the accompanying utterance, Holle et al. results suggest that the greater activations in the left STS reflect gesture and speech interactive comprehension rather than simple hand movement perception. In contrast, the weaker activations in the left STS when speech came with simple grooming movement suggest that they did not interact in a meaningful way.

In another fMRI study, Willems et al. (2007) modulated the semantic relationship between an iconic gesture and the verb of the sentence in order to increase the semantic integration load. For example, the gesture and the verb could be congruent and semantically correct in the speech context (i.e. the condition in which the semantic integration was easier). In the worst case, gesture and verb were both semantically ambiguous respect to the speech context (i.e. the condition in which the semantic integration load was the highest). The results showed an effect of semantic

integration load particularly in the left Infero Frontal Gyrus (left IFG) where activations were decreased when gesture and verb semantically matched the speech context, respect to the other more semantically demanding conditions. Interestingly, the left IFG is thought to be engaged in the non-specific unification of multimodal complementary streams to facilitate language comprehension as well as semantic processing in sentence context (Hagoort, 2005; Hagoort, 2003; Friederici et al., 2003). Here, the left IFG appeared to be sensitive to the semantic relationship between gesture and corresponding speech (Willems et al., 2009; Dick et al., 2009; Willems et al., 2007; Skipper et al., 2007). Finally, Willems et al. (2009) investigated the influence of the degree of dependency between meaningful gestures and speech on the neural activations during perception. They compared the effect of semantic incongruence on neural activations for speech accompanied either by iconic gestures (speech dependent) or pantomimes (easily understood without speech). The fMRI data revealed differences of sensitivity to incongruence between speech and the type of co-occurring gesture. Specifically, the authors found that the posterior STS/Medial Temporal Gyrus (i.e. post STS/MTG) was only sensitive to incongruence between speech and pantomimes. In contrast, the left IFG activations were modulated both by the incongruence between speech accompanied by iconic gesture and speech accompanied by pantomime. The modulation of activation in the post STS/MTG only when speech comes with pantomimes suggests that speech accompanied by pantomime convey two stable representations in both audio and visual modalities, engaging lower

levels of multimodal stimulus processing (the pantomime explicitly describes the verb contained in speech). The sensitivity of the left IFG to incongruence, irrespectively to the type of gesture supports the hypothesis that gestures in general are perceived as complementary information processed with speech stream to facilitate comprehension (Hagoort, 2005; Hagoort, 2003; Friederici et al., 2003). Further, the engagement of the left IFG reflects higher levels of semantic integration as the unification of gesture with speech requires the construction of an entire multimodal representation in the case of iconic gestures.

Although different degrees of semantic relationship between gesture and speech engage distinct neural correlates, a recent meta-analysis of neuroimaging studies attempted to determine a common neural network of gestures in general (Marstaller & Burianová, 2014). Based on six studies including iconic, metaphoric and beat gestures, the authors identified a restricted neural network responding to the multimodal (speech accompanied with gestures) in contrast to unimodal (speech or gesture) speech perception, and that engages two main mechanisms. A first component of this network would include the temporal regions related to auditory and movement perception with increased BOLD responses in the right auditory cortex as well as the left posterior STS for gesture-speech perception. The right auditory cortex has been hypothesized to sample the spectral auditory signal and extract prosodic aspects of speech, in particular the Planum Temporale (Griffiths & Warren, 2002; Zatorre & Gandour, 2008). Gesture may be processed with

prosodic features during perception, facilitating the segmentation and low level processing of speech. In line with it, some ERP studies demonstrated that the semantic processing is effectively sensitive to the temporal synchrony between gesture and speech (Habets et al., 2011; Obermeier, Holle & Gunter, 2011; Obermeier & Gunter, 2014). The left STS has been shown to participate in audiovisual speech integration, as previously explained, but might also support the processing of biological movement per se (Pavlova, 2012; Pelphrey et al., 2005). A second component of this network would include fronto-parietal regions related to action understanding (de Lange et al., 2008), that exhibit greater activations in the ventral premotor and the infero-parietal cortices when speech comes with gestures as compared to unimodal presentations. This may reflect the perception of gestures as intentional communicative movements (Marstaller & Burianová, 2014; Wagner et al., 2014). The basic gesture correlates can be seen in the figure 2.

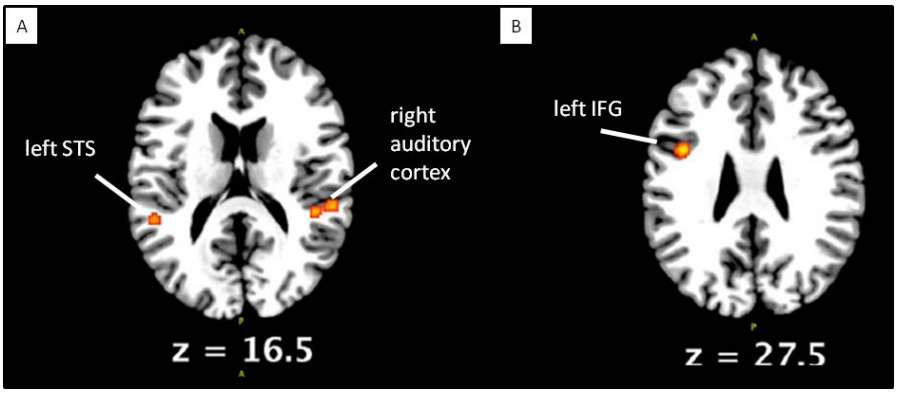

Figure 2. Basic neural correlates of gestures (adapted from Marstaller & Burianová, 2014).

1.5.3. The starting point for us: Need for new approaches to investigate neural correlates of gestures

The results discussed above report precious evidence about the influence of gestures on the listener's side. These studies are pioneering, as neuroimaging research on the topic of gestures is scarce. They have allowed establishing the time course of the impact of gestures on speech processing and part of their neural correlates, depending on their semantic relationship or even shape content.

Nevertheless, the experimental procedures by which gestures were presented may lack ecological validity. Indeed in the field of gesture investigation, most of the paradigms used short audio videoclips in with a sentence is generally accompanied by an isolated gesture produced in a discrete manner. Instead, in natural conversations or public addressees, speech and gestures constitute two continuous streams that unfold temporally and semantically aligned. This explains why it turned out difficult to determine distinct gesture categories and led to the establishment of at least four different continua (McNeill, 2000; McNeill, 1992). Further, speakers normally embody successive gestures in a common concept to discuss a point. Presenting a single and spatiotemporally well delimited gesture, aligned with short speech fragments without previous context may have artificially increased the saliency and modulate the legitimacy (i.e. would one really have produced this gesture to describe this particular sentence?) of the gesture respect

to natural situations. Finally, as previously evoked, almost all the studies focused on meaningful gestures (iconic, pantomimes or metaphoric) to investigate the neural correlates of gesture-speech processing. As far as we know, only three studies used non-elaborated gestures (beats) to investigate gesture correlates (Wang & Chu, 2013; Holle et al., 2012; Hubbard et al., 2009), which is quite surprising as they are the most frequent type of gestures in narrative discourses (McNeill, 1992). This may be explained in part by the fact that, in controlled conditions (i.e. lab conditions), iconic gestures have a clear stroke phase that matches well the corresponding utterance segment, whereas beats are difficult to isolate without losing their functionality. Or else, iconic gestures looked more appealing respect to simple flick of the hand. An alternative manner to investigate the neural correlates of gestures might be to actually focus on these beat gestures conveying less semantic content in their hand shape, but whom the flow of production is maintained integrated with continuous speech. Adopting a more ecological approach may preserve the natural function of gesture accompanying speech and how listeners perceive them normally when attending to the speaker.

In the following section, I will first describe beat gestures and report empirical evidences suggesting that they impact speech processing at behavioral and neural levels as well. At the same time, I will underline the fact that the same methodological issues raised for iconic gestures studies, apply on beat studies as well. Second, I will present how public speeches (e.g. political discourses)

constitute a valuable context to present beat gestures because they conserve the temporal and pragmatic alignments with verbalization, allowing investigating beats correlates in close-to-natural conditions.

1.6 Beat gestures: General description

Although beats are simple hand gestures, appearances are misleading. Beat gestures are typically rapid biphasic flicks of the hand(s) in one dimension like up and down, or back and forth movements (McNeill, 1992). The hand shape is independent from speech content. But the fact that beats do not explicitly convey any semantic in their shape does not mean that they do not have any communicative value. Usually, speakers produce beat gestures to emphasize relevant information, or to accompany words when they want to make a digression during the narrative (accompanying the conjunction '*but*', for instance). Consequently, beats serve to bring additional information that is not explicitly present in speech, conferring them a pragmatic function that requires a mutual comprehension from both speaker and listener.

As beats are rapid, their core functionality may reside in the high temporal alignment between speech envelop and beats' apexes (the maximum extension point of the arm before retraction, corresponding to the functional phase of the gesture). Naturally and with a great consistency, speakers synchronize beats' apexes with the stressed syllable of the affiliated words. Using audiovisual recordings of three different speakers Yasinnik, Renwick and

Shattuck-Hufnagel (2004) marked separately beats apexes from the video and prosodic cues from the audio (pitch accents and intentional phrase boundaries). The authors found that, in more than ninety per cent of the cases, the gesture apex occurred with a pitch-accented syllable (raise in F0, i.e. fundamental frequency). The authors suggested that beat gestures temporally align with the prosodic structure of the verbal utterance (F0 height), suggesting that when speakers plan the prominent patterns of their speech, they do the same in the gestural modality as well. Interestingly, the production of a co-occurring beat with its corresponding word has significant acoustic consequences on the corresponding syllable. Krahmer and Swers (2007) investigated the influence of beat production on the prominence of the accompanied words in the verbal utterance (i.e. the strength of the accentuation). Ten participants were instructed to utter short sentences in a neutral manner, or stressing the pitch accent on one of two possible target words. Additionally, they had to produce a beat congruent or incongruent with the pitch accentuation. Results showed that beats modulated the acoustic properties (length, F2 frequency) of the corresponding syllable in a similar manner, as did the pitch accentuations, even when the syllable was not voluntarily stressed.

The fact that speakers naturally produce beats in accordance with the prosody structure of speech and this production modulates the acoustic properties of the corresponding segments raises the following questions: Do beat gestures modulate speech processing on the listener's side as well? If so, which levels of speech

processing are modulated by beats and what are their neural correlates?

1.7 Beats impact speech perception

1.7.1. Behavioural evidence for the effect of beats on speech processing

Only a handful of studies have investigated the effects of beat gestures at behavioural level. Here, I report evidence supporting the assumption that listeners integrate the speakers beat gestures with the speech signal, as a (visual) part of the language stream, rather than simple hand movements.

At first, if listeners associate beats with prosody during speech perception then they should be sensitive to asynchrony between the two streams. That is, they have a representation of the normal timing between both modalities and thus, detect deviations from this alignment that eventually affects the processing. Treffner, Peter and Kleidon (2008) investigated the effect of speech-beat timing on sentence perception by presenting participants with audiovisual sentences in which the speaker produced a unique beat gesture. The temporal beat-speech alignment was shifted to gradually synchronize the apex from a word to the following. Listeners had to determine which word was the intended focus of the sentence. Results demonstrated clearly that the perceived prominence shifted with the beat-speech alignment from one word to the other. As prosodic information was removed from speech,

these results suggests that listeners can infer an intended focus from the kinetics of the beats, but also that beats can modulate the interpretation of sentences only by their temporal alignment with speech. Later, Leonard and Cummins (2012) measured the sensitivity of listeners to the temporal relation between beats and speech. In short audiovisual clips, the authors gradually shifted the video from 0 to 800 ms respect to audio in both directions (gesture either preceded or lagged respect to the corresponding speech segment). For each clip, participants were instructed to decide if audio and video were synchronized or not. Results showed that listeners were sensitive to an asynchrony between beat and corresponding word in both directions. Particularly, when gesture lagged respect to audio, listeners were able to detect asynchronies as short as 200 ms. Further, the authors performed a qualitative analysis of the relation between speech and beat gesture to determine the anchor points in speech (vowel onset, pitch peak…) that have the more stable temporal relation with relevant kinetic landmarks in the gesture (gesture onset, velocity peak, apex…). Their results confirmed that the gesture's apex and the pitch peak in the stressed syllable of corresponding word exhibited the most stable temporal alignment between all the different possible speech/gesture anchors.

Both studies described above show that listeners are sensitive to the temporal alignment between beat gestures apexes and stressed syllable of the corresponding affiliate word. But, one further question is: How this co-occurrence affects speech

processing in the listener? Back to Krahmer and Swers (2007), the second part of their study evaluated the influence of beat gestures on the listener's perception of prominence. The authors showed that beats significantly increased the perceived prominence of the accompanied word (when it was pronounced with a pitch accent) and decreased the prominence of the accented word in case of mismatch (i.e. when the beat targeted the other word). When none of the two target words were accented, these effects of beat were still true. Finally, in their study, Khramer and Swers (2007) also evaluated if targeted words were perceived as more prominent in audiovisual conditions (seeing the speaker) than in audio only conditions. Results showed that beats effectively improved perceived prominence of accented words and decreased perceived prominence of the other word (in case of mismatch) as compared to prominence perception in audio only conditions. These second results suggested higher pragmatic level functions, because the fact that listeners perceived greater prominence of word when speakers' hands were visible implies that they understand the communicative value of beats even if not explicit in the utterance. Co-occurring beats modulate speech processing at phonological processes and seem to establish a mutual pragmatic synchronization between the speaker and the listener by emphasizing the audio prosody ("I know that the utterance accompanied by the beat is important").

One manner to evaluate the impact of beat gestures on listeners is to find a behavioural index to qualitatively evaluate speech processing. So, Chen-Hui and Wei-Shan (2012) investigated

the mnemonic effect of beat gestures in adults and children to measure the quality of audiovisual speech processing. In a first experiment, adults were presented with three types of lists of isolated words pronounced by a speaker presented audio-visually: words accompanied either with an iconic gesture, with a beat gesture, and without a gesture (words pronounced alone). After the presentation, participants were asked to recall as many words as possible. Results showed that listeners recalled more words that had originally been accompanied with iconic gesture than words alone. But more interesting, words accompanied with a beat gesture were remembered the same as words accompanied with iconic gesture (thus, more than words alone). In a second experiment, the authors ran a similar procedure with 4-5 year old children. As with adults, words accompanied with iconic gestures were better recalled than words alone. In contrast, children did not recall more words accompanied with beat gesture than words alone. Taken together, these results showed that beat gestures improved memory recall for words in adults, suggesting that they improved encoding during speech processing. As beats have been shown to influence the perception of speech prosody and to increase the perceived prominence of corresponding affiliate words (Krahmer & Swers, 2007), one possible explanation of this advantage may be that beats cross-modally modulate activity in the auditory cortex during speech perception (Marstaller & Burianová, 2014; Hubbard et al., 2009), in a similar fashion as what has been suggested for visual speech (van Wassenhove, Grant & Poeppel, 2005; Nath and Beauchamp, 2012; Calvert et al., 2000; Callan et al., 2004;

Macaluso et al., 2004; Meyer et al., 2004; Campbell, 2008). However, the fact that their mnemonic effect was not found in children suggests that beats engage higher cognitive processes as well, that are needed to interpret (enable) their communicative value (i.e. at pragmatic levels). These social skills may require longer communicative experiences, later after the first life years (So et al., 2012; McNeill, 1992).

1.7.2. Neuroimaging evidence of beats effects on speech processing

Very few studies have investigated the time course of beats processing during speech perception and their neural correlates. In a previous section (*1.5.1 The time course of gesture and speech processing*), I have already discussed an ERP study from Holle et al. (2012) that investigated the possible role of beat gestures in syntactic analysis during ambiguous sentences comprehension. They showed that the presence of a beat gesture on a critical word in the complex form of ambiguous sentences facilitated the syntactic analysis, as the P600 component was significantly decreased. More recently, another study investigated the possible role of beat gestures on semantic processing during speech perception. Using the ERPs, Wang and Chu (2013) compared the semantic processing of a critical word in short sentences, when it was accompanied by a beat gesture, a control hand movement or pronounced alone. Results showed that beats elicited more positive waveforms than the word presented alone or with control hand movements around the critical word onset. Further, in the N400

time window, beats elicited less negative waveforms (that is, again a positive shift) than the word alone or accompanied with control movements. As the N400 strength is measured as a negative shift, this result suggested that beats facilitated semantic processing of the affiliated word during sentence perception. Moreover, this result supports the hypothesis that, even if very rudimentary, beats carry communicative intentions from the speaker and are perceived differently from simple hand movements. This is in line with a previous study suggesting that beats engage higher cognitive (So et al., 2012) to interpret implicit aspects of speech. Finally, only one study investigated the neural correlates of beat gestures using fMRI (Hubbard et al., 2009). Listeners were presented with audiovisual clips featuring a speaker who produced spontaneous beats unaware of the purpose of the experiment while speaking. So, in these materials, gesturing occurred in a natural, speech context. In three additional conditions, the original video was replaced by another in which the speaker produced either non-communicative gestures (like scratching), sign language gestures, or simply stood still. Results showed greater activations in the left STG/S in response to speech when it was accompanied by beat gestures as compared to when it was presented with unrelated sign language gestures. The authors also reported greater BOLD responses in the bilateral posterior STG/S, including the Planum Temporale (PT), when participants listened to speech accompanied by beats compared to a still body. When speech was removed in control conditions, beats did not modulate BOLD responses differently from simple hand movements. These results are in line with previous

behavioural/ERPs studies as they showed that beats engaged multisensory (left STG/S) and acoustic processing (PT) areas, and were processed differently from non-communicative movements. Further, they showed that beats have to accompany congruent speech to be processed as linguistic information.

1.7.3. Methodological issues and need for new materials

All together, these studies (both behavioural and neuroimaging studies) demonstrated that beats are a valid model to study gestures processing and their neural correlates. Nonetheless, as previously discussed for iconic gestures, these studies investigated beats in very artificial and restricted contexts of production (except for Hubbard et al., 2009). The speakers were often aware of the goal of the study and were instructed to produce a deliberate beat gesture on a particular word (So et al., 2012; Krahmer & Swers, 2007). Trying to voluntarily execute a pre-planned beat at a particular point of a sentence is difficult, and especially challenging if the goal is to synchronize the apex with a particular accented syllable and make it sound natural. This defeats the very essence of "spontaneous" beat gestures in natural speech. Further, the materials consisted of short sentences containing one single beat gesture which, from my viewpoint, raises two principal issues. First, these sentences may not constitute a natural semantic or syntactic context in which one normally would have produced a beat gesture (Wang & Chu, 2013). In Wang and Chu (2013) for example, the beat always accompanied grammatically critical words but not other classes of words. However, beats often come with conjunction words as well (for

example “but”) when the speaker adds pragmatic information (McNeill, 1992). Also, as sentences were isolated, the poor semantic context (and the absence of a previous context) did not allow to fully understand why a beat had to be produced with a noun and it may result trivial to listeners. Considering that beats convey pragmatic and emotional information reflecting the engagement of the speaker, it may appear artificial to produce a salient beat to accompany short sentences. In other words, because beats do not appear at their normal rate, syntactic context and with the normal variability, this may induce subjects to pay attention to them in a different manner than they would do normally, becoming artificial temporal cues. As beats are highly temporally aligned with prosody (i.e. rhythmic modulations of acoustic envelop of speech), this implies a certain continuity to establish a stable and fluent congruence between gestures and speech streams. Consequently, there is evident need for searching new speech contexts to present beats in more natural conditions. Taking into account all these issues, we attempted to find natural situations that may be particularly suitable for the production of spontaneous beats, to overcome the restrictions imposed by the laboratory conditions.

Actually, there is context of public addressees in which beat gestures are the most frequent gestures: the political discourses. In his book, McNeill (1992) described “*Political speeches are accompanied by an incessant beat presence*” and “*The beat is accordingly the politician’s gesture per excellence*” (p16). During their public speeches, politicians produce a lot of beat gestures which have two principal functions: First, discrete beats serve to

highlight the discontinuities in the narrative, to introduce details or focus attention on important information. Second, successive beats (also called "*cohesive*") serve to mark a series of points belonging to a common argument. In this case, the cohesive beats tend to have the same trajectories/hand shape to underline the repetition and the continuation of the idea. Politicians can also use beats to organise their ideas and structure the narrative discourse. For instance, Casasanto and Jasmin (2009) examined the gestures produced by politicians during public debates (as they looked at the one hand gestures, most of them were beats) and showed that they associated their dominant hand with positive points and their non-dominant hand with negative point. These results suggested that beats provide also implicit information on how the speaker feels about the content of corresponding speech. Political discourses provide particularly well suited material if one considers gestures and speech as both complementary sides of a common language system (McNeill, 1992; Kelly, Creigh and Bartolotti, 2010) because they maintain the continuous flow of both visual and audio streams fully functional. Further, beats come in a more spontaneous way with their natural frequency as they would be embodied in a discourse respecting the narrative rules like adding details, successive arguments about a common point for example, in which beats play an important role (McNeill, 1992). Then, even if political speeches are sometimes well trained by coaches, they appear to be an interesting compromise also because people are familiar with this particular format of communication.

1.8 Scope of the present thesis: The current goals and overview of the experimental section

1.8.1. Hypothesis of the present thesis

The overall goal of this dissertation was to develop alternative experimental procedures to investigate the neural mechanisms related to gesture-speech integration during continuous speech perception. We designed new experimental paradigms combining the presentation of real AV political discourses with electrophysiology (ERPs/EEG) and neuroimaging (fMRI) recording techniques. In doing so, we focused on a particularly underestimated gesture type (i.e. beats) that predominates in public addressees. This new approach allowed to investigate gestures at a natural frequency of production and correctly contextualized by the accompanying continuous verbalization.

To test for our original approach, we focused on the temporal aspects of the relation between beats and utterance. As previously described, beats are initiated before corresponding words onset and their apexes co-occur with pitch peaks of speech prosody. Keeping this in mind, we made three hypotheses that we tested to demonstrate that beats are visual linguistic information of speech (and can be considered as visual prosody matching the speech envelope modulations):

(1) Beats modulate early stages of audio processing during continuous speech perception.

Previous studies reported that beats onset always precede corresponding words onset (Treffner, Peter & Kleidon, 2008; Leonard & Cummins, 2012). Further, the production of a beat modulates significantly the acoustic properties of the accented syllable (increase of pitch accent, and loudness and duration), increasing the saliency of the corresponding word (Krahmer & Swers, 2007). Consequently, listeners perceived affiliated words as more prominent in short sentences. Based on these evidences, we first hypothesized that the presence of a beat may affect the phonological processing of corresponding words during speech perception. At neural levels, we expected to find an influence on the ERPs reflecting phonologic stages of affiliated word integration during continuous speech perception. More precisely, we predicted an effect of beats in an early time window corresponding to the N1/P2 ERP component reflecting the multisensory integration and phonological processing of AV speech (Stekelenburg & Vroomen, 2007; van Wassenhove, Grant & Poeppel, 2005; Näätänen, 2001; Rugg & Coles, 1995). Such an effect may rely on attentional mechanisms by driving the listener's focus on relevant information during speech perception and support this hypothesis, originally developed by McNeill (1992).

(2) Beats bear a predictive value within the speech signal.

Second, we hypothesized that as gestures bear a possible predictive value on associated speech segments, they might be susceptible to diminish the uncertainty about when the corresponding acoustic cues will occur, to facilitate continuous speech processing (Arnal & Giraud, 2012). The consistency and the recurrence of perception order (a beat starts before the corresponding word and its apex falls on the pitch peak of the accented syllable) allow listeners to anticipate the relevant segments in the utterance marked by beats. We tested such a predictive value could be measured through the modulations of low frequency oscillatory activities as a possible neural signature of the integration between beats and auditory information. First, theta activity has been shown to mirror speech segmentation during its processing with an increasing of phase synchronization at word/syllable onsets (Giraud & Poeppel, 2012; Peelle & Davis, 2012; Luo and Poeppel, 2007; Greenberg, 1999). Second, it has been argued that this resonance between theta oscillatory activity and regular relevant acoustic cues can be modulated by stable preceding visual information reflecting temporal anticipation and facilitation (Arnal & Giraud, 2012; Lakatos et al., 2008; Schroeder & Lakatos, 2009; Schroeder et al., 2008). We predicted that beats might influence temporal anticipation through a greater increase of theta phase synchronization around affiliated word onsets, than equivalent words pronounced in the absence of a concurrent beat.

(3) Beats convey communicative value and are perceived as visual prosodic information.

Beats may be part of the same language system with speech, providing visual prosody when synchronized with utterance prosody during speech perception. First, we hypothesized that if the temporal alignment between beats apexes and pitch accents is broken by an asynchrony, then neural activations in language related areas may be modulated as well because beats are automatically integrated with prosody in normal conditions of speech processing. Based on previous fMRI studies (Marstaller & Burianova, 2014; Hubbard et al., 2009), we expected a modulation of neural activations in the left Inferior frontal Gyrus (left IFG) and left Superior Temporal Sulcus/Gyrus (left STS/G) when the temporal alignment between beats and speech prosody is affected.

To go further, we addressed whether the potential prosodic role of beats relies only on their emphasizing trajectories (velocity, directions and apexes) aligned with auditory envelop modulations or, whether beats engage a specialized mechanism because they convey additional communicative intentions of the speaker. To address this question, we added a manipulation in which we replaced the speaker's hands by moving discs that reproduced the original kinematics and spatio-temporal properties of beats (this manipulation will be described in details in the corresponding article). We hypothesized that simple emphasising spatiotemporal trajectories of arbitrary visual stimuli may not be enough to accomplish the same linguistic function that gestures have when combined with speech. At neural levels, we expected qualitatively distinct modulations of BOLD responses in the language related

areas by an asynchrony between speech and beats, or speech and discs.

1.8.2. Overview of the experimental section

The experimental section (section 4) of this thesis includes the three articles that report the results of these investigations published in international scientific journals. I will present each article individually, ordered according to the previously presented hypothesis. The articles will be:

2.1. Biau, E., & Soto-Faraco, S. (2013). Beat gestures modulate auditory integration in speech perception. *Brain and Language*, 124(2), 143–52.

In this article, we addressed the hypothesis (1). To do so, we investigated the time course of beat-speech integration during perception of a running discourse, to highlight the levels at which the co-occurrence of accompanying beat gestures may influence speech processing. We recorded EEGs from participants as they watched a pre-recorded TV broadcast of a political discourse. We extracted the ERPs time-locked to the onset of words synchronized with beat gestures, and compared them to ERPs from equivalent words pronounced without accompanying gestures in the same discourse. The latencies of the modulations will inform as to the level of processing at which gestures express their influence on speech processing.

2.2. 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.

The second article tested the hypotheses (1) and (2). Here, we presented participants with a natural audiovisual speech discourse while recording their EEG, and investigated low frequency activities profiles at the onsets of words either accompanied by a beat gesture or not. Following the temporal evolution of low frequency synchronizations provided evidences on when and how beats modulated the auditory processing of the affiliated word, complementing the results from the first ERP study.

2.3. Biau, E., Moris Fernandez, L., Holle, H., Avila, C., & Soto-Faraco, S. (Submitted). Spontaneous beat gestures as prosody: an asynchrony with speech affects language processing. *Neuroimage*

In the third article, we addressed the hypothesis (3). We combined the presentation of AV clips taken from a broadcasted discourse with fMRI neuroimaging to investigate the neural correlates of beat gestures. Beats may be part of the same language system with speech, providing visual prosody when aligned with spoken prosody during speech perception. First, we hypothesized that if the temporal alignment between beats apexes and pitch accents is broken by an asynchrony, then neural activations in language

related areas may be modulated as well. Based on previous fMRI studies (Marstaller & Burianova, 2014; Hubbard et al., 2009), we expected different BOLD responses particularly in the left Inferior frontal Gyrus (left IFG) and left Superior Temporal Sulcus/Gyrus (left STS/G) when beats were synchronized as compared to desynchronized with speech.

Second, we addressed whether the potential prosodic role of beats relies only on their emphasizing trajectories (velocity, directions and apexes) aligned with auditory envelop modulations or, whether beats engage a specialized mechanism because they convey additional communicative intentions of the speaker. To address this question, we added a manipulation in which we replaced the speaker's hands by moving discs that reproduced the original kinematics and spatio-temporal properties of beats. We hypothesized that simple emphasising spatiotemporal trajectories of arbitrary visual stimuli may not be enough to accomplish the same linguistic function that gestures have when combined with speech. At neural levels, we expected qualitatively distinct modulations of BOLD responses in the language related areas by an asynchrony between speech and beats, or speech and discs.

2. EXPERIMENTAL SECTION

2.1 Beats modulate early stages of audio processing during continuous speech perception

Biau, E., & Soto-Faraco, S. (2013). Beat gestures modulate auditory integration in speech perception. *Brain and Language*, 124(2), 143–52.

Biau E, Soto-Faraco S. Beat gestures modulate auditory integration in speech perception. Brain Lang. 2013 Feb;124(2): 143-52. DOI 10.1016/j.bandl.2012.10.008

2.2 Beats bear a predictive value within speech signal

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

Biau E, Torralba M, Fuentemilla L, de Diego Balaguer R, Soto-Faraco S. Speaker's hand gestures modulate speech perception through phase resetting of ongoing neural oscillations. Cortex. 2014; 68:76-85. DOI 10.1016/j.cortex.2014.11.018.

2.3 Beats convey communicative value and are perceived as linguistic visual information

Biau, E., Moris Fernandez, L., Holle, H., Avila, C., & Soto-Faraco, S. (Submitted). Spontaneous beat gestures as prosody: an asynchrony with speech affects language processing.

Spontaneous beat gestures as prosody: an asynchrony with speech affects language processing.

Emmanuel Biau[1], Luis Moris Fernandez[1], Henning Holle[3], César Avila[4] and Salvador Soto-Faraco[1,2]

1. *Center for Brain and Cognition (CBC), University Pompeu Fabra, Barcelona, Spain*
2. *Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain*
3. *Department of Psychology, University of Hull, UK.*
4. *Department of Psychology, Universitat Jaume I, Castelló de la Plana, Spain.*

Hand gestures as visual prosody: BOLD responses to audio-visual alignment are modulated by the communicative nature of the stimuli

Emmanuel Biau [a]
Luis Moris Fernandez [a]
Henning Holle [c]
César Avila [d]
Salvador Soto-Faraco [a, b]

[a] Multisensory Research Group, Center for Brain and Cognition, Universitat Pompeu Fabra, Barcelona, Spain.

[b] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.

[c] Department of Psychology, University of Hull, UK.

[d] Department of Psychology, Universitat Jaume I, Castelló de la Plana, Spain.

Corresponding author: Emmanuel Biau

Dept. de Tecnologies de la Informació i les Comunicacions
Universitat Pompeu Fabra
Roc Boronat, 138
08018 Barcelona
Spain

+34 691 752 040

emmanuel.biau@free.fr

ABSTRACT

During public addresses, speakers accompany their discourse with spontaneous hand gestures (beats) that are tightly synchronized with the prosodic contour of the discourse. It has been proposed that speech and beat gestures originate from a common linguistic process, with both speech envelope and beats serving to emphasize relevant information. In this study, we measured BOLD responses to a natural discourse where the speaker used beat gestures. We hypothesized that breaking the consistency between beats and prosody, by introducing an asynchrony between gesture apexes and pitch accents, has an impact on the activity of language-related brain areas sensitive to the integration of beat and speech information. In order to identify brain areas specifically involved in processing hand gestures with communicative intention, beat synchrony was evaluated against arbitrary visual cues bearing equivalent rhythmic and spatial properties compared to the gestures. Our results revealed that left MTG and IFG were specifically sensitive to speech synchronized with beats, compared to the control vision-speech pairing with discs. Interestingly, these areas seemed to exhibit opposing patterns of activity when the speaker's hands were replaced by discs bearing the same trajectories. Our results suggest that listeners confer beats a function of visual prosody, complementary to the prosodic structure of speech. We conclude that the emphasizing function of beat gestures in speech perception is instantiated through a specialized brain network sensitive to the communicative intent conveyed by a speaker with his/her hands.

Beat gestures; Audiovisual speech; Multisensory Integration; left MTG; fMRI.

1. INTRODUCTION

In everyday life, people communicate with each other in social contexts where speaker and listener share information through acoustic, as well as visual channels. Although the verbal utterance is sufficient to convey information between two persons (as it is well illustrated by phone conversations), most communicative interactions involve also visual information. Listeners have visual access to the speaker's lips, head, body posture and spontaneous hand gestures. Here we concentrate on the communicative impact of a certain type of cospeech gestures, which are the hand movements produced by the speaker while talking to someone. McNeill (1992) defined different categories of gestures according to their hand shape or relationship with speech. Subsequent studies showed that gestures modulate various levels of speech processing. By combining behavioral and physiological measures like event-related potentials (ERPs), many studies demonstrated for example that gestures describing an object or an action (i.e. iconic gestures) can alter semantic processing of speech (Kelly et al., 2004; Kelly et al., 2009; Wu & Coulson, 2010) or help disambiguate semantically complex sentences (Holle et al., 2007). These studies suggest that gestures provide additional visual information not present in the verbal modality, supporting the idea that both streams of information are in fact components of a common underlying language system (McNeill, 1992; Kelly, Creigh & Bartolotti, 2009).

The intrinsic relationship between gesture and speech processing was illustrated in fMRI studies that investigated the degree to which gesture and speech recruit similar brain areas. For instance, the Superior Temporal Sulcus (STS) and adjacent Middle and Superior Temporal Gyri (MTG/STG), which are well known to respond to audiovisual (AV) speech (Nath and Beauchamp, 2012; Calvert et al., 2000; Callan et al., 2004; Macaluso et al., 2004; Meyer et al., 2004; Campbell, 2008), were found to be sensitive to the semantic relationship and congruency between gestures and the spoken message (Marstaller & Burianova, 2014). Greater BOLD responses in the STS, inferior parietal lobule and precentral sulcus were found for the perception of spoken sentences accompanied by corresponding iconic gestures, as compared to meaningless

movements or auditory-only versions (Holle et al., 2010; Holle et al., 2008). Willems et al, (2009) also found greater activations in the left STS/MTG when spoken sentences were presented with simultaneous pantomimes (i.e. gestures depicting objects or action that can be understood even without speech) whose shape matched the verb of the utterance in meaning, as compared to incongruent pantomimes. Additionally, the left Inferior Frontal Gyrus (IFG) has been often found to respond to the manipulation of the semantic relationship between gesture and speech (Marstaller & Burianova, 2014; Willems et al., 2009; Willems et al., 2007), suggesting that this area plays a role in the integration of both streams of information to support sentence comprehension (Glaser et al., 2013; Uchiyama et al., 2008; Willems et al., 2007; Hagoort, 2005). In other words, studies exploring the contribution of gestures to semantic integration during speech comprehension have established the implication of a fronto-temporal network of language-related areas, including the STS/G and the left IFG (for more details, see also Dick et al., 2014).

Although very relevant, these studies focused on the neural correlates of hand gestures conveying semantic content, leaving aside the function of gestures as prosodic markers of speech. Additionally, in these past studies, the spoken as well as gestural stimuli were realized in a highly constrained context. Participants were typically presented with short sentences containing an isolated gesture corresponding to a critical word; a context that is far from ecological in production and perception. Therefore, so far these studies did not help understanding the perception of gestures as they are normally produced in continuous, natural social interactions. If one considers gestures and speech as two complementary sides of a common language system (McNeill, 1992; Kelly, Creigh and Bartolotti, 2009), the continuous flow of both visual and audio streams might need to be maintained for the system to remain fully functional (Hubbard et al., 2009; Biau & Soto-Faraco, 2013).

In the present study, we address the neural correlates of the prosodic (rhythmic) function of co-speech gestures. We were interested in spontaneous gestures with less sophisticated hand form (as they bear prosodic but no semantic information) which are embedded in a continuous, natural speech context. We investigated the potential role of gestures in the analysis of the

speaker's narrative structure from the listener's point of view. We focused on the most frequent type of gestures produced in natural political discourse, the so-called beats (McNeill, 1992). Beats are rapid biphasic flicks of the hand (with no semantic content in their shape) that serve to highlight relevant information and structure the narrative discourse (McNeill, 1992; Casanto and Jasmin, 2010).

The production of a co-occurring beat gesture has been shown to influence the prominence of affiliate words in production by modulating the acoustic properties of the accentuated syllable (Krahmer & Swerts, 2007), and to improve a listener's word retrieval in memory tasks (So et al., 2012). Recent ERP studies have shown that beats can effectively modulate the processing of affiliate words. For instance, Biau & Soto-Faraco (2013) presented an entire natural audiovisual discourse to observers while recording their EEG signal and found that beat gestures modulated early ERPs time-locked to affiliate words, suggesting an early attentional effect of beat gestures. Wang & Chu (2013) showed that beats facilitated semantic processing by reducing the amplitude of the N400 component when synchronized with a critical word in sentences. Additionally, an fMRI study by Hubbard et al. (2009) investigated the neural correlates of beats using naturalistic stimuli. In this study, observers watched a speaker producing spontaneous beats while speaking. unaware of the purpose of the experiment. The authors reported greater activations in the left STG/S in response to speech when it was accompanied by beat gestures as compared to when it was presented with unrelated sign language gestures (Hubbard et al., 2009). The authors also reported greater BOLD responses in the bilateral posterior STG/S, including the Planum Temporale (PT) when subjects listened to speech accompanied by beats relative to listening to speech accompanied by a still body. Using beats from an actual fragment of continuous discourse ensured that gestures were produced in a legitimate context and frequency, instead of being isolated or placed in out-of-context sentences. In addition, using spontaneous conditions of speech production ensured that the temporal relationship between the continuous beats stream and the rhythm of speech was maintained as in natural language conversation (Biau et al., *in press*).

Despite their simple appearance, linguistic hand beats may convey visual aspects of the speaker's conception of his discourse and language-related characteristics. Here, we address whether temporal characteristics of the speaker's beats may impact continuous speech processing by the listener. This question is relevant because it is widely accepted that beat gestures may play a role in prosodic processing (see for example Guellaï, Langus & Nespor, 2014). Indeed, the functional phases of beat gestures - the brief maximum extension moments of the movement (i.e. the "apex") - was consistently reported to be temporally aligned with auditory prosody and particularly with the pitch accents of the corresponding spoken word (McNeill, 1992; Krahmer & Swerts, 2007; Treffner and al., 2008). For instance Yasinnik, Renwick and Shattuck-Hufnagel (2004) reported a consistent overlap of gesture apex and pitch accent when labelling audio and visual streams independently across several speakers. Leonard and Cummins (2010) reported that participants could detect asynchronies as small as 200ms when pitch accentuations lagged with respect to gesture apexes. From the listener's point of view, this association of beat gestures with the prosody of the spoken message suggested that they might convey relevant information for syntactic parsing. It is well known that prosody and syntax interact during comprehension (Eckstein & Friederici, 2005, 2006). Recently, Guellaï et al. (2014) showed that a mismatch between prosody and beats increased the difficulty to comprehend syntactically ambiguous sentences. At a neural level, Holle et al. (2012) found that one isolated beat can modulate a component of the Event Related Brain Potential (ERP) known to reflect syntactic analysis, depending on the beat's precise alignment with the accentuated syllable of the relevant noun within syntactically ambiguous sentences.

Scope of the present study

In the present study, we hypothesized that beat gestures are produced as an integral part of the language system and therefore, can convey linguistic information to the perceiver by means of providing visual prosody when aligned with the spoken prosodic contour during speech perception. If this is true, the fronto-temporal language-related network described in previous fMRI studies on

co-speech gestures (at least left STS/G and left IFG) may be sensitive to a breach in the temporal synchrony between beats with respect to their speech affiliates (Marstaller & Burianova, 2014; Hubbard et al., 2009). To test this hypothesis, we used fMRI while participants were presented with video clips in which the video was either synchronized with the audio track or lagged. With this manipulation, we assumed that when beat's apexes fall out of synchrony with their affiliated speech accentuations, their highlighting function is cancelled. Importantly, we addressed whether this potential prosodic function of beat gestures when aligned with pitch accents relates to a generic mechanism of visual emphasis or, alternatively, whether beats engage a specialized mechanism. Suggestive of the latter account, in the aforementioned study by Holle et al. (2012), an influence of visual emphasis on a syntax-associated ERP component was not found when the beats were replaced with a disc following the same trajectory in the visual display. Based on this result, Holle et al. concluded that beat gestures bear additional communicative intention influencing language comprehension that distinguishes them from simple visual emphasis. Besides visual prosody, beats may convey speaker's emotions and intentions, whereas simple discs do not. Here, we hypothesized that the simple emphasis conferred by the spatiotemporal trajectory of arbitrary visual stimuli may differ from the linguistic function that gestures have when combined with speech (i.e. when beat emphasis is synchronized with the speech prosody). If beat gestures effectively engage language processing because of their value in communicative intention, then one should expect disparate effects of audio-visual asynchrony for beat gestures as compare to other visual cues. To test this we created a design in which we replaced the speaker's hands by moving discs that reproduced the original kinematics and spatio-temporal properties of beat gestures.

We set up a 2x2 design with the factors AV synchrony (synchronous or asynchronous) and visual information (hands or discs) to test how the temporal alignment affects the integration of speech with either type of visual information. The interaction between synchrony and visual information is of particular interest because it allows isolating brain areas in which the impact of asynchrony depends on which kind of visual information (beats or discs)

accompanies speech prosody. If the hypothesis that beats confer a special communicative value to the spoken message is true, then brain areas related to this specialized integration should exhibit greater response to the synchrony manipulation when speech is presented with beats compared to moving discs. Thus, this study will concentrate on brain areas where such an interaction arises. According to prior literature, these areas might (though not exclusively) correspond to the ones previously shown to be sensitive to gesture-speech integration, such as the left STS/G but also the left IFG (Holle et al., 2007; Willems et al., 2007; Hubbard et al., 2009; Holle et al., 2010; Marstaller & Burianova, 2014).

2. MATERIAL AND METHODS

2.1 Participants

Nineteen native speakers of Spanish (12 female, age range 19-29) took part in the current study. All participants were right-handed with normal auditory acuity as well as normal or corrected-to-normal vision. Participants gave informed consent prior to participation in the experiment and the study was approved by the University's ethics committee. Due to a technical problem, two participants could not listen to the speech stream during fMRI data acquisition and were therefore excluded from the statistical analysis. Thus, data from 17 participants (12 females, age range: 22.4 ± 2.4 years old) were included in the imaging analysis.

2.2 Material and stimuli

We extracted 44 video clips (18 s duration each) from a political discourse of the former Spanish President Luis Rodríguez Zapatero, recorded at the palace of La Moncloa and available on the official website (*Balance de la acción de Gobierno en 2010*, 12-30-2010; http://www.lamoncloa.gob.es). During the whole public address, the speaker stood behind a lectern, with the upper part of the body in full sight. The video clips were edited using Adobe Premiere Pro CS3.

We visually inspected the entire discourse to select relevant segments of speech, containing only beats and cohesive gestures (series of beats that link successive points to a common concept) according to McNeill's definition. Clear iconic gestures were not found but as gesture categories sit along a continuum with fuzzy boundaries, some gestures may fall into multiple categories. Therefore one cannot be absolutely certain that our stimuli never included a minimum of semantic content in the hand shape. However, hand movements always conformed to McNeill's definition of beat gestures. To avoid abrupt onsets and offsets, we introduced 1 second audio-visual fade-in and –out at the beginning and end of each clip (respectively). In all the AV clips, the head of the speaker was masked with a superimposed ellipse-shaped patch in order to remove any facial information, such as lips or eyebrow movements, as well as head movements. After editing, videos were exported using the following parameters: video resolution 960x720, 25 fps compressor Indeo video 5.10, AVI format; audio sample rate 48 kHz 16 bits Mono. As explained below, we created four different versions for each video, corresponding to the four conditions of our experimental design: Beat Synchronous (Bs), Beat Asynchronous (Ba), Disc Synchronous (Ds) and Disc Asynchronous (Ds) (Fig. 1).

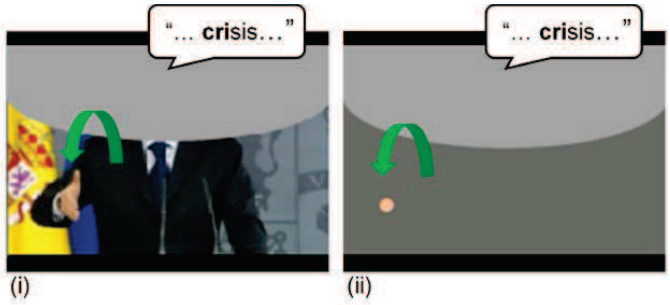

Figure 1. Screenshots from (i) Beat and (ii) Disc conditions. Audio and video streams were either synchronized (Bs and Ds conditions) or desynchronized (audio lagged video by 32 frames, corresponding to 800 ms) with respect to audio in the Ba and Da conditions). Green arrow illustrates the trajectory of a beat gesture and the corresponding disc. The apex of the movement coincided in this case with the Spanish word 'crisis'.

Beat conditions: We selected 44 segments (18s each, 450 frames) of the discourse in which the speaker naturally produced spontaneous beats (McNeill, 1992). For each clip, the speaker produced a minimum of 8 beats within the 18

s (mean number of gestures per clip: 12.8 ± 4.2). To create the Beat-Synchronous condition, audio and visual information remained synchronized as in the original discourse, with the speaker's hands fully visible (beat synchrony, Bs). For the beat asynchrony (Ba) condition, audio and visual information were desynchronized by inserting a lag of 800 ms (32 frames), leading to speech preceding beat gestures.

Disc conditions: To create the disc conditions, the video was removed and the hands were replaced by two discs that followed the hand trajectories of the original clips. We defined the junction between the index and the thumb as the reference point of both hands. We used *Skin Color Estimation Application* and *ELAN* software to detect pixel coordinates of hands frame-by-frame in each Beat video (http://tla.mpi.nl/tools/tla-tools/elan; Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands; Wittenburg et al., 2006). Reference point coordinates were reviewed and corrected were necessary for both hands using custom-made scripts for Matlab (MATLAB Release 2012b, The MathWorks, Inc., Natick, Massachusetts, United States). The two discs representing the hands had a 40 pixel diameter size and were flesh-colored (Red, Green, Blue color values: 246, 187 and 146) at their corresponding reference point. The background color was set to the average value of a still frame of the speaker (Red Green Blue Value: 110, 114, and 104). We then created a synchronized (Disc Synchrony, Ds) and a desynchronized (Disc Asynchrony, Da) condition following the same process as in the beat condition.

Target videos: To ensure that stimuli were attended, participants performed an auditory detection task. For this, we used two clips from each experimental condition to create 8 targets. For each target video, the fundamental pitch of the original audio tracks was artificially shifted up three semitones (high pitch) for one syllable using Adobe's PitchShift filter while the intensity remained the same. In total, each participant was presented with 36 experimental and 8 target videos.

2.3 Procedure and Instructions

Participants were presented with 44 trials using E-Prime2 software. The order of trials was pseudo-randomized to avoid direct repetition of experimental conditions. Each trial consisted of a fixation cross with variable duration (from 7.5 to 8.5 seconds in steps of 0.25 seconds, uniformly distributed) followed by a video clip. The next trial began automatically after the end of the preceding video. A total of four experimental lists were created, counterbalanced for the four experimental conditions. Each participant saw one of the four lists.

Participants were instructed to perform an auditory detection task and press a button of the fMRI-compatible controller as soon as they detected an artificial pitch change in the voice of the speaker. The hand holding the controller (left or right hand) was counterbalanced across participants (even though target trials were not included in the statistical analysis). Participants were also instructed to always look at the screen during the whole experiment as if they were watching television. Before the fMRI acquisition, participants performed a rapid training with an extra target video presented in both Bs and Ds conditions as an example of artificial pitch change. After the scanning session, participants were given a questionnaire, asking 1) Did you perceive any asynchrony between video and speech during the experiment? 2) What could the moving discs represent? This questionnaire served to ensure that participants correctly attended to all videos. More importantly, it allowed us to evaluate if they could perceive the asynchrony between video and speech.

2.4 fMRI acquisition

Imaging was performed in a single session on a 1.5 T Siemens scanner. We first acquired a high-resolution T1-weigthed structural image (GR\IR TR=2200ms, TE=3.79ms, FA=15º, 256 x 256 x 160, 1mm isotropic voxel size). Functional data was acquired in a single run consisting of 610 Gradient Echo EPI functional volumes (TE = 50 ms, TR = 2000 ms) not specifically co-planar with the Anterior Commisure – Posterior Commisure line, acquired in an interleaved ascending order using a 64× 64 acquisition matrix with a FOV = 224. Voxel size was 3.5 x 3.5 x 3.5 mm with a 0.6 mm gap between slices,

covering 94.3 mm in the Z axis.. The functional volumes were placed attempting to cover the whole brain in 23 axial slices. The first four volumes were discarded to allow for stabilization of longitudinal magnetization.

2.5 Imaging data analysing

FMRI data were analyzed using SPM12b (www.fil.ion.ucl.ac.uk/spm) and Matlab R2013b (MathWorks).

2.5.1. Preprocessing

Standard spatial preprocessing was performed for all participants using the following steps: Horizontal AC-PC reorientation; realignment and unwarp using the first functional volume as reference, a least squares cost function, a rigid body transformation (6 degrees of freedom) and a 2^{nd} degree B-spline for interpolation, creating in the process the estimated translations and rotations occurred during the acquisition; slice timing correction using the middle slice as reference using SPM8's Fourier phase shift interpolation; coregistration of the structural image to the mean functional image using a normalized mutual information cost function and a rigid body transformation. The image was then normalized into the Montreal Neurological Institute (MNI) space (Voxel size was changed during normalization to isotropic 3.5 × 3.5 × 3.5 mm and interpolation was done using a 4^{th} B-spline degree). Functional data was smoothed using an 8-mm full width half-maximum Gaussian kernel to increase signal to noise ratio and reduce inter subject localization variability. To add an extra quality control to the movement in participants, we used the Artifact Detection tools (ART) (http://www.nitrc.org/projects/artifact_detect/) with which the composite movement was calculated. This provides a single measure that comprises the movement due to rotation and translation between volumes. All volumes with a composite movement of more than 0.5 mm or more than 9 standard deviations away from the global mean signal of the session were considered as outliers (On average, 1.4% of the volumes per participant were detected as outliers). One regressor per outlier was added at the first level to discard any possible influence of these volumes in the final analysis.

2.5.2. fMRI analysis

The time series for each participant were high-pass filtered at 128 s and pre-whitened by means of an autoregressive model AR(1). At the first level (subject-specific) analysis, box-car regressors modelling the occurrence of the four conditions of interest (Bs, Ba, Ds and Da) and a fifth regressor for trials containing a target, all modelled as 18s blocks, were convolved with the standard SPM12b hemodynamic response function. Additionally, several regressors of no interest were included, including the six movement regressors provided by SPM during the realign process, the extra composite movement regressor calculated with ART and one regressor for each of the volumes considered as outliers. The resulting general linear model produced an image estimating the effect size of the response induced by each of the conditions of interest. The images from the first level were used for the planned critical contrasts in a second level analysis (inter-subject). At the second (inter-subject) level, these images were entered into a random effects factorial design with five levels, corresponding to the four critical conditions, plus an additional subject constant to account for non-condition-specific inter-subject variance. Correction for non-sphericity (Friston et al., 2002) was used to account for possible differences in error variance across conditions and any non-independent error terms for the repeated measures. Statistical images were assessed for cluster-wise significance using a cluster-defining threshold of $p<0.001$. The 0.05 Family-wise error correction critical cluster size was 31 voxels and was determined using random field theory (Data smoothing FWHM: 11.4mm, 11.2mm, 11.3 mm. Resel Count: 749.2), considering the whole brain as a volume of interest. Contrasts vectors assessing the two main effects and the interaction were used. Although the whole interaction statistical parametric map is presented, the discussion is limited to the clusters that showed an effect of Beat gestures compared to Discs (Bs+Ba > Ds+Da), as our main interest is focused on the parts of the brain that are involved in beat processing (for unmasked results and additional contrasts, please see supplementary online materials).To achieve this, we masked the interaction contrast, corrected as explained above, with the Beat > Discs contrast (p-threshold (unc.) <0.05). MNI

coordinates were classified as belonging to a particular anatomical region using the SPM Anatomy Toolbox (Eickhoff et al., 2005).

3. RESULTS

3.1 Behavioral results

Participants correctly detected pitch deviation targets on 65.4% ± 31.7 % of the target trials and gave False Alarm (FA) responses only on 7.0% ± 13.6 % of the non-target trials.

3.2 Post-scanning questionnaire

When asked, after the scanning session, whether they perceived any asynchrony between video and speech during the experiment, 12 participants responded “yes”; 3 participants responded “yes, but not in the disc condition” and 2 participants responded “no”. With respect to the second question (“What could the moving discs represent?”), all participants responded “the hand of the speaker. This suggests that the asynchrony between beats and speech was noticeable, even though facial information was removed from videos. Furthermore, this consistent response confirmed that the spatiotemporal characteristics of disc movements successfully mimicked the hand trajectories in the Disc conditions. Both the behavioural and post-scanning questionnaire results suggest that participants were attentive to the AV stimuli.

3.3 fMRI results

3.3.1 Differential effect of AV synchrony depending on visual information

The first contrast of interest concerns the interaction between synchrony and visual information [(Bs-Ba) – (Ds-Da)]. This contrast is of particular interest as it highlights the brain areas where the impact of synchrony depends on which kind of visual information (beats or discs) accompanies speech. We studied this

interaction in the areas that showed an effect of Beat > Disc (uncorrected mask p<0.05), as explained in the methods section (see Table 1). This restricts our analysis to areas that are related to beat processing. The results revealed a significant interaction in BOLD responses in two different clusters of the left Middle Temporal Gyrus and Superior Temporal Sulcus (MTG/STS), one more posterior and one more anterior (respectively, pMTG and aMTG/STS). Additionally, significant interactions in left IFG and left occipital cortex (Brodmann area 18) were observed.

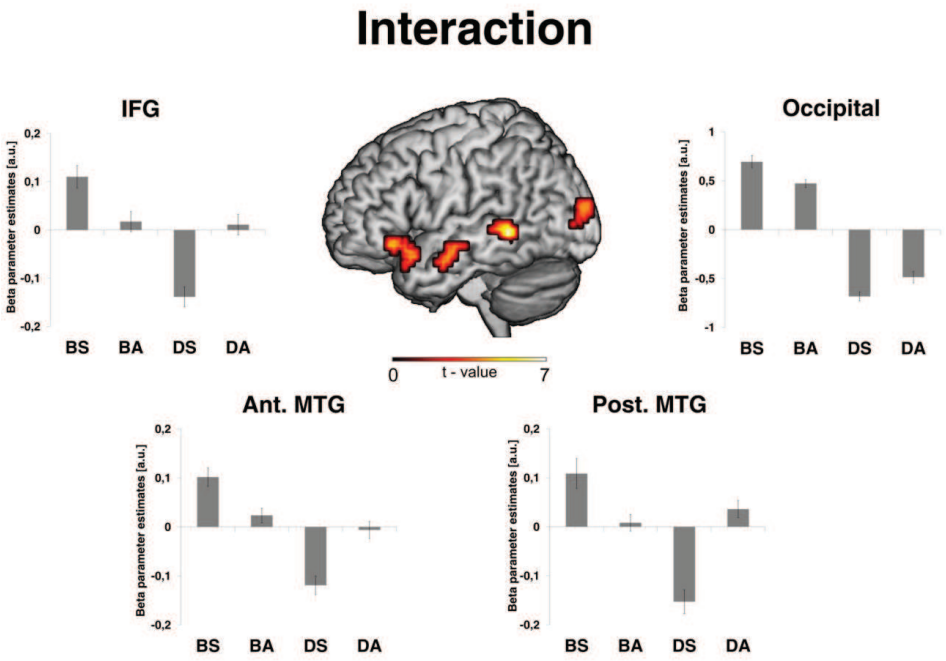

Figure 2. Interaction contrast [(Bs- Ba) – (Ds – Da) inclusively masked with the main effect of Beat (Bs+Da) compared to Disc (Ds+Da) using a p<0.05 cluster-corrected threshold with a minimum cluster size k = 31 and rendered on a 3D brain surface in MNI space (Left hemisphere). Error bars show 1 S.E.M of parameter estimates. IFG: Inferior frontal gyrus (-41 32 -11); Ant.MTG: anterior Middle temporal gyrus (-52 -7 -18); Post. MTG: posterior MTG (-59 -46 -4); Occipital (-20 -95 14).

These results suggest that audio-visual synchrony differentially affects speech integration, depending on the content of visual information. In particular, speech-gesture synchrony seems to recruit left-hemisphere brain areas

preferentially, as compared to other visual cues which share the same spatio temporal properties but are arbitrary. Please note that following up on the pattern of simple main effects in the areas relevant for this interaction would involve post-hoc analysis whose interpretation, according to some authors, would incur in circularity (Kriegeskorte et al., 2009). Thus, albeit their pattern follows an expected trend (see Figure 2; see the significance of post-hoc simple main effects in the Supplementary Material), we will refrain from interpreting them. Nevertheless, it is worth noting that the areas which display this pattern (MTG, IFG and Occipital cortex in the left hemisphere) and the directionality of the numerical effects of beat synchrony are well in line with previous studies investigating gesture perception (Hubbard et al., 2009; Willems et al., 2009; Skipper et al., 2007; Holle et al., 2008, 2010), which further reassures the interpretation of these activations.

3.3.2 Effect of type of visual information within temporal synchrony

Looking at the main effect of type of visual cue within the synchronous conditions can reveal differences arising from the type of visual stimulus. The contrast Beat Synchronous > Disc Synchronous revealed a greater BOLD response in various brain areas when speech was accompanied by synchronized beats (Bs), relative to synchronized discs (Ds) (see figure 3 and table 1). Not surprisingly, the greatest difference was observed in the occipital cortex likely due to a pure difference in visual information between conditions. The contrast also revealed differences in beyond visual brain areas, such as a significantly greater BOLD activity in the left MTG/STS, as well as in the left Inferior frontal Gyrus (left IFG) and left hippocampus. The contrast Ds>Bs revealed greater BOLD activity when speech was accompanied by synchronous discs rather than synchronous hand beats in the Superior Parietal areas bilaterally and right Angular Gyrus (see figure 3 and table 1).

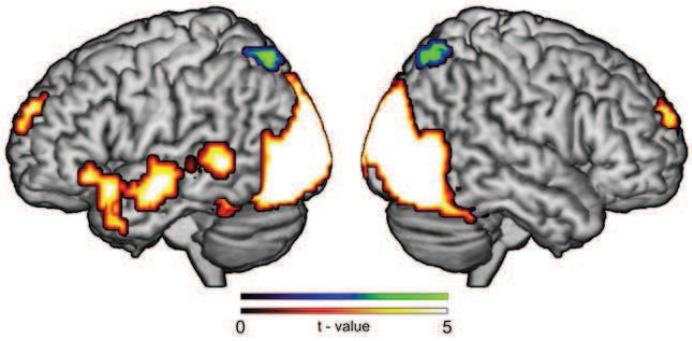

Figure 3. Main effect of Beat Synchronous (Bs) compared to Disc Synchronous (Ds). Statistical maps are thresholded at *P*-uncorrected <0.001 with a minimum cluster size k = 31 and rendered on a 3D brain surface in MNI space. From left to right: left hemisphere, right hemisphere and an axial cut at z=0. Hot colours indicate Bs > Ds. Cold colours indicate Ds> Bs.

3.3.3 Effect of asynchrony between beat gestures and speech

The contrasts involving the comparisons Bs>Ba and Ba>Bs, restricted within the beat gesture conditions, revealed no main effect of synchrony, when performed at the whole brain level. Note that this particular result deviates from Hubbard et al. (2009), who reported an effect of synchrony in the left STS/G area. However, it must be mentioned that in Hubbard's study not only the actual synchrony, but also the nature of the gestures themselves was different between the synchronous and asynchronous condition (beats vs. ASL gestures in the control condition, respectively). In any case, our result implies that despite the BOLD responses for synchronous gestures tend to be larger than the BOLD responses for asynchronous gestures in the areas of significant interaction (as revealed in the interaction analysis), this effect can only be interpreted safely relative to the responses of these areas to the disc synchrony/asynchrony condition.

Hemisphere	Region	Corrected Cluster P-Value	Number of Voxels[a]	Z - Score	Coordinates (mm)[b] x	y	z

Interaction [(Bs-Ba) – (Ds-Da)] masked with Beat > Disc (mask p-value <0.05)

L	Middle Temporal Gyrus	0,043	32	5,93	-59	-46	-4
L	Inferior frontal gyrus	0,048	31	4,36	-41	32	-11
L	Temporal Pole			4,35	-45	14	-18
L	Middle Temporal Gyrus	0,048	31	4,20	-52	-7	-18
L	Middle Temporal Gyrus			4,10	-59	-11	-14
L	Middle Temporal Gyrus			4,09	-59	-4	-21
L	Middle Occipital	0,039	33	4,04	-20	-95	14
L	Inferior Occipital			3,38	-31	-88	4
Beat Synchronous > Disc Synchronous							
R	Lingual Gyrus	0,000	3080	Inf	8	-88	4
L	Cuneus			Inf	-10	-98	18
L	Calcarine			Inf	-3	-88	-4
L	Middle Temporal Gyrus	0,000	151	5,22	-62	-11	-14
L	Temporal Pole			4,75	-48	18	-14
L	Inferior Frontal Gyrus			4,33	-41	28	-11
L	Thalamus	0,006	52	5,20	-24	-28	0
L	Middle Temporal Gyrus	0,001	75	4,90	-55	-46	0
L	Middle Temporal Gyrus			3,93	-48	-32	0
Disc Synchronous > Beat Synchronous							
L	Superior Parietal	0,006	50	4,75	-16	-70	56
R	Superior Parietal	0,009	47	3,73	22	-66	59
	Angular Gyrus			3,49	22	-56	49
	Superior Parietal			3,40	15	-59	63
Beat Synchronous > Beat Asynchronous							
No significantly activate regions							
Beat Asynchronous > Beat Synchronous							
No significantly activate regions							

Table 1.[a] Number of voxels exceeding a voxel-height threshold of $p < 0.001$ using a $p < 0.05$ cluster-extend FWE correction. [b] First three maximum peaks more than 8 mm apart are reported for each cluster.

4. DISCUSSION

In the present study, we investigated the neural correlates of spontaneous beat gestures accompanying continuous natural discourse. Based on previous reports (McNeill, 1992; Yasinnik et al., 2004; Guellaï et al., 2014; Biau et al., *in press*), we hypothesized that beats act as a visual counterpart of prosody. If this is the case, then breaking up the consistency between beat apexes and speech prosody may affect speech processing. At the neural level, we hypothesized that if beats are treated as linguistically relevant information, then activations in language-related areas, including left STS/G and IFG, may reflect the sensitivity to an asynchrony between visual and audio streams (Holle et al., 2008; Willems et al., 2007; Hubbard et al., 2009; Holle et al., 2010;

Marstaller & Burianova, 2014). Critically, we also addressed whether mere audio-visual spatio-temporal synchrony is sufficient to affect language areas, or whether beats convey additional communicative aspects above and beyond arbitrary visual cues (discs) sharing the same spatiotemporal properties (Holle et al., 2012). We hypothesized that beats translate speaker intentions to emphasize relevant segments of speech, which are available for listeners during speech perception (So et al., 2012; Casasanto & Jasmin, 2009). If this is the case, the effect of audio-visual synchrony in previously known audio-visual areas such as left MTG and IFG should be qualitatively different for beats as compared to discs (i.e., an interaction between synchrony and visual Information should occur). Indeed, we found the interaction that indicates that the temporal asynchrony of beats with speech prosody has a differential impact on neural activations in these language related areas, compared to other kinds of visual information. The tendencies in the pattern of the interaction contrasts suggest greater activations when beats and speech were presented in synchrony as compared to asynchrony. In contrast, the opposite pattern was observed when speech was accompanied by discs sharing the same spatio-temporal properties as the original hand gestures. Based on this significant interaction pattern, we interpret that, in addition to their emphasizing trajectory, beats also convey communicative aspects that simple discs are arguably lacking.

One surprising finding of our study is that the effect of synchrony for beats (i.e., greater activity for synchronous as compared to asynchronous beats in left IFG and MTG) was not simply absent for the moving discs, but actually tended to be reversed (i.e., trend for reduced activity for synchronous as compared to asynchronous discs in left IFG and MTG). When interpreting this cross-over interaction, it is also useful to take into account whether the neural response in these areas represents an activation or deactivation, relative to the implicit fixation cross baseline (see parameter estimates in Fig. 2). Relative to this fixation cross baseline, only speech accompanied by synchronous beats elicited activation in IFG, aMTG and pMTG. This is consistent with the idea that IFG and posterior temporal lobe are crucially involved in comprehending co-speech gestures (Holle et al., 2008, 2010, Willems et al., 2007, 2009). In contrast, a visual emphasis cue presented in asynchrony with speech

(regardless of whether emphasis consisted of beats or moving discs) did not activate these areas, which may reflect that temporally incongruent AV stimuli are less likely to be integrated and may even cause suppression in multisensory areas (Noesselt et al., 2007). Interestingly, processing speech accompanied by temporally congruent discs elicited a reduction of activity in IFG, aMTG and pMTG, relative to fixation baseline. Such a deactivation could possibly reflect a phasic inhibitory influence onto IFG, aMTG and pMTG whenever speech is accompanied by temporally congruous but unfamiliar visual emphasis cues, such as moving discs. An influence of stimulus familiarity on AV integration in the temporal lobe has been demonstrated before (Hein et al., 2007) and may extend to unfamiliar speech-accompanying visual emphasis cues, such as moving discs.

Our results are in line with previous fMRI studies which investigated neural correlates of iconic gestures (Holle et al., 2010; Holle et al., 2008; Willems et al., 2009; Willems et al., 2007). Particularly, one previous fMRI addressed natural hand beats co-occurring with continuous speech (Hubbard et al., 2009) and reported a greater engagement of the STS compared to speech alone, an area comparable to the one found in the present study. The authors also reported greater BOLD activation in the left STS/G when speech was presented with the corresponding beat as compared to when presented with unrelated hand movements. Please note that this comparison does not allow one to infer whether the difference in left STS activation was produced by the lack of synchrony between control gestures and speech, the lack of communicative value of control gestures, or an unknown combination of the two. When Hubbard et al. compared speech accompanying beats to beats presented without speech, no difference was observed, suggesting that the modulations in the left STS/G reflect not only processing of biological movement but also integration of speech with the synchronized beat gestures. Indeed, the STS is sensitive to various types of cross-modal correspondence including AV speech (sound-lip correspondence) in various previous studies (Nath and Beauchamp, 2012; Calvert et al., 2000; Callan et al., 2004; Macaluso et al., 2004; Meyer et al., 2004).

In the present study, the interaction contrast suggests that BOLD response in the left MTG was greater when speech was accompanied by beats as compared to discs (regardless of whether they were synchronized or not with speech). At first glance, the greater response to stimuli containing beats in occipital areas compared to those with discs may reflect a pure bottom-up effect of richness of visual information (Figure 3). However, the interaction (Figure 2) revealed also that the significant difference of BOLD activity in the visual areas between beat and disc were dramatically reduced under asynchronous presentations. This suggests that mere physical differences between beats and discs conditions were not sufficient to explain their respective impact of asynchrony in language-related areas. The difference between beats and discs might bring about more profound consequences. For example, in a previous ERP study, Holle et al. (2012) showed that a beat modulated the P600 component reflecting syntactic parsing, whereas a disc following the equivalent trajectory did not. The authors suggested that the lack of communicative intention may explain the failure of simple discs to affect the neural correlates of syntactic parsing. Here, the significant simple contrast Bs>Ds supports this claim as it revealed greater activations not only in the occipital areas (certainly due to differences of visual information), but also in the left MTG and left IFG areas. Indirectly, this result also converges toward the idea a differential response to synchrony for using discs that are not functionally associated with speech as part of a common language system.

According to the effect of interaction on the neural activations, it seems that the MTG responded to some additional language-related aspects associated with beat gestures during speech perception. Previous behavioral studies suggested that some implicit pragmatic and intentional information from the speaker could be extracted from beats, and influence speech encoding. For example, So et al, (2012) showed that adult observers managed to remember more words from a spoken list when the words had previously been accompanied by a beat gesture. As this memory improvement was not found in children, the authors concluded that beat gestures conveyed communicative information but the effect was functionally dependent on experiencing social interactions during development (McNeill, 1992). For example, listeners learn to interpret the speaker's intention to underline relevant information with a beat

through social experience. This association of communicative aspects between beats and pitch accentuations was highlighted by Krahmer and Swerts (2007) who showed that listeners perceived words as more salient when accompanied with a beat gesture compared to same words presented in isolation. What is often missing in these studies is whether the value of gestures and their integration of speech simply depended on the general salience of the stimulus, or whether co-speech gestures engaged a more specialized system. Although the listeners in the present study could associate moving discs with movements of the hands and participants were able to detect an asynchrony between discs and speech, synchronized gestures and synchronized discs elicited qualitatively distinct patterns of brain activation (see contrast Bs>Ds). This suggests that during perception listeners distinguished visual information functional related to some aspect of speech (beats) from arbitrary visual cues (discs). Here, this information may require additional processes reflected by the differences of activations in the MTG between beats and discs conditions.

In addition to the above explanation, the possible linguistic aspects engaged when beats are present may be directly related to human movement understanding and body postures, over and above to their interaction with speech. The STS was found to respond to point-light representations of biological movements (Grossman et al., 2004; Pelphrey et al., 2004), actions executed by humans (Thioux et al., 2008) and social visual cues (for reviews, see Nummenmaa & Calder, 2009; Allison, Puce & McCarthy, 2000). Herrington et al, (2009) showed that the posterior STS was significantly more activated for trials in which participants perceived human point-light representations of actions compared to non-human movements. In the present study, the discs did not clearly represent a human form but clearly mimicked the trajectories described by hands during speech. In reference to the present study, listeners could have associated discs trajectories with hands (as they identified in the post-task questionnaire). Yet, whatever aspect of biological motion engaged by left MTG activations in the disc conditions, it was more strongly expressed during beat conditions. Please note, however, that this possible perceptual difference between beat gestures and discs in biological motion cannot explain the whole pattern of results we found in the left MTG, because the interaction

term [(Bs – Ba) – (Ds – Da)] effectively controls for the different amounts of biological movement in the beat and disc conditions.

The present results also revealed an interaction between synchrony and visual information effects in the left IFG. Several fMRI studies have showed that the left IFG is sensitive to the semantic relationship between gesture and corresponding speech (Skipper et al., 2007; Willems et al., 2007; Willems et al., 2009; Dick et al., 2009) and may be engaged in the unification of visual (gestures) and audio (speech) complementary streams to facilitate comprehension (Willems et al., 2007; Hagoort, 2005). Recently, a meta-analysis investigating the neural correlates shared between different types of gestures reported a common engagement of the left IFG during the perception of speech accompanied with gestures as compared to a still body (Marstaller & Burianova, 2014). However, beat gestures do not convey semantic content, therefore the IFG responses observed in the present study cannot be explained in terms of semantic integration. Beyond meaning integration, the left IFG was also shown to be involved in the process of syntactic analysis during sentence comprehension (Glaser et al., 2013; Meyer et al., 2012; Obleser et al., 2011; Uchiyama et al., 2008). As beats play a role in syntactic parsing (Holle et al., 2012), our results might correspond to an engagement of this area in the integration of beat information toward the parsing of the spoken stream, as compared to moving discs. When beats were delayed (Ba condition), their apexes felt out from synchrony with pitch accents and likely out of the time window of gesture-speech integration, potentially affecting the AV speech processing load (Habets et al., 2011; Obermeier et al., 2011; Obermeier & Gunter, 2014).

It is worth noting that the simple main effect of synchrony for beat stimuli (contrast Bs vs Ba) in left MTG, IFG and occipital cortex did not reach significance in the whole brain analysis, but is suggested by the patterns of activations in the interaction contrasts following up on the interaction. Yet, the interpretability of post-hoc simple main effects restricted to the interaction areas is controversial, and we have chosen not to include it in the main text (see, Supplementary Materials, for completeness). In consequence, the interpretation

of synchrony effects for beat gestures must be for now linked to its effects relative to the disc condition. In other words, the disc synchrony manipulation can be seen as a baseline for the beat-synchrony manipulation. Yet, if we go by the results of previous studies, and extant knowledge the neural correlates of speech, we feel safe in interpreting this pattern in line with the results of the interaction that suggested a difference between synchronous and asynchronous beat conditions (see Figure 2). Note, for example that a similar effect of AV synchrony involving gestures in the left STG/S was reported in Hubbard et al. (2009). In their study, however, as mentioned earlier, Hubbard et al. used unrelated sign language movements as a control condition, which not only constitute a more dramatic asynchrony manipulation altogether (as speech and gestures had completely different rhythms), but also changed the very nature of the visual stimuli from the synchronous to the asynchronous condition. Here, we have looked at these two effects (confounded in Hubbard) separately, and therefore it is not surprising that their individual neural correlates are more subtle. That is, in the present study, although delayed with respect to speech, the rhythm of beats was maintained and might still be associable with the global speech envelope. This may have diminished the detrimental impact of desynchronized gestures on a listener's perception. This may also explain why we did not observe any effect of synchrony in the right auditory cortex related to auditory processing and prosody, as it was reported in Hubbard et al.'s results. A further relevant aspect in our study is that participants were asked to simply focus on an auditory detection task, instead of explicitly monitoring speech-gesture synchrony. This is interesting because our results cannot be attributed to the explicit (meta-linguistic) task of monitoring speech-gesture synchrony but, as a consequence our task may have decreased attention on visual information and effectively weakened the expression of beat synchrony on speech processing networks.

Taken together, the present results provide new insights about the specificity of left MTG and IFG in the processing of multimodal language (for a review, see Campbell, 2008; Özürek, 2014). As participants were not explicitly asked to pay attention to the speaker's hands, this suggests that the temporal correspondence between beats and speech prosody may be picked up

automatically. This is in line with previous proposals considering speech and gestures as two side of a same underlying language system (McNeill, 1992; Kelly, Creigh and Bartolotti, 2009). Beats appear to convey additional communicative value such as speakers' intentions, which are not available (or at least, not extracted) from simple visual stimuli (Holle et al., 2012; So et al., 2012; Casasanto & Jasmin, 2009; McNeill, 1992). The access to concurrent gestures during speech perception may engage the listeners and provide a better alignment between listener and speaker, improving speech processing and information encoding. Finally, the fact that the speaker was a well-known former Spanish president may have engaged some political sensitivity from listeners. However, such a possible bias is unlikely to influence our results, since participants viewed the same speaker across all four experimental conditions.

5. CONCLUSION

We investigated the neural correlates of spontaneous beat gestures produced in continuous speech. Our results revealed that the asynchrony affected language-related areas activations differently according to the visual information accompanying speech during perception. We concluded that beats conveyed visual aspects of language by their trajectories aligned with speech prosody, but also communicative intentions of the speaker.

AKNOWLEDGMENTS

This research was supported by the Ministerio de Economia y Competitividad (PSI2013-42626-P), AGAUR Generalitat de Catalunya (2014SGR856), and the European Research Council (StG-2010 263145).

REFERENCES

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: role of the STS region. *Trends in Cognitive Sciences*, 4(7), 267–278.

Biau, E., & Soto-Faraco, S. (2013). Beat gestures modulate auditory integration in speech perception. *Brain and Language*, 124(2), 143–52.

Biau, E., Torralba , M., Fuentemilla, L., de Diego Balaguer, R., & Soto-Faraco, S. (in press). Speaker's hand gestures modulate speech perception through phase resetting of ongoing neural oscillations. *Cortex*.

Brett, M., Anton, J-L., Valabregue, R., & Poline, J-B. Region of interest analysis using an SPM toolbox [abstract] Presented at the 8th International Conference on Functional Mapping of the Human Brain, June 2-6, 2002, Sendai, Japan. Available on CD-ROM in NeuroImage, Vol 16, No 2.

Callan, D. E., Jones, J. A., Callan, A. M., & Akahane-Yamada, R. (2004). Phonetic perceptual identification by native- and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory-auditory/orosensory internal models. *NeuroImage*, 22(3), 1182–94.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology : CB*, 10(11), 649–57.

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 363(1493), 1001–10.

Casasanto, D., & Jasmin, K. (2010). Good and bad in the hands of politicians: spontaneous gestures during positive and negative speech. *PloS One*, 5(7), e11805.

Dick, A. S., Mok, E. H., Raja Beharelle, A., Goldin-Meadow, S., & Small, S. L. (2014). Frontal and temporal contributions to understanding the iconic co-speech gestures that accompany speech. *Human Brain Mapping*, *35*(3), 900–17.

Dick, A. S., Goldin-Meadow, S., Hasson, U., Skipper, J. I., & Small, S. L. (2009). Co-speech gestures influence neural activity in brain regions associated with processing semantic information. *Human Brain Mapping*, 30(11), 3509–26.

Eckstein, K., & Friederici, A. D. (2005). Late interaction of syntactic and prosodic processes in sentence comprehension as revealed by ERPs. *Brain Research. Cognitive Brain Research*, *25*(1), 130–43.

Eckstein, K., & Friederici, A. D. (2006). It's early: event-related potential evidence for initial interaction of syntax and prosody in speech comprehension. *Journal of Cognitive Neuroscience*, *18*(10), 1696–711.

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., & Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage*, *25*(4), 1325–35.

Friston, K. J., Glaser, D. E., Henson, R. N. A., Kiebel, S., Phillips, C., & Ashburner, J. (2002). Classical and Bayesian inference in neuroimaging: applications. *NeuroImage*, *16*(2), 484–512.

Glaser, Y. G., Martin, R. C., Van Dyke, J. A., Hamilton, A. C., & Tan, Y. (2013). Neural basis of semantic and syntactic interference in sentence comprehension. *Brain and Language*, *126*(3), 314–26.

Grossman, E. D., Blake, R., & Kim, C.-Y. (2004). Learning to see biological motion: brain activity parallels behavior. *Journal of Cognitive Neuroscience*, 16(9), 1669–79.

Guellaï, B., Langus, A., & Nespor, M. (2014). Prosody in the hands of the speaker. *Frontiers in Psychology*, *5*, 700.

Habets, B., Kita, S., Shao, Z., Ozyurek, A., & Hagoort, P. (2011). The role of synchrony and ambiguity in speech-gesture integration during comprehension. *Journal of Cognitive Neuroscience*, *23*(8), 1845–54.

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends in Cognitive Sciences*, 9(9), 416–23.

Hein, G., Doehrmann, O., Müller, N. G., Kaiser, J., Muckli, L., & Naumer, M. J. (2007). Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *27*(30), 7881–7.

Herrington, J. D., Nymberg, C., & Schultz, R. T. (2011). Biological motion task performance predicts superior temporal sulcus activity. *Brain and Cognition*, 77(3), 372–81.

Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience*, 19(7), 1175–92.

Holle, H**.**, Gunter, T. C., Ruschemeyer, S. A., Hennenlotter, A., & Iacoboni, M. (2008). Neural correlates of the processing of co-speech gestures. *Neuroimage, 39*(4), 2010-2024.

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.

Holle, H., Obleser, J., Rueschemeyer, S.-A., & Gunter, T. C. (2010). Integration of iconic gestures and speech in left superior temporal areas boosts speech comprehension under adverse listening conditions. *NeuroImage*, 49(1), 875–84.

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(3), 1028–37.

Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language*, 89(1), 253–60.

Kelly, S. D., Ozyürek, A., & Maris, E. (2010). Two sides of the same coin: speech and gesture mutually interact to enhance comprehension. *Psychological Science*, 21(2), 260–7.

Kelly, S. D., Ward, S., Creigh, P., & Bartolotti, J. (2007). An intentional stance modulates the integration of gesture and speech during comprehension. *Brain and Language*, 101(3), 222–33.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. Nature Neuroscience, 12(5), 535–40.

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(3), 396–414.

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

Macaluso, E., George, N., Dolan, R., Spence, C., & Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech: a PET study. *NeuroImage*, 21(2), 725–32.

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.

Meyer, M., Steinhauer, K., Alter, K., Friederici, A. D., & von Cramon, D. Y. (2004). Brain activity varies with modulation of dynamic pitch variance in sentence melody. *Brain and Language*, 89(2), 277–89.

Noesselt, T., Rieger, J. W., Schoenfeld, M. A., Kanowski, M., Hinrichs, H., Heinze, H.-J., & Driver, J. (2007). Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *27*(42), 11431–41.

Nummenmaa, L., & Calder, A. J. (2009). Neural mechanisms of social attention. *Trends in Cognitive Sciences*, 13(3), 135–43.

Obermeier, C., Holle, H., & Gunter, T. C. (2011). What iconic gesture fragments reveal about gesture-speech integration: when synchrony is lost, memory can help. *Journal of Cognitive Neuroscience*, *23*(7), 1648–63.

Obermeier, C., & Gunter, T. C. (2014). Multisensory Integration: The Case of a Time Window of Gesture-Speech Integration. *Journal of Cognitive Neuroscience*, 1–16.

Obleser, J., Meyer, L., & Friederici, A. D. (2011). Dynamic assignment of neural resources in auditory comprehension of complex sentences. *NeuroImage*, *56*(4), 2310–20.

Özyürek, A. (2014). Hearing and seeing meaning in speech and gesture: insights from brain and behaviour. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *369*(1651), 20130296.

Pelphrey, K. A., Morris, J. P., & McCarthy, G. (2004). Grasping the intentions of others: the perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. *Journal of Cognitive Neuroscience*, 16(10), 1706–16.

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(3), 260–77.

So, W. C., Sim Chen-Hui, C., & Low Wei-Shan, J. (2012). Mnemonic effect of iconic gesture and beat gesture in adults and children: Is meaning in gesture important for memory recall? *Language and Cognitive Processes*, 27(5), 665–681.

Straube, B., Meyer, L., Green, A., & Kircher, T. (2014). Semantic relation vs. surprise: the differential effects of related and unrelated co-verbal gestures on neural encoding and subsequent recognition. *Brain Research*, *1567*, 42–56.

Thioux, M., Gazzola, V., & Keysers, C. (2008). Action understanding: how, what and why. *Current Biology : CB*, 18(10), R431–4.

Treffner, P., Peter, M., & Kleidon, M. (2008). Gestures and Phases: The Dynamics of Speech-Hand Communication. *Ecological Psychology*, *20*(1), 32–64.

Uchiyama, Y., Toyoda, H., Honda, M., Yoshida, H., Kochiyama, T., Ebe, K., & Sadato, N. (2008). Functional segregation of the inferior frontal gyrus for syntactic processes: a functional magnetic-resonance imaging study. *Neuroscience Research*, *61*(3), 309–18.

Wang, L., & Chu, M. (2013). The role of beat gesture and pitch accent in semantic processing: an ERP study. *Neuropsychologia*, 51(13), 2847–55.

Willems, R. M., Ozyürek, A., & Hagoort, P. (2007). When language meets action: the neural integration of gesture and speech. *Cerebral Cortex (New York, N.Y. : 1991)*, 17(10), 2322–33.

Willems, R. M., Ozyürek, A., & Hagoort, P. (2009). Differential roles for left inferior frontal and superior temporal cortex in multimodal integration of action and language. *NeuroImage*, 47(4), 1992–2004.

Wu, Y. C., & Coulson, S. (2010). Gestures modulate speech processing early in utterances. *Neuroreport*, 21(7), 522–6.

Yasinnik, Y. (2004). The timing of speech-accompanying gestures with respect to prosody. *Proceedings of From Sound to Sense,MIT.* MIT.

3. GENERAL DISCUSSION

The general aim of the present thesis was to gain a better understanding of beat gesture processing and its neural correlates during speech perception. We adopted a novel approach by using conditions of speech production that are closer to real life speech than it has been normally done in this field of research. The advantage of such approach is that we presented listeners with a natural, continuous AV speech stream instead of isolated syllables, words or audio only continuous speech. To do so, we designed an experimental protocol based on electrophysiology (EEG/ERP) and neuroimaging (fMRI) on the real-life recordings of political discourses. Public speaking favours the production of a particular type of spontaneous gestures (i.e. beats) in a legitimate semantic context, but also maintains the integrity of the natural, rhythmic co-occurrence between gesture (beats' apexes) and speech (prosody) in terms of their frequency and synchrony. The use of different neuroimaging techniques allowed us to investigate the neural correlates of beat-speech integration in both its temporal (ERPs and oscillatory activities) and spatial (fMRI) dimensions. Adopting this approach allowed us to develop new experimental procedures, and use beats produced with a continuous discourse as a good model of gesture-speech processing, and finally to bring about some new empirical data of cospeech neural correlates, published in the scientific articles included in this dissertation. This thesis considers three main hypotheses related to language processing with accompanying gestures.

(1) **Beat gestures modulate early stages of auditory processing during continuous speech perception.** Previous studies have reported that the production of beat modulated significantly the acoustic properties of the accented, co-occurring syllable (increase of pitch accent, and loudness and duration) of the corresponding word (Krahmer & Swers, 2007). Consequently, listeners perceived affiliate words as more prominent in short sentences. But, more interestingly, when participants saw a speaker producing a visual beat on a word, they perceived it as more prominent than when they did not see the beat gesture on the same word (and hence, acoustically identical). These results suggested not only a trivial effect of pure loudness perception, but also that listeners internally emphasized word prominence with the mere sight of a beat. Based on these evidences, we hypothesized that beats may affect the phonological processing of corresponding words during speech perception. At a neural level, we expected ERP modulations at early latencies, corresponding to the N1/P2 time window, reflecting such an effect (Stekelenburg & Vroomen, 2007; van Wassenhove, Grant & Poeppel, 2005; Näätänen, 2001; Rugg & Coles, 1995).

(2) **Beats may bear a predictive value within speech signal as they are temporally aligned with speech prosody and anticipate the corresponding word** (Leonard & Cummins, 2011; Treffner, Peter & Kleidon, 2008). Indeed, beats may be

susceptible to reduce the uncertainty about when relevant acoustic cues will occur, and hence facilitate corresponding speech processing (Arnal & Giraud, 2012). We hypothesized that if beats had a predictive value on associated words, then they should modulate timing coding and low-frequency entrainment process at word onsets. At a neural level, we addressed this possible effect by measuring synchronization of low-frequency activity around word onsets as a neural signature of the integration between beats and auditory information.

(3) **Beats convey communicative value and are perceived as linguistic visual information.** First, we hypothesized that if this statement is true, then an asynchrony between beats' apexes of hand movements and pitch accents in speech may affect BOLD responses in language-related areas such as left IFG and left STS (Marstaller & Burianova, 2014; Hubbard et al., 2009). Second, in order to determine the specificity of this interaction, we addressed whether any visual cue (i.e., simple discs instead of hands) may be enough to accomplish the same linguistic function that gestures have when combined with speech. Instead, beats may convey additional communicative intentions of the speaker engaging a specialized mechanism. At a neural level, we expected qualitatively distinct modulations of BOLD responses in the language related areas by synchrony between speech and beats, versus synchrony between speech and other visual cues (discs).

3.1 About new experimental procedures

One of the challenging aspects of the present thesis was to setup new experimental paradigms, based on what was already done in terms of conditions and contrasts (classically, comparing an AV processing with audio and visual only conditions). As previously discussed, beats may be seen as simple biphasic flicks of the hands, but are exquisitely timed with the speech signal in ways that may not be trivial. We thought that it was necessary to come up with new experimental protocols that would maintain the gestures' function intact. To date, investigations have mostly presented isolated beats in short sentences (Krahmer & Swerts, 2007; Holle et al., 2012; Wang et al., 2013) or even lists of isolated words (So et al., 2012). This approach, although necessary, may have virtually increased the prominence of beats and "forced" attention to the hands in an artificial way, affecting any conclusions about the reported automaticity of their integration with associated speech. Additionally, these protocols may have disrupted the essence of the function that beats have in a fluent, integrated audiovisual speech, rendering them potentially trivial in off-context staged productions. As an alternative, we advocated that gestures should preserve their temporal and semantic alignment with speech to be integrated naturally integral to the message, as the visual stream of speech. One limiting factor in many studies using staged materials is that beats are very difficult to produce on demand in a discrete manner, as the speaker without any explicit purpose of the experiment

should spontaneously produce them. Thus, we opted for AV material that satisfied both aspects of gesture-speech production: Aspect1, Aspect2. Public addressees revealed to be a good compromise as they maintained the temporal alignment between beats and speech flows, and beats were spontaneous and varied (because the speaker produced naturally discrete or cohesive beats, in a large display of hand shapes). The production of three scientific articles based on the presentation of TV broadcasted political speeches (section 2.1, 2.2 and 2.3), validated this approach and the experimental paradigms that we set to investigate gesture-speech processing.

To date, only two studies had investigated the time course of beat gestures, one in terms of syntactic parsing (Holle et al., 2012) and the other in terms of semantic processing (Wang & Chu, 2013) with ERPs, and their neural correlates during speech processing with fMRI (Hubbard et al., 2009). These studies provided first evidence that beats are perceived differently from simple hand movements without communicative intention and actually modulate certain aspects of speech processing (at syntactic and semantic stages). Through our three articles, we provided converging ERP, oscillations (EEG) and fMRI evidence favouring the idea that beats are effectively perceived as visual linguistic information of speech, as part of a common language system. In general, our results support earlier, theoretical accounts that go in the same direction (McNeil, 1992). Our results are also in line with previous behavioural evidences that investigated beat gestures' impact at

phonological processing level (section 4.1) and the crucial temporal alignment between beats and speech prosody (sections 4.2 and 4.3). In the following sections of the general discussion, I will relate our main findings with these previous reports and discuss their possible interpretation.

3.2 Beat gestures modulate phonological level in speech processing: possible acoustic and attentional effects

In our first study (section 4.1), we demonstrated that the spontaneous beats modulated the auditory integration by mean of a naturalistic continuous speech presentation and ERPs analysis. The ERPs time-locked to the onset of words accompanied by a beat gesture were significantly less negative than equivalent words pronounced without gesture. Importantly, the presence of beats modulated the signal processing at early stages of auditory integration in a temporal window (220-280 ms) coinciding with the auditory ERP component P200 (P2) of the N1/P2 ERP classic complex. This neural correlate is in line with previous behavioural evidences showing that listeners perceive gesture-associate words as more prominent than the rest, in short sentences (Krahmer & Swers, 2007). As the production of a beat modulates the acoustic properties of the corresponding word (Krahmer & Swers, 2007) and apexes co-occur with pitch accents (Leonard & Cummins, 2011; Yasinnik, Renwick & Shattuck-Hufnagel, 2004), we see at least two possible interpretations of our results. In the first one (that may be

the most evident), the production of a co-occurring beat with relevant verbal information modulates it in a significant way that is perceivable from the listener's side. In other words, when the speaker accompanies the utterance with a beat, he pronounces it louder and modifies the prosody (increasing pitch accents for example), impacting pure auditory aspects of the signal and hence, processing on the listener's side (reflected at P2 corresponding time window) without engaging any particular pragmatic process. In the literature, the P2 were often described as a classic ERP component reflecting auditory processing (Colin et al., 2002; Näätänen, 2001; Rugg & Coles, 1995) and audiovisual integration (Stekelenburg & Vroomen, 2007; van Wassenhove, Grant & Poeppel, 2005). Thus, the modulation observed at P2 in the signal integration in our first study reflected a pure difference of acoustic properties between words pronounced with a beat and equivalent words pronounced without beat.

Nevertheless, this "direct" and simple influence of gestures on the signal is only part of the whole story. Other evidences suggest that the modulations in the ERP signal might instead, or in addition, reflect attentional effects of beats on the listener's processing of their corresponding affiliate words. First, in our experiment (section 4.1), we controlled for various acoustic properties of words accompanied with a beat and their equivalent pronounced without beat (loudness, syllable length, and F0, F1 and F2). We did not find any significant difference between the two kinds of words. This may be explained by the lower quality of

sound of the broadcasted AV speech taken from Internet. Another explanation may be that the conditions in which the speech was recorded were not sensitive enough to notice the subtle acoustic variations in the speaker's voice (the distance between the speaker and the microphone for example, was not as controlled as in experimental conditions). In any case, the important thing is that the acoustic properties of words between the two conditions could not explain the modulation of ERPs at P2. Second, in the audio only modality of speech presentation (when we removed the visual information for both words accompanied by a beat or not), the difference at P2 time window disappeared, suggesting that the effect was due to visual information in the AV modality of speech presentation. Third, in our second study (section 4.2), we analysed the EEG signal before the word onsets. This is very relevant information as it allowed us to investigate the period between the beat onset and the corresponding word onset. Thus, any modulation of the auditory signal integration during word processing before its onset may be attributed to preceding visual information and possibly to the beat gesture. More generally, it would suggest that the visual context in which the following word is processed was modulated by relevant preceding visual information. McNeill (1992) already hypothesized that beat gestures play the role of highlighters and help the speaker to attract listener's attention on relevant parts of speech. The time course of the ERP modulations in the first study (section 4.1) supports this claim, in the light of previous ERP studies investigating the attentional effects on auditory integration that have reported effects at P2 time window

(Hillyard et al., 1973; Näätänen, 1982; Picton & Hillyard, 1974). In an ERP study, Astheimer and Sanders (2009) showed that acoustic probes placed around word onsets elicited greater amplitudes in the N1 time window of the N1/P2 ERP component, as compared to other probes placed in non-relevant sites. These results suggested the influence of these probes varied in function of their position in speech because listeners did not allocate the same attentional load during all the speech, but rather around informative cues (word onsets). The authors concluded that temporally selective attention was attracted by relevant acoustic information during audio speech perception and modulated the auditory integration at N1/P2 time window. Pilling (2009) presented isolated syllables and showed that N1/P2 amplitudes of the ERPs time-locked to their onset were significantly reduced in the AV respect to audio only modality. The author suggested that the preceding visual information (lip movement) provided an alerting cue for the following onset of corresponding auditory speech, and then, involving attentional processes. These results were in line with previous report in which van Wassenhove et al. (2005) also suggested an attentional effect of preceding visual information on auditory integration during AV processing of isolated syllables. The author advanced that visual information allows making predictions monitored by attentional processes on the upcoming auditory speech, and the N1/P2 amplitude reflects the prediction facilitation. Then, in line with these results, the modulation of signal at N1/P2 corresponding time window in our ERPs data may reflect the influence of the preceding

beat gesture on the following targeted word, through attentional mechanisms as well.

Although less exciting, the first "acoustic" hypothesis is compatible with the attentional hypothesis, as speakers modulate their intensity when producing a beat, and this might be probably reflected during processing at phonological levels. In the present thesis, I favour the attentional explanation over the purely acoustical one. This interpretation is based on several sources of evidence. First, Krahmer and Swerts (2007) showed that listeners perceived words when produced with a beat more prominent than the exact same ones without beat. This strongly suggests that listeners kind of "simulate" the emphasizing weight of preceding beat, even without acoustic difference. Second, the results reported in the second study empirical of this thesis (section 4.2) also conformed to the attentional hypothesis. When visual information was available during speech perception, a decrease of alpha synchronization was observed around the onset of words pronounced with a beat gesture as compared to words pronounced alone. Previous studies demonstrated that a desynchronization of alpha activity reflected attentional deployment on stimuli (Thut, Nietzel, Brandt & Pascual-Leone, 2006; Rohenkohl & Nobre, 2011 for examples). Importantly, words onsets are relevant anchors for segmentation during speech perception, as the theta phase was showed to match the spectro-temporal structure of the utterance (Luo & Poeppel, 2007). Then it makes sense to assume that alpha desynchronization at word onset may reflect beneficial attentional effects. Further,

alpha desynchronization is related to speech processing. For example, Krause et al. (1997) showed that auditory speech perception decreased alpha synchronization. Further, this desynchronization was independent from the intelligibility of speech in the 8-10 Hz band (presenting participants with either normal or backward auditory speech), which was actually our frequency band of interest (in the second study). Krause et al. hypothesized that the alpha desynchronization reflected pure attentional processes, independently from speech content analysis but rather on the analysis of general stimuli spectro-temporal structure (i.e. speech envelope). Regarding to McNeill (1992) and our studies (sections 4.1 and 4.2), we hypothesized that the potential attentional influence of beats on relevant content during speech perception relies on two things. First the temporal alignment of beats and their corresponding words that maintains a systematic order or perception (beat then corresponding utterance). Second, as suggested (McNeill, 1992; So et al., 2012; Holle et al., 2012), some pragmatic and communicative intentions probably acquired through social experience are extracted from simple beats during speech perception ("I know why you put a beat at this moment because I would have done the same"; "what follows is important because you initiated a beat"). This may be because listeners also gesture when become speakers, or because of a mutual behavioural synchronization between both partners of the conversation. Consequently, beats may be perceived as visual cues indicating often what is important and needs more attention from listeners. Thus, beats are able to attract or guide the focus of listeners'

attention in particular moments and modulate the processing of following corresponding auditory segment.

Finally, the results from the study 1 and 2 in this thesis suggest the automaticity of beat-speech processing, and the weight of beats on attention attraction of the listeners. That is, in those 2 studies (section 4.1 and 4.2), participants attended to a continuous AV speech in which other visual information was available (speaker's face, background, etc…). Additionally, they were not explicitly asked to pay attention to the speaker's hands, as they had to do a memory task on the content after speech perception. Still, beats modulated the ERPs/low-frequency activities of corresponding words when visual information was available suggesting that listeners naturally give weight to beats. However, one can argue that, as listeners knew that they were about to be evaluated on a memory task, they paid more attention to speech content and possibly beats than normally. When information was not available (e.g., audio alone modality), ERPs and oscillatory activities were not different between words pronounced or not with a beat, suggesting the visual attention modulation by beat gestures. If it is the case (although not testable here), this would mean that listeners used all available speech information in AV modality, and beats constituted reliable visual information indicating when relevant utterance segments were coming, conforming with McNeill's hypothesis. Such automaticity goes with previous behavioural studies that suggested that gestures and verbalizations are part of the same language system and their integration together

is systematic (Kelly, Creigh and Bartolotti, 2009). However, these studies used material with short sentences or isolated words presented with unique, salient beats (Krahmer & Swerts, 2007; Holle et al., 2012; So et al., 2012; Wang et al., 2013), which may have artificially forced listeners to take into account beats (or else, they could not ignore beats and inferred their task-specific relevance). Here, for the first time, results obtained with a more realistic approach presenting spontaneous beats integrated with a natural, continuous speech seem to confirm the automaticity hypothesis without the previous caveats.

3.3 Beats as road signs: The possible predictive value of beats on critical corresponding words

As previously discussed, the systematic order and rather precise timing between beat's initiation and the subsequent corresponding affiliate word's onset confers the gesture the potential to influence how the affiliated speech segment is integrated during perception. Actually, the attentional hypothesis developed in the sections above may relate to which weight listeners give to beats in AV speech perception. In other words, beats attract the focus of attention because they are robust visual cues relevant for the online segmentation of the continuous auditory speech. Regarding the order of presentation (beat then word), we assumed that beats facilitate the anticipation of relevant words during online speech processing. We addressed this aspect in the second study (section 4.2), in which we hypothesized that beats bear

a predictive value on salient acoustic cues in the auditory speech signal (i.e. the onsets of affiliated words). We assumed that if the gesture allows directing listener's attention during the lag between gestures and word onsets (i.e. around 200ms, see Biau & Soto-Faraco, 2013) on affiliate words, this may be reflected by a change in the brain's state initiated within this short time window and lasting after the following word's onset. The modulations of oscillatory activity in the theta/alpha bands around the incoming word onset indexed the anticipatory effect of beat gestures at neural levels. Our results revealed an increase of theta phase synchronization with a co-occurring decrease of alpha phase synchronization around onsets when words were preceded by the preparation phase of a beat gesture, as compared to equivalent words pronounced without beat. We concluded that words were better anticipated by the presence of a preceding beat (reflected by the theta phase synchronization), and this effect probably engaged attentional processes reflected by a modulation of the alpha activity. Importantly, these differences of phase synchronizations in theta and alpha bands were not found when words pronounced or not with a beat were presented without visual information, suggesting an effect of congruent visual information on auditory speech integration. These conclusions were in line with previous proposal stating that theta phase synchronization around periodic acoustic features may be enhanced by a stable preceding (predicting) visual cue during speech perception (Arnal & Giraud, 2012). By prediction, the authors meant the process that decreases the uncertainty about the likelihood for periodic relevant cues to occur,

then facilitating their processing. For example, Arnal, Wyart, and Giraud (2011) showed that a mismatch between lip movements and auditory speech generated a violation of predictions, reflected by different patterns of low-frequency activities, as compared to congruent presentation between audio and visual modalities. Concerning beat gestures, we argued that predictive visual information from the speaker's hands was integrated with the spoken signal through theta synchronization at word onsets. More generally, theta frequency and oscillatory activity in brain are intrinsically related with speech processing. On the signal's side first, speech can be segmented as a chain of discrete units, syllables (200 ms long), corresponding to a period of 4-8Hz theta oscillatory activity (Ghitza & Greenberg, 2009; Greenberg, 1999; Greenberg, Carvey, Hitchcock, & Chang, 2003; Peelle & Davis, 2012). Second, theta phase synchronization has been proposed as a potential mechanism enabling predictive coding and reflecting the anticipation of certain auditory features of speech (Arnal & Giraud, 2012; Lakatos et al., 2008; Schroeder & Lakatos, 2009; Schroeder et al., 2008). Theta activity tunes to syllabic periodicity and time-frequency architecture of speech envelope (Luo and Poeppel, 2007). Consequently, an increase of theta synchronization at word/syllable onsets suggests that correlated preceding visual information (i.e. beats) leads to excitable states alternating predictably, thereby improving the sensory processing when the relevant audio input comes at the right moment (Busch, Dubois & VanRullen., 2009; Engel, Fries, & Singer, 2001; Lakatos et al., 2008; Schroeder & Lakatos, 2009; Schroeder et al., 2008). Nevertheless, further

experiments are needed to correlate behavioural evidence of speech analysis facilitation and low-frequency activity modulation in the context of beat gestures. A recent article arising from this thesis, Biau & Soto-Faraco (*in press*), presents this perspective. The article has been included in an annex, at the end of the present thesis (see Annex 4).

Beyond this interpretation of the results so far, one further question is: What makes listeners attribute a predictive value to the speaker's hand beats so that they direct attention on the highlighted acoustic cues for speech processing? As they do not convey semantic content, beats cannot help prediction on the following speech content. Rather, it seems that listeners attribute predictive value because they know *why* the speaker gestures at precise moments. Being temporally aligned with suprasegmental features of audio speech (i.e. prosody) may confers beats the same role, and listeners base on preceding beats to anticipate following corresponding audio prosody modulations. This suggests that, nevertheless, simple beats engage complex cognitive processes and, by inference, bear additional communicative information, as audio prosody. I develop this aspect in the following section.

3.4 Beats as visual prosody: Gestures may convey additional communicative information.

The attentional effect of beats and their potential predictive value rely on how listeners acquired experience through social

interactions, and depend on the communicative weight that listeners attribute to the gestures. McNeill (1992) already described that beats accompany words that are often more relevant for the external context of the narrative rather than directly related to the immediate context. Beats will serve to add detail which may not be fundamental in the sentence itself, but for the whole story. Also, beats can be used to introduce a new character in the narrative, which will not be important for what he is doing in the present sentence, but for the rest of the story. Beats also serve to underline additional information related to a central character (in this case, a beat accompanies the name, then the surname, etc...). In any case, there is a common implicit consensus between the speaker and the listener. The correct production and interpretation of beats requires knowledge about the narrative structure. According to McNeill (1992), a narrative is not a succession of short episodes, but rather, a continuous shift in time and space, with a change of distance between the speaker, the discourse and the listener, leading to different levels. Here in the present thesis, I will not describe them in detail but when speaking, the speaker alternates different moments that constitute the whole narrative: 1) The narrative level that constitutes the story properly. The speaker describes exactly what happened in the sequential order of the actual story. 2) The metanarrative level at which the speaker contextualizes the story with sentences that add information on the characters or when the story takes place. Thus, the metanarrative clauses do not respect the temporal order of the events, as the speaker decides when something has to be signalized to facilitate comprehension. 3) The

paranarrative level at which the speaker refers to his own experience and expresses impressions or emotions, out of the storytelling. The speaker also implies the listener (for example: "*have you seen this film, right*"). The preponderance of the paranarrative level highly depends on the relationship between the speaker and the listener (as it serves to synchronize and put them on the same page).

The different types of gestures serve preferentially one narrative level. For example, the speaker produces more iconic gestures when he is engaged in the narrative level because he needs complementary visual information to describe actions and objects from the story. In our case, the speaker produces more beat gestures for metanarrative and paranarrative levels (beats shift between both levels) as he needs to maintain the attention of the speaker and to involve him in the conversation (making sure they are on the same page). Thus, albeit simple, beat gestures engage complex cognitive processes and may bear the communicative intentions from the speaker that the listener has to interpret to follow the discourse (and distinguish narrative moments from the others). McNeill evoked the fact that before five years old, children do not produce beats, which remain sporadic until 11 years old. First, this suggests that young children still have not developed the narrative structure (with meta and paranarrative levels developing probably even later). Second, that beats, even if very easy in terms of motor production, serve complex speech processes that require social interactions and experience. On the listener's side, a study investigated the effects of

beats on the encoding of isolated words, in adults and 4-5 years old children. So, Chen-Hui and Wei-Shan (2012) showed that listeners managed to recall more words when they had been accompanied with iconic or beat gestures than equivalent words pronounced without gestures. In contrast, children only benefited from iconic gestures whereas beats had no effect on word recall as compared to words pronounced alone. Taken together, these results are in line with McNeill (1992) as they showed that beats interpretation require developed linguistic skills to be fully functional, and that they are relevant visual information in adults even if they did not bring any explicit semantic content.

One may question whether these processes (leading to attentional and, meta- / paranarrative functions) are triggered simply by the mere temporal alignment between the kinematics of gestures with the acoustic envelope. If so, then any kind of sufficiently salient visual cue correlated with the acoustic properties of the speech signal would then be enough. Alternatively, do beat gestures differentially engage these processes (as opposed to simple visual cues aligned with auditory signal envelope)? As previously described in the introduction, one ERP study investigated the potential effect of beats on syntactic analysis during ambiguous sentences (Holle et al., 2012). In this study, they found an effect on the P600 component with a significant decrease when the beat accompanied the critical word for disambiguation in complex form sentences. This first result suggested that beats effectively helped for disambiguation, as the P600 is a positive going wave reflecting

some aspects of the syntactic analysis during sentence processing (van de Meerendonk et al., 2010; Haupt et al., 2008; Friederici, 2002; Frisch et al., 2002). More relevant here, the authors found no equivalent effect on the P600 when the beat was replaced by a moving dot following the exact same spatiotemporal trajectories as the hands in the gestures. These results suggest that the simple temporal alignment of the movement with auditory speech rhythm is not enough to confer visual information a linguistic value during speech perception. In contrast, the beats are conveyed by the hand attached to the speaker body and may transmit communicative information (e.g. emotion, intentionality). Our fMRI results in the third study presented in this thesis (section 4.3) go with Holle et al. (2012) and So, Chen-Hui and Wei-Shan (2012), as they suggest that equivalent moving discs are processed differently from real beat gestures. We showed that breaking up the temporal alignment between auditory speech and visual information affected differently the BOLD responses in the language related areas. In particular, we found reversed patterns of modulation when the auditory speech signal came with beats as compared to moving discs following the exact same trajectories as the hands in the beat gestures. In particular, the effects of synchrony that was selective for hand beat vs. discs were seen in the left IFG, left MTG and occipital cortex. Further analysis suggested greater activations in these areas when beats and auditory speech were presented in synchrony than asynchrony. The exact reverse pattern was observed with discs instead of beats, conforming to the hypothesis that beats convey additional information that engages other cognitive processes that

with trivial visual cues. Yet, the post-scan session questionnaire revealed that participants associated moving discs with the speaker's hands, suggesting first that our methodology was good (conserving the physical properties of hands' movements). But also, that the correct velocity, acceleration and trajectories of a simple visual cue in the peripersonal space of a speaker, are enough to associate it with a body part (i.e. the hands) without apparently conferring it the social value. In fact, Hubbard et al. (2009) showed in another fMRI study that when beats' characteristics were fully conserved but presented without speech (e.g. the speaker was visible but the audio was removed), beats were not processed differently from non-sense movements anymore. Then, from Hubbard's results and ours, it appears that to bear interpretable communicative intentions, beats' kinematics have to be contextualized by a speech context.

Together, the previous ERPs study (Holle et al., 2012) and our fMRI data bring neural evidence to a recently published study that investigated the behavioural modulation of beats in syntactic parsing (Guellaï, Langus & Nespor, 2014). The authors presented participants with sentences with two possible meaning depending on the prosody in audio only or AV modality. After each sentence, participants were asked on their interpretation according to the prosody (the answer was considered as correct if it followed the prosody). Guellaï and colleagues' results showed that correct responses were decreased when the beats mismatched the auditory prosody in the AV modality, as compared to audio only or AV

matching modalities. These results showed that congruent beats with prosody did not help to better comprehend ambiguous sentences than audio alone. This is not very surprising as auditory speech in itself convey already the semantic context and the syntactic structure sufficient for comprehension (think about phone conversations for example). But more important for us, beats mismatching prosody significantly decreased the correct response rates. First, these results conformed to the hypothesis that during speech perception, listeners use beats and perceive them as part of the same language system. Second, as hypothesized in our third study (section 4.3), the auditory prosody extends to visual prosody through beat gestures. It is now well established that speakers can manipulate prosody to serve communicative purposes. For example, they can modulate pitch accents to introduce a distance between speech content and their state of mind (e.g. irony). They can also produce vocal inflections to accompany sarcasms, or to clarify the speech act they want to make (i.e. question or affirmation). In any case, the subtle interpretation of prosody requires complex cognitive processes on the listener's side. Thus, regarding behavioural and neural evidence, beat gestures as visual prosody probably engage speaker's intentions as well, and maybe help to explicit them with auditory prosody. Finally, beats gestures belong to a big family of "beats" conveyed through different body parts. Krahmer and Swers (2007) compared beat gestures to eyebrow and head movements for instance. They found comparable effects on speech production and perception of accented words accompanied by these three body parts (i.e. modulation of acoustic properties of

the accented syllable, and significant relevance of target word). Yehia et al. (2002) reported that natural head movements of the speaker correlate with the fundamental frequency (F0, i.e. pitch accents) and amplitude (loudness) modulations. More precisely, Munhall et al. (2004) showed that head movements during speech production match the pitch accents with a frequency of around 3Hz that corresponds to the prosody in the auditory signal. Additionally, the authors showed that the sight of the speaker's head movements improved significantly intelligibility of speech when head beats were congruent with pitch accents in noisy conditions. Thus, beat gestures (as performed with eyebrows and head) can be considered as corporal language and convey, even implicitly, some aspects of the speaker's mind that listeners can perceive and interpret once necessary language skills have developed in the early years of life (So et al., 2012; McNeill, 1992). In other words, if beats share the same temporal characteristics of auditory prosody, they may convey the same communicative intentions as well.

3.5 Do the present neural modulations reflect specific beat effects, or biological motion?

Altogether, our three studies brought relevant spatiotemporal neural data to understand the role of gesture in communication, and conform to previous neuroimaging results dealing with beats and gesture-speech processing in general (Marstaller & Burianová, 2014; Hubbard et al., 2009; McNeill, 1992). Nevertheless, as in many studies using neuroimaging techniques to investigate gesture-

speech processing, we could not combine behavioural measures in our own work (I will comment a set of behavioural studies related to this thesis, later on). Thus, one can argue that our neural effects on the time course of beat processing (section 4.1 and 4.2) and their neural correlates (section 4.3) may be possibly due to the perception of biological motion. This is an important alternative hypothesis to consider in our interpretation. However, we have a series of arguments in favour of a specific effect of beat gesture perception with co-occurring speech segments:

First, in our three experiments we always compared the beat condition (i.e. AV speech in which beats were naturally aligned with speaker's prosody) with an equivalent AV condition in which only the co-occurrence between apexes and pitch accents were naturally absent (section 4.1 and 4.2) or shifted in time (section 4.3). In the two first studies, the pairs of words (either pronounced with or without beat) came from the exact same AV speech, only that in the no gesture condition, the speaker although visible, did not accompanied the critical word with a beat. Except for the gesture, the average biological motion was highly similar in both conditions. Further, even if not gesturing with the hand, the speaker eventually moved the rest of the upper part of the body, compensating the absence of an explicit beat in terms of visual modulation in the no gesture condition. Previous studies investigating the biological motion perception with ERPs or oscillatory activity have reported quite different modulations (in patterns and time courses of modulations), with respect to our results (as commented in our discussions, respectively in sections

4.1 and 4.2). Further, the ERP and PLV effects observed in our studies peaked around word onset, which coincides well with other studies that investigated the gesture-speech integration time course and reported a time window centred on word onset of -200 to + 120 ms (Habets et al., 2011; Obermeier & Gunter, 2014; Obermeier et al., 2011). In our third study (section 4.3), using fMRI, the audio and visual information was exactly the same in conditions X and Y, only that in the critical condition (Beat synchronous), we maintained the natural synchrony between beats and prosody, whilst in the contrast condition (Beat asynchronous), we artificially induced a lag between prosody and beats' apexes. Thus any difference between both conditions resulted from a synchrony effect between audio and video, but could not result from a difference in the amount of biological motion (present in both to the same degree). However, biological movements or point-light representations of biological movements were already shown to engage the STS (Grossman et al., 2004; Pelphrey et al., 2004), actions executed by humans (Thioux et al., 2008) and social visual cues (Nummenmaa & Calder, 2009; Allison, Puce & McCarthy, 2000 for reviews). However, the STS is also a classic multisensory site (Nath and Beauchamp, 2012; Calvert et al., 2000; Callan et al., 2004; Macaluso et al., 2004; Meyer et al., 2004; Campbell, 2008), in particular, audiovisual speech processing (Sekiyama et al., 2003; Calvert et al., 2000). Then, we cannot fully discard a contribution of biological movement perception in the BOLD responses modulation observed in the left MTG, but we believe that most of the contribution conforms more probably to multimodal speech

processing. In the present dissertation, I only reported three studies that were published in international scientific journals. But in parallel to thi work, we set various behavioral experiments to test for our attentional hypothesis. Unfortunately, none of these experiments led to conclusive results.

First, we adapted a mispronunciation detection task in which participants listened to short AV spoken sentences and had to detect as soon as possible words for which the first consonant had been mispronounced (leading to a non word). Based on our ERP results (section 4.1), we hypothesized that if beat gestures locally attracted the focus of attention of listeners, they may facilitate the processing of the corresponding utterance segment, and then, improve the detection of corresponding mispronounced (non)words. Our results from two ful experiments using different levels of masking noise revealed that non words accompanied with beat gesture were not significantly better detected than equivalent non words pronounced without beat (for a more detailed description of the experiments, see Annex 1). Second, based on So et al. (2012), we used the mnemonic effect of beats to test if their potential attentional effect was local or global during speech perception. We hypothesized that perceiving a speaker tending to produce many spontaneous beats during continuous speech may involve more the listener, engaging more his attention on speech content. If the attentional effect was local, we expected that listeners would recall more words pronounced with beats than others words pronounced without beat from the same AV clips. If the attentional effect was global, we

expected listeners to better recall words in AV speech condition than audio only condition, in general. However, the results did not reveal any difference of recall between words pronounced with beats than words pronounced without beat in the AV modality. Further, we found not significant difference of word recall between AV modality and A only modality of speech presentation. Thus, these results did not allow concluding on the attentional effect of beat gestures (for a more detailed description of the experiment, see Annex 2). Finally, in another behavioral study we tested the hypothesis developed in the third empirical study of this thesis (section 4.3). Namely, that beats play the role of visual prosody because of the robust temporal alignment between apexes and pitch peaks during speech production. Based on Holle et al. (2012) that used the syntactic parsing to index the role of beat gestures, we designed an experiment in which participants were presented with syntactically ambiguous sentences. These sentences had two possible interpretations but could be disambiguated following the auditory prosody (i.e. the placement of prosodic information like pauses and pitch peaks). We measured the interpretation preference in audio only condition, and compared it with the AV condition in which auditory prosody was removed and replaced by the equivalent beat placement (i.e. apexes corresponded with original pitch accents on critical words for disambiguation). In the audio only condition, we obtained very good results as listeners inversed their preference of interpretation of the sentences, according to the placement of auditory prosodic pauses. In contrast (and unfortunately for us), in the AV modality, the placement of the beat

gesture did not modulate the interpretation preference of sentences. These results suggested that, in this context of presentation, listeners did not took into account beat gestures to compensate missing auditory prosody, and consequently did not perceived them as relevant visual prosody. Thus, we could not conclude much more from those behavioral results (for a more detailed description of the experiment, see Annex 3). Nevertheless, further experimental procedures have to be set to make the link between neural correlates (Holle et al., 2012, our fMRI study) and behavioral modulation of beats on speech processing (So et al., 2012; Guellaï et al., 2014) to fully detangle between a clear beat effect or partially explained by biological motion.

3.6. Summary and final conclusions

The experiments presented in the present dissertation aimed to advance the knowledge of gesture-speech processing and their neural correlates, proposing alternative methodology both with less considered beats and new experimental procedures. The main conclusions of this thesis are the following:

1. Spontaneous beat gestures presented in continuous speech constitute a good model to investigate the neural correlates of gesture-speech integration.

2. The temporal alignment of beats with auditory prosody and the systematic order of presentation confer a

predictive value to beats. Then, listeners perceive beats as relevant visual information that attracts attention on associated elements in the utterance.

3. Consequently beats modulate the auditory processing of accompanied words early, possibly at a phonological stage.

4. Beats are visual prosody for their spatiotemporal relationship with auditory prosody, but also because they convey additional communicative information that is not present in simple equivalent moving discs.

In conclusion, our results conform to the original assumption stating that gestures and verbalization are part of the same language system as they showed that beats influenced auditory speech correlates during speech processing, and that when the natural relationship between both modalities was affected, modulations were found in some language related areas. In the future, I believe this alternative methodology will be exploited and improved, combining neuroimaging techniques with behavioral measures to investigate gesture processing in more natural conditions. Overall, the findings reported in the present thesis confirm the importance of non-verbal information in human spoken (and by extension, social) interactions. I believe that communicating is not just the verbalization of the mind's content, but speakers/listeners convey/decode information across different kinds of available channels (i.e. utterance, voice, hands and posture to name some), in

order to maximize the successful transmission/decoding of the message. Sometimes unconsciously or sometimes voluntarily exaggerated, the general posture is enough to convey the emotional value of the discourse or intentions of the speaker that may become hidden from the strict acoustic content. In turn, listeners are experts in interpreting this source of visual information and, they have to perpetually juggle with concomitant information coming from distinct modalities. Seen under this angle, it appears evident that future investigations will have to consider AV speech, not only as a richer version of the same audio only speech, but as a multifaceted communication format. I always think in a very common situation in which I look at someone greeting at a third person that I cannot see. Based only on his posture (gazing, smiling, direction) and hand shaking, I systematically turn the head in the same direction to reach the third person. I believe this reflects perfectly the high-level cognitive processes that non-verbal information engages (inferring intentions, direction, who responds by another posture), and why AV speech is more than the simple sum of A plus V information at low processing levels.

References

Aguiar-Conraria, L, Nuno Azevedo, N & Soares, M, A. (2008). Using wavelets to decompose the time–frequency effects of monetary policy. *Physica A*, 387, 2863-2878

Alibali, M. W., Kita, S., & Young, A. J. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language and Cognitive Processes*, *15*(6), 593–613.

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: role of the STS region. *Trends in Cognitive Sciences*, 4(7), 267–278.

Arnal, L. H., & Giraud, A. L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences, 16*(7), 390-398.

Arnal, L. H. Wyart, V., & Giraud, A.L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. Nature Neuroscience, 14(6), 797-801.

Astheimer, L. B., & Sanders, L. D. (2009). Listeners modulate temporally selective attention during natural speech processing. *Biological Psychology, 80*(1), 23-34.

Aydore, S., Pantazis, D., & Leahy, R. M. (2013). A note on the phase locking value and its properties. *Neuroimage, 74*, 231-244.

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559.

Bavelas, J. B., Chovil, N., Lawrie, D. A., & Wade, A. (1992). Interactive gestures. *Discourse Processes*, *15*(4), 469–489.

Beattie, G., & Coughlan, J. (1999). An experimental investigation of the role of iconic gestures in lexical access using the tip-of-the-tongue phenomenon. *British Journal of Psychology (London, England : 1953)*, *90 (Pt 1)*, 35–56.

Berens, P. (2009). CircStat: A MATLAB Toolbox for circular statistics. *Journal of Statistical Software*, 31(10).

Biau, E., & Soto-Faraco, S. (2013). Beat gestures modulate auditory integration in speech perception. *Brain and Language*, *124*(2), 143-152.

Biau, E., Torralba , M., Fuentemilla, L., de Diego Balaguer, R., & Soto-Faraco, S. (in press). Speaker's hand gestures modulate speech perception through phase resetting of ongoing neural oscillations. *Cortex.*

Brett, M., Anton, J-L., Valabregue, R., & Poline, J-B. Region of interest analysis using an SPM toolbox [abstract] Presented at the 8th International Conference on Functional Mapping of the Human Brain, June 2-6, 2002, Sendai, Japan. Available on CD-ROM in NeuroImage, Vol 16, No 2.

Busch, N. A., Dubois, J., & VanRullen, R. (2009). The phase of ongoing EEG oscillations predicts visual perception. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *29*(24), 7869-7876.

Callan, D. E., Jones, J. A., Callan, A. M., & Akahane-Yamada, R. (2004). Phonetic perceptual identification by native- and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory-auditory/orosensory internal models. *NeuroImage*, 22(3), 1182–94.

Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology : CB*, 10(11), 649–57.

Calvert, G. A., & Thesen, T. (2004). Multisensory integration: methodological approaches and emerging principles in the human brain. *Journal of Physiology, Paris*, *98*(1-3), 191–205.

Calvert, G. A., Spence, C., & Stein, B. E. (2004). The handbook of multisensory processing. Cambridge: MA:MIT Press.

Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 363(1493), 1001–10.

Carpenter, R. L., Mastergeorge, A. M., & Coggins, T. E. The acquisition of communicative intentions in infants eight to fifteen months of age. *Language and Speech, 26 (Pt 2)*, 101-116.

Casasanto, D., & Jasmin, K. (2010). Good and bad in the hands of politicians: Spontaneous gestures during positive and negative speech. *PloS One, 5*(7), e11805.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7).

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk-MacDonald effect: A phonetic representation within short-term memory. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology, 113*(4), 495-506.

Cook, S. W., Yip, T. K. Y., & Goldin-Meadow, S. (2012). Gestures, but not meaningless movements, lighten working memory load when explaining math. *Language and Cognitive Processes*, *27*(4), 594–610.

Cowan, N., Elliott, E. M., Scott Saults, J., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. A. (2005). On the capacity of attention: its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, *51*(1), 42–100.

Engle, R. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11*, 19-23.Davis, C., & Kim, J. (2006). Audio-visual speech perception off the top of the head. *Cognition*, *100*(3), 21–31.

De Lange, F. P., Spronk, M., Willems, R. M., Toni, I., & Bekkering, H. (2008). Complementary systems for understanding action intentions. *Current Biology : CB*, *18*(6), 454–7.

De Ruiter, J. (2000). The production of gesture and speech. *Language and Gesture*. Cambridge: Cambridge University Press, 284-311.

Dick, A. S., Mok, E. H., Raja Beharelle, A., Goldin-Meadow, S., & Small, S. L. (2014). Frontal and temporal contributions to understanding the iconic co-speech gestures that accompany speech. *Human Brain Mapping*, *35*(3), 900–17.

Dick, A. S., Goldin-Meadow, S., Hasson, U., Skipper, J. I., & Small, S. L. (2009). Co-speech gestures influence neural activity in brain regions associated with processing semantic information. *Human Brain Mapping*, 30(11), 3509–26.

Driver, J., & Spence, C. (2000). Multisensory perception: beyond modularity and convergence. *Current Biology : CB*, *10*(20).

Eckstein, K., & Friederici, A. D. (2005). Late interaction of syntactic and prosodic processes in sentence comprehension as

revealed by ERPs. *Brain Research. Cognitive Brain Research*, *25*(1), 130–43.

Eckstein, K., & Friederici, A. D. (2006). It's early: event-related potential evidence for initial interaction of syntax and prosody in speech comprehension. *Journal of Cognitive Neuroscience*, *18*(10), 1696–711.

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., & Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage*, *25*(4), 1325–35.

Engel, A. K., Fries, P., & Singer, W. (2001). Dynamic predictions: Oscillations and synchrony in top-down processing. *Nature Reviews. Neuroscience, 2*(10), 704-716.

Esteve-Gibert, N., & Prieto, P. (2014). Infants temporally coordinate gesture-speech combinations before they produce their first words. *Speech Communication, 57,* 301-316.

Feyereisen, P., & Lannoy, J.-D. de. (1991). *Gestures and Speech: Psychological Investigations*.

Fisher, NI. (1993). Statistical Analysis of Circular Data, *Cambridge University Press*.

Friston, K. J., Glaser, D. E., Henson, R. N. A., Kiebel, S., Phillips, C., & Ashburner, J. (2002). Classical and Bayesian inference in neuroimaging: applications. *NeuroImage, 16*(2), 484–512.

Frisch, S., Schleswesky, M. Saddy, D., Alpermann, A. (2002). The P600 as an indicator of syntactic ambiguity. *Cognition 85*: 83-92.

Fuentemilla, L., Marco-Pallares, J., & Grau, C. (2006). Modulation of spectral power and of phase resetting of EEG contributes differentially to the generation of auditory event-related potentials. *Neuroimage, 30*(3), 909-916.

Gillespie, M., James, A. N., Federmeier, K. D., & Watson, D. G. (2014). Verbal working memory predicts co-speech gesture: evidence from individual differences. *Cognition*, *132*(2), 174–80.

Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience, 15*(4), 511-517.

Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica, 66*(1-2), 113-126.

Glaser, Y. G., Martin, R. C., Van Dyke, J. A., Hamilton, A. C., & Tan, Y. (2013). Neural basis of semantic and syntactic

interference in sentence comprehension. *Brain and Language*, *126*(3), 314–26.

Goldin-Meadow, S., Alibali, M. W., & Church, R. B. (1993). Transitions in concept acquisition: using the hand to read the mind. *Psychological Review*, *100*(2), 279–97.

Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining math: gesturing lightens the load. *Psychological Science*, *12*(6), 516–22.

Goolkasian, P., & Foos, P. W. (2005). Bimodal format effects in working memory. *The American Journal of Psychology*, *118*(1), 61–77.

Grant, K. W., & Seitz, P.-F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*(3), 1197.

Gratton, G., & Coles, M. G. H. (1989). Generalization and evaluation of eye-movement correction procedures. *Journal of Psychophysiology, 3,* 14-16.

Greenberg, S. (1999). Speaking in shorthand, A syllable-centric perspective for understanding pronunciation variation. Speech Communication, 29, 159-176.

Greenberg, S., Carvey, H., Hitchcock, L. & Chang, S. (2003). Temporal properties of spontaneous speech—a syllable-centric perspective. *Journal of Phonetics*, 31, 465-485.

Griffiths, T. D., & Warren, J. D. (2002). The planum temporale as a computational hub. *Trends in Neurosciences*, *25*(7), 348–53.

Grossman, E. D., Blake, R., & Kim, C.-Y. (2004). Learning to see biological motion: brain activity parallels behavior. *Journal of Cognitive Neuroscience*, 16(9), 1669–79.

Guellaï, B., Langus, A., & Nespor, M. (2014). Prosody in the hands of the speaker. *Frontiers in Psychology*, *5*, 700.

Guthrie, D., & Buchwald, J. S. (1991). Significance testing of difference potentials. *Psychophysiology*, *28*(2), 240-244.

Habets, B., Kita, S., Shao, Z., Ozyurek, A., & Hagoort, P. (2011). The role of synchrony and ambiguity in speech-gesture integration during comprehension. *Journal of Cognitive Neuroscience*, *23*(8), 1845–54.

Hagoort, P. (2003). How the brain solves the binding problem for language: a neurocomputational model of syntactic processing. *NeuroImage*, *20 Suppl 1*, S18–29.

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends in Cognitive Sciences*, 9(9), 416–23.

Hasson, U., Malach, R., & Heeger, D. J. (2010). Reliability of cortical activity during natural stimulation. *Trends in Cognitive Sciences*, *14*(1), 40-48.

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*(1), 54–96.

Hein, G., Doehrmann, O., Müller, N. G., Kaiser, J., Muckli, L., & Naumer, M. J. (2007). Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *27*(30), 7881–7.

Herrington, J. D., Nymberg, C., & Schultz, R. T. (2011). Biological motion task performance predicts superior temporal sulcus activity. *Brain and Cognition*, 77(3), 372–81.

Hillyard, S. A., Hink, R. F., Schwent, V. L., & Picton, T. W. (1973). Electrical signs of selective attention in the human brain. *Science (New York, N.Y.)*, *182*(4108), 177-180.

Hinojosa, J. A., Martín-Loeches, M., & Rubia, F. J. (2001). Event-related potentials and semantics: an overview and an integrative proposal. *Brain and Language*, *78*(1), 128–39.

Hirai, M., Fukushima, H., & Hiraki, K. (2003). An event-related potentials study of biological motion perception in humans. *Neuroscience Letters*, *344*(1), 41-44.

Holle, H., & Gunter, T. C. (2007). The role of iconic gestures in speech disambiguation: ERP evidence. *Journal of Cognitive Neuroscience*, 19(7), 1175–92.

Holle, H., Gunter, T. C., Ruschemeyer, S. A., Hennenlotter, A., & Iacoboni, M. (2008). Neural correlates of the processing of co-speech gestures. *Neuroimage, 39*(4), 2010-2024.

Holle, H., Obleser, J., Rueschemeyer, S.-A., & Gunter, T. C. (2010). Integration of iconic gestures and speech in left superior temporal areas boosts speech comprehension under adverse listening conditions. *NeuroImage*, 49(1), 875–84.

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.

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*(3), 1028-1037.

Igualada, A., Bosch, L., & Prieto, P. (2015). Language development at 18 months is related to multimodal communicative strategies at 12 months. *Infant Behavior & Development*, *39*, 42–52.

Iverson, J. M., & Goldin-Meadow, S. (1998). Why people gesture when they speak. *Nature, 396*(6708), 228.

Iverson, J. M., & Goldin-Meadow, S. (2001). The resilience of gesture in talk: gesture in blind speakers and listeners. *Developmental Science, 4*(4), 416–422.

Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science, 16*(5), 367–71.

Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language, 89*(1), 253-260.

Kelly, S. D., Ward, S., Creigh, P., & Bartolotti, J. (2007). An intentional stance modulates the integration of gesture and speech during comprehension. *Brain and Language*, 101(3), 222–33.

Kelly, S. D., Creigh, P., & Bartolotti, J. (2010). Integrating speech and iconic gestures in a stroop-like task: Evidence for automatic processing. *Journal of Cognitive Neuroscience, 22*(4), 683-694.

Kelly, S. D., Ozyurek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science, 21*(2), 260-267.

Kendon, A. (1988). Sign Languages of Aboriginal Australia: Cultural, Semiotic and Communicative Perspectives. Cambridge: Cambridge University Press.

Kendon, A. (2004). Gesture: Visible Action as Utterance. Cambridge: Cambridge University Press.

Kita, S., van Gijn, I., & van der Hulst, H. (1998). Movement Phases in Signs and Co-speech Gestures, and Their Transcription by Human Coders. *Gesture and Sign Language in Human-Computer Interaction, (1371).*

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language, 48*(1), 16–32.

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*(3), 396-414.

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*(1), 373-383.

Krause, C. M., Porn, B., Lang, A. H., & Laine, M. (1997). Relative alpha desynchronization and synchronization during speech perception. *Brain Research.Cognitive Brain Research, 5*(4), 295-299.

Krauss, R. M.(1998). Why Do We Gesture When We Speak? *Current Directions in Psychological Science, 7(2),* 54-60.

Krauss, R., & Hadar, U. (1999). The role of speech-related arm/hand gestures in word retrieval. *Gesture, Speech, and Sign*.

Lachaux, J. P., Rodriguez, E., Martinerie, J., & Varela, F. J. (1999). Measuring phase synchrony in brain signals. *Human Brain Mapping, 8*(4), 194-208.

Lakatos, P., Chen, C.-M., O'Connell, M. N., Mills, A., & Schroeder, C. E. (2007). Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron, 53*(2), 279–92.

Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science (New York, N.Y.), 320*(5872), 110-113.

Leonard, T., Cummins, F. The temporal relation between beat gestures and speech. (2011). *Language and Cognitive Processes, 26,* 10.

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron, 54*(6), 1001-1010.

Macaluso, E., George, N., Dolan, R., Spence, C., & Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech: a PET study. *NeuroImage*, 21(2), 725–32.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164*(1), 177-190.

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.

Massaro, D.W. (1998). Perceiving Talking Faces: From Speech Perception to a Behavioral Principle. MIT Press: Cambridge, MA.

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

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

Mehler, J., Dommergues, J.Y., U. Frauenfelder, U. & Seguí, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, 20, 298–305.

Meyer, M., Steinhauer, K., Alter, K., Friederici, A. D., & von Cramon, D. Y. (2004). Brain activity varies with modulation of

dynamic pitch variance in sentence melody. *Brain and Language*, 89(2), 277–89.

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*(2), 133-137.

Murillo, E., & Belinchón, M. (2012). Gestural-vocal coordination: Longitudinal changes and predictive value on early lexical development. *Gesture*, *12*(1), 16–39.

Muthukumaraswamy, S. D., & Johnson, B. W. (2004). Primary motor cortex activation during action observation revealed by wavelet analysis of the EEG. *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology*, *115*(8), 1760–6.

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), 1-21.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Alho, K. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature, 385*(6615), 432-434.

Näätänen, R. (1982). Processing negativity: An evoked-potential reflection of selective attention. *Psychological Bulletin, 92*(3), 605-640.

Nagels, A., Chatterjee, A., Kircher, T., & Straube, B. (2013). The role of semantic abstractness and perceptual category in processing speech accompanied by gestures. *Frontiers in Behavioral Neuroscience, 7*, 181.

Noesselt, T., Rieger, J. W., Schoenfeld, M. A., Kanowski, M., Hinrichs, H., Heinze, H.-J., & Driver, J. (2007). Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 27*(42), 11431–41.

Nummenmaa, L., & Calder, A. J. (2009). Neural mechanisms of social attention. *Trends in Cognitive Sciences*, 13(3), 135–43.

Obermeier, C., Holle, H., & Gunter, T. C. (2011). What iconic gesture fragments reveal about gesture-speech integration: when synchrony is lost, memory can help. *Journal of Cognitive Neuroscience, 23*(7), 1648–63.

Obermeier, C., & Gunter, T. C. (2014). Multisensory Integration: The Case of a Time Window of Gesture-Speech Integration. *Journal of Cognitive Neuroscience*, 1–16.

Obleser, J., Meyer, L., & Friederici, A. D. (2011). Dynamic assignment of neural resources in auditory comprehension of complex sentences. *NeuroImage*, *56*(4), 2310–20.

Özyürek, A. (2014). Hearing and seeing meaning in speech and gesture: insights from brain and behaviour. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *369*(1651), 20130296.

Partan, S. R. (2013). Ten unanswered questions in multimodal communication. *Behavioral Ecology and Sociobiology*, *67*, 1523–1539.

Pavlova, M. A. (2012). Biological motion processing as a hallmark of social cognition. *Cerebral Cortex (New York, N.Y. : 1991)*, *22*(5), 981–95.

Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology*, *3*, 320.

Pelphrey, K. A., Morris, J. P., & McCarthy, G. (2004). Grasping the intentions of others: the perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. *Journal of Cognitive Neuroscience*, 16(10), 1706–16.

Pelphrey, K. A., Morris, J. P., Michelich, C. R., Allison, T., & McCarthy, G. (2005). Functional anatomy of biological motion

perception in posterior temporal cortex: an FMRI study of eye, mouth and hand movements. *Cerebral Cortex (New York, N.Y. : 1991)*, *15*(12), 1866–76.

Pfurtscheller, G., Neuper, C., & Krausz, G. (2000). Functional dissociation of lower and upper frequency mu rhythms in relation to voluntary limb movement. *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology*, *111*(10), 1873–9.

Picton, T. W., & Hillyard, S. A. (1974). Human auditory evoked potentials. II. effects of attention. *Electroencephalography and Clinical Neurophysiology, 36*(2), 191-199.

Pilling, M. (2009). Auditory event-related potentials (ERPs) in audiovisual speech perception. *Journal of Speech, Language, and Hearing Research : JSLHR, 52*(4), 1073-1081.

Ping, R., & Goldin-Meadow, S. (2010). Gesturing saves cognitive resources when talking about nonpresent objects. *Cognitive Science*, *34*(4), 602–19.

Quandt, L. C., Marshall, P. J., Shipley, T. F., Beilock, S. L., & Goldin-Meadow, S. (2012). Sensitivity of alpha and beta oscillations to sensorimotor characteristics of action: an EEG study of action production and gesture observation. *Neuropsychologia*, *50*(12), 2745–51.

Rauscher, F. H., Krauss, R. M., & Chen, Y. (1996). Gesture, speech, and lexical access:. The Role of Lexical Movements in Speech Production. *Psychological Science, 7*(4), 226–231.

Rohenkohl, G., & Nobre, A. C. (2011). Alpha oscillations related to anticipatory attention follow temporal expectations. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 31*(40), 14076-14084.

Schroeder, C. E., & Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in Neurosciences, 32*(1), 9-18.

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., & Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Sciences, 12*(3), 106-113.

Sebastián-Gallés, N., Martí, M.A., Carreiras, M., Cuetos, F. LEXESP: Léxico informatizado del Español, Edicions Universitat de Barcelona, Barcelona, 2000.

Sekiyama, K., Kanno, I., Miura, S., & Sugita, Y. (2003). Auditory-visual speech perception examined by fMRI and PET. *Neuroscience Research, 47*(3), 277–87.

Shah, A. S., Bressler, S. L., Knuth, K. H., Ding, M., Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2004). Neural dynamics and the fundamental mechanisms of event-related brain potentials. *Cerebral Cortex (New York, N.Y.: 1991), 14*(5), 476-483.

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(3), 260–77.

So, W. C., Chen-Hui, C. S., Wei-Shan, J. L. (2012). Mnemonic effect of iconic gesture and beatgesture in adults and children: Is meaning in gesture important for memory recall? *Language and Cognitive Processes*, 27(5), 665-681.

Spence, C., & Driver, J. (2004). *Crossmodal Space and Crossmodal Attention*.

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience*, *19*(12), 1964-1973.

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses. Cognitive neuroscience.*

Straube, B., Meyer, L., Green, A., & Kircher, T. (2014). Semantic relation vs. surprise: the differential effects of related and unrelated co-verbal gestures on neural encoding and subsequent recognition. *Brain Research*, *1567*, 42–56.

Streltsova, A., Berchio, C., Gallese, V., & Umilta', M. A. (2010). Time course and specificity of sensory-motor alpha modulation during the observation of hand motor acts and gestures: A high density EEG study. *Experimental Brain Research*, *205*(3), 363-373.

Sumby, W., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*(2), 212-215.

Thioux, M., Gazzola, V., & Keysers, C. (2008). Action understanding: how, what and why. *Current Biology : CB*, 18(10), R431–4.

Thut, G., Nietzel, A., Brandt, S. A., & Pascual-Leone, A. (2006). Alpha-band electroencephalographic activity over occipital cortex indexes visuospatial attention bias and predicts visual target detection. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 26*(37), 9494-9502.

Treffner, P., Peter, M., & Kleidon, M. (2008). Gestures and Phases: The Dynamics of Speech-Hand Communication. *Ecological Psychology*, *20*(1), 32–64.

Uchiyama, Y., Toyoda, H., Honda, M., Yoshida, H., Kochiyama, T., Ebe, K., & Sadato, N. (2008). Functional segregation of the inferior frontal gyrus for syntactic processes: a functional magnetic-resonance imaging study. *Neuroscience Research*, *61*(3), 309–18.

Van de Meerendonk, N., Kolk, H. H. J., Vissers, C. T. W. M., & Chwilla, D. J. (2010). Monitoring in language perception: mild and strong conflicts elicit different ERP patterns. *Journal of Cognitive Neuroscience, 22*(1), 67–82.

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*(4), 1181-1186.

Vatikiotis-Bateson, E., & Yehia, H. (1996). Physiological modeling of facial motion during speech. *Trans. Tech. Com. Psycho. Physio. Acoust*.

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

Wang, L., & Chu, M. (2013). The role of beat gesture and pitch accent in semantic processing: An ERP study. *Neuropsychologia*, *51*(13), 2847-2855.

Willems, R. M., Ozyürek, A., & Hagoort, P. (2007). When language meets action: the neural integration of gesture and speech. *Cerebral Cortex (New York, N.Y. : 1991)*, 17(10), 2322–33.

Willems, R. M., Ozyurek, A., & Hagoort, P. (2009). Differential roles for left inferior frontal and superior temporal cortex in multimodal integration of action and language. *NeuroImage*, *47*(4), 1992-2004.

Wu, Y. C., & Coulson, S. (2007). Iconic gestures prime related concepts: An ERP study. *Psychonomic Bulletin & Review*, *14*(1), 57-63.

Wu, Y. C., & Coulson, S. (2010). Gestures modulate speech processing early in utterances. *Neuroreport, 21*(7), 522-526.

Wu, Z., & Gros-Louis, J. (2014). Infants' prelinguistic communicative acts and maternal responses: Relations to linguistic development. *First Language*, *34*(1), 72–90.

Yasinnik, Y. (2004). The timing of speech-accompanying gestures with respect to prosody. *Proceedings of From Sound to Sense,MIT*. MIT.

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, *26*(1-2), 23–43.

Zatorre, R. J., & Gandour, J. T. (2008). Neural specializations for speech and pitch: moving beyond the dichotomies. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *363*(1493), 1087–104.

ANNEX 1.

The potential local attentional effect of beat gestures on the corresponding auditory segment.

1. Introduction

Spontaneous beat gestures arise naturally as part of the situation of communication. Beats are rapid biphasic movements, and even though they do not present a discernible meaning, they seem to engage complex cognitive processes to be correctly interpretated (McNeill, 1992; So et al., 2012; Holle et al., 2012; Guellaï, Langus & Nespor, 2014). On the production side, the speaker uses beats to accompany relevant information and structure his narrative and contrast the different levels (mostly metanarrative and paranarrative levels, McNeill, 1992). On the listener's side, although beats have received little attention, there is evidence that they play a role in the perceived prominence of a word in a spoken utterance (Yasinnik, Renwick and Shattuck-Hufnagel, 2004; Krahmer & Swerts, 2007; Treffner, Peter & Kleidon, 2008; Guellaï, Langus & Nespor, 2014). Beats can therefore be viewed as temporal highlighters, and structure related on both sides as roles get reversed all the time during the conversation. In our previous study, (Biau & Soto-Faraco, 2013; section 4.1 in the present dissertation) we suggested that gestures emphasize the focus of attention locally on the affiliated utterance while perceiving speech. Using the ERP technique, we investigated the time course of beat-speech

integration during natural continuous speech perception. Compared to the auditory alone condition, beats elicited a positive shift at an early attentional stage as well as at the P2 time window (corresponding to auditory processing and more precisely to the N1-P2 component). This modulation was interpreted as local attentional highlighter affecting early sensory/phonological stages of processing (Hillyard et al., 1973; Näätänen, 1982; Picton & Hillyard, 1974; Astheimer & Sanders, 2009). These results suggested that listeners allocated attention onto the words that are uttered specifically with a beat gesture because they are marked as relevant by the speaker (and they know it).

Scope of the present study:

The object of this study is to test the local attentional effect of beats and their potential facilitation on the processing of corresponding auditory speech segments. In an ERP study, Astheimer et al. (2009) evidenced a selective processing of sounds at relevant timings (word onsets) based on acoustic information. The authors found significant modulations of N1-P2 component around word onsets, respect to other segments in the auditory speech. They suggested that listeners allocate more attention on relevant acoustic information to facilitate following auditory processing during speech perception. Based on these results and Biau & Soto-Faraco (2013), we hypothesized that gestures can be considered as visual linguistic information signaling relevant acoustic segment for the allocation of attentional resources, potentially at affiliated word

onsets. Their temporal alignment with prosody (co occurrence of beats' apexes and pitch accents) and the systematic order of presentation (gestures are initiated 200ms before corresponding word onset) suggest that listeners perceive beats as robust markers of following relevant acoustic cues, having interest to allocate enough local attention on those visual signals to improve auditory speech processing facilitation. Then, the acoustic relevance of word onsets accompanied by a beats might be enhanced, respect to equivalent words pronounced without beat in audiovisual speech.

To test this hypothesis, we adapted a mispronunciation detection task (Cole, 1973). A mispronunciation is defined as a change of a segmental feature (e.g. the first syllable of the word), leading to the transformation of a word into a non-word. Listeners are asked to detect when those small phonetic changes occur. If beats indeed work as highlighters, we expect them to increase the listener's attention on the related word onsets and facilitate the mispronunciation (MP) detection. At behavioral levels, this facilitation may be evidenced by shorter reaction times, and greater correct response rates (listeners may detect more MP and miss them less).

2. Material and method

2.1. Material

We created short audiovisual clips presenting a speaker uttering isolated sentences (one sentence per clip). A native Italian speaker

was filmed using a digital camera at a rate of 50 frames/sec while pronouncing sentences of about 10 words in Spanish (duration 5.5ec). Each sentence was recorded twice. In one version, he had to mispronounce (MP) the first syllable of a critical word, without gesturing (MP+G- condition). In the second version, he was asked to pronounce correctly the whole sentence, producing a beat in synchrony with the critical word (MP-G+ condition). The same video was later synchronized with the sentences containing a mispronunciation to create an additional condition in which there was a mispronunciation accompanied by a beat (MP+G+ condition). The critical contrast of the present experiment relied on the comparison of performance between these two conditions MP+G+ and MP-G+. To avoid any possible risk of strategy (due for example to the higher probability of beat occurrence in isolated sentences respect to natural situations of conversations), we added "filler" conditions for which sentences appeared with MP and G desynchronized (MP+G+des condition); or with no G nor MP (MP-G- condition). Finally a condition for which sentences contained a gesture but no mispronunciation was added to complete all the possible situations (MP-G+ condition). A total of 210 sentences were recorded. In addition, we count 10 training sentences (two of each type). The conditions are summarized in the Table 1 below.

As the literature pointed out a certain number of factors on mispronunciation detection scores and RTs, which are as many possible biases in this study, we controlled for word frequency, word position in the sentences, as well as the position of MP in the critical word (i.e. the first syllable) and speaker's accent (Italian

speaker) across conditions (for more details, see Schmid et al., 1999).

Condition	Mispronounciation (MP)	Gesture (G)	Number of sentences
of interest	+	+	25
control 1	+	-	25
control 2	-	+	45
control 3	-	-	25
fillers	+	desync	90

Table 1. Summary of the experimental material, in five conditions.

At last, phonetic parameters were treated carefully. The error should be of one feature only (manner, place, voicing, nasality). Because some changes are easier to detect (e.g. place rather than voicing), we maintained the proportions of those changes across the four conditions. Finally, in the MP+G+ condition, beats and critical non-word were synchronized, by aligning the frame containing the apex of the gesture with the pitch peak of the accented syllable (F0) according to previous literature (Yasinnik, Renwick and Shattuck-Hufnagel, 2004).

2.2. Procedure

25 Participants (native Spanish speakers) were told that they were about to hear an Italian speaker sometimes making a mispronunciation in sentences (not all the time, and only once a

sentence). Participants were asked to press as fast as possible the space bar whenever they heard a mispronunciation. The experimental interface and recording of the results was made using E-prime. Each trial was displayed as followed:

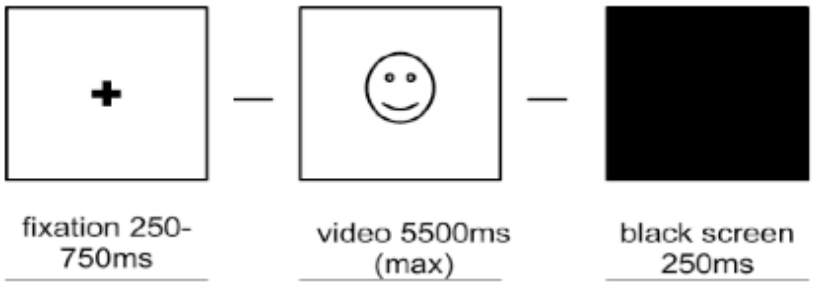

Fig.1: Linear display of each trial.

The fixation duration varied from 250ms to 750ms to maintain the participant's attention and keep him/her ready when the video began. In order to give the subject a feedback, when he pressed Space bar, the video stopped and jumped to the next trial. The trials were divided into five blocks of approximately four minutes each, allowing the participants to take the break time they judged acceptable in between. All these parameters taken in account, the experiment should last no more than 30 minutes. Performances were recorded in terms of response type (response/no response) and reaction time, starting from the onset of each target word (previously manually extracted). False alarms, including early responses, and late responses (over 1.5sec, following Schmidt et al. (1999)), were excluded before statistical analysis.

3. Results

3.1. Correct response rates

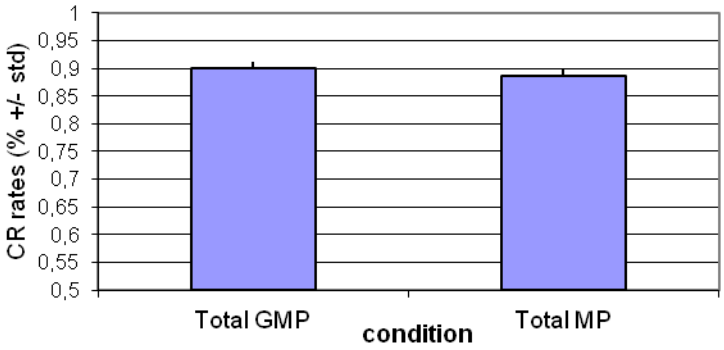

Fig.2: Correct response rates (% +/- std) for both MP+G+ (left) and MP+G- (right) conditions.

Results revealed no significant differences of correct responses rates between the two conditions (t-test p-values > 0.05). This suggested that participants did not better detect mispronunciations when they were accompanied with a beat gesture as compared to alone.

3.2. Reaction times

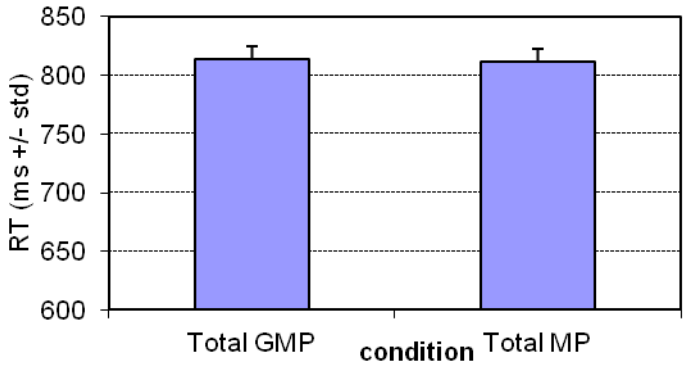

Fig.3: Mean reaction times (ms +/- std) for both MP+G+ (left) and MP+G- (right) conditions.

Results revealed no significant differences of reaction times between the two conditions (t-test p-values > 0.05). This suggested that participants were not faster at detecting mispronunciations when they were accompanied with a beat gesture as compared to alone.

4. Discussion

The present study aimed at investigating the possible local attentional effect of beat gestures on the processing of acoustic relevant part of associated utterance (word onsets) during speech perception. To do so, we adapted a mispronunciation detection task in an audiovisual version for which non-words came with a beat gesture or not. We hypothesized that beats may increase the natural local attention allocated at word onsets during speech perception, as they are robust visual linguistic information. At behavioral levels, we expected that this effect might be reflected by an increase of mispronunciation detection and a decrease of reaction times of listeners, indexing a facilitation of speech processing. Results were not conclusive as they showed no effects of beat gestures on MP detection performance. We saw different reasons that may explain the null effect. First, listeners did not rely on visual information to perform the task. As it was an auditory task, participants did not pay attention to additional visual modality, even if they were told to attend to the stimuli as if they were watching TV (they were explicitly asked to not close the eyes during the

entire procedure). But in this particular experimental context, visual information was irrelevant and beats were maybe underestimated. Second, even if listeners perceived beats, the fact to introduce conditions in which beats and MP were voluntary unrelated (i.e. MP+G+ desynchronized condition) may have confused them. Consequently, they decided that visual information was not helpful (or else distracting) to perform the task. Finally, the absence of effect could also be explained by the behavioral measure, which was not be fine enough, or adapted, to reflect the integration between speech and gestures. Alternatively, the task was too easy and participants' performances reached a ceiling effect in both conditions. However, we replicated the same experiment in noisy conditions (we added a white noise in all the clips) and obtained the exact same patterns of performances (with an additional effect of noise that decreased performances in general). We focused on the accompanied word onsets processing but as beats' apexes are temporally aligned with pitch peaks, it may be an alternative way to investigate the beat-gesture integration.

ANNEX 2.

Do beats have a mnemonic effect on continuous speech processing? Behavioral study.

(This project was part of my 4-months internship, realized at the Psychology department of the university of Hull, UK, under the co-supervision of Henning Holle)

1. Scope of the study

The main topic of this study was to use the memory recall to index the possible affect of perceiving beat gestures produced in a natural and legitimate speech context. We hypothesized that in a continuous AV speech, spontaneous beats influence how listeners select relevant information by underlining the accompanied segments. If beats help to form a more coherent global representation of speech (McNeill, 1992), we assumed that listeners may encode better relevant information and improve memory recall.

First, we wanted to investigate if beat gestures, as natural visual prosodic information, are special or if simple visual discs following comparable trajectories affect speech encoding in a similar manner. When perceiving natural beats, listeners can extract intentions of the speaker to emphasize important parts of speech, as they also gesture when speaking. Seeing someone gesturing may involve more the listener because he can infer some meta-cognitive aspects of the body posture as emotions. We hypothesized that mnesic performances may be improved for speech accompanied by natural

beats compared to artificial moving dots. Second, we wanted to test if the possible effect of beat gestures on speech encoding is local or global. If beat gestures have a local effect, the synchrony between the gesture and the corresponding speech segment has to be maintained to attract attention at correct moments. Then listeners may have a better mnesic trace when speech and gestures are synchronized than desynchronized. In contrast, if beat gestures have a global attentional effect, the simple fact to attend to someone gesturing is enough to improve attention on general speech content. If so, the asynchrony between beats and speech should not affect memory performances.

To do so, we designed an experimental paradigm that allowed measuring the memory recall of participants soon after AV speech perception. We adapted the word recognition task described by Roediger & McDermott (1995) called the DRM paradigm. In this task, participants were first presented with lists of words, semantically related. Then, in the recognition task, they were represented with new lists of words and they had to say if they heard or not each of these words in the first presentation. The interest of this recognition task was that, as words of lists are semantically related, one can induce false recognitions by adding new related words in the second presentation. Then, it increases the difficulty of the task and allows avoiding ceiling effect in memory performances. In our experiment, 30 participants were presented with short AV clips. Soon after the clip end, they were asked to answer if they heard or not a target word in the previous speech. They were instructed to respond “yes” when they were sure to

recognize words. When they were guessing, they had to respond "no".

2. Material and procedure

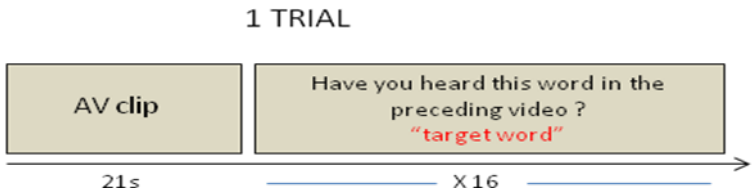

Fig.1: Experimental procedure.

1) AV Clips:

We created 20 AV clips of 19s (+2s of fade in, fade out) each, in which the Prime Minister David Cameron answers questions from the opposition at the House of the Commons. Each video deals with only a particular main topic (for example horse meat scandal, accident of helicopter in London, etc). All visual information is available, except for the head that is artificially hidden when the speaker is visible during speech. The same point of view of the camera has also been selected to create equivalent clips. For each clip, we created 5 versions:

- Audio: speech + a picture of the speaker.
- Beat_sync: speech + speaker, synchronized.
- Beat_async: speech + speaker, asynchrony (1s of lag, A>V).
- Disc_sync: speech + discs, synchronized.
- Disc_async: speech + discs, asynchrony (1s of lag, A>V).

2) Recognition task:

We created one list of 16 words for each video clip, as following.

- 4 OG: Old words pronounced with a Gesture during speech.
- 4 ONG: Old words pronounced with No Gesture during speech.
- 4 NR: New words Related to the topic of the clip, not pronounced during speech.
- 4 NU: New words Unrelated to the topic of the clip, not pronounced during speech.

3) Measure of the memory quality:

	response “YES”	Response “No”
OG + ON	Hit (H)	miss
NR + NU	false alarm (FA)	Correct rejection

To evaluate the accuracy of word recognition we used 2 characteristics of the mnesic trace: 1) the proportion of old words (OG+ON) effectively recognized as previously heard (Hits), that reflects the quality of speech encoding (“I know what I’ve heard in the previous clip”). 2) the proportion of new words (NR+NU) correctly rejected, that reflects how participants compare the mnesic trace of speech encoding to new inputs and decide what is new or not (I know I haven’t heard it before). Implicitly, we used 1-Correct rejection to measure the proportion of FA, to calculate the d-prime and evaluate the global accuracy of the task recognition. The d-prime for each condition was normalized with the d-prime of the audio condition (base). We applied a 2x2 ANOVA with factors Condition (Beats or Discs) and Synchrony (Sync or Async).

3. Results

3.1. General performances across conditions and word categories

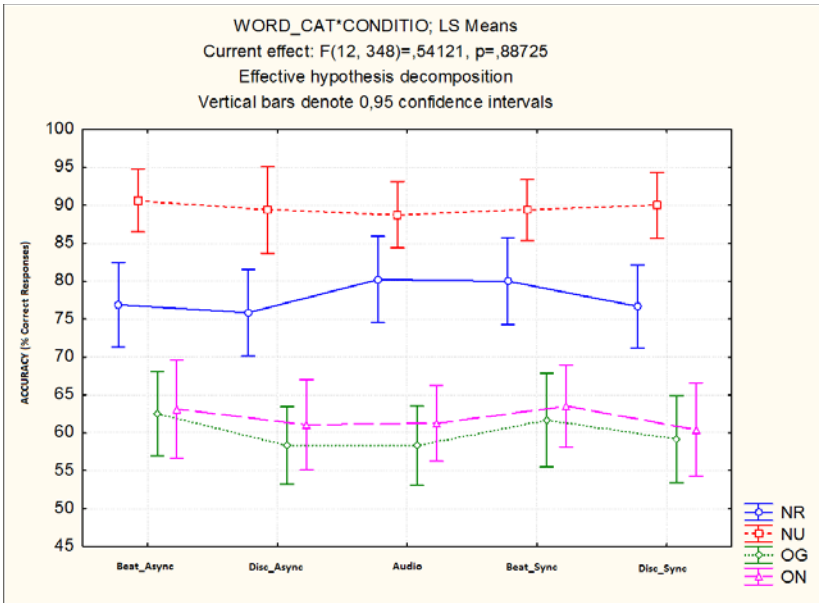

Fig.2: Mean accuracy (correct response rates, % +/- std) per word categories across the five experimental conditions (Beat_async, Disc_async, Audio, Beat_sync and Disc_sync): New related words (blue line), new unrelated words (red line), old gesture words (red dashed line) and old no gesture words (pink dashed line).

Results show that the accuracy was only affected by the word category across conditions ($F(3, 87)=40, 32$; $p<0,0001$). Participants were significantly better at rejecting new related (NR) or unrelated (NU) words than recognizing old words with (OG) or no (NG) gesture. Results showed no significant effect of conditions on word recall ($F(4, 116)=1, 23$; $p=0, 30$) nor interaction between word categories and conditions ($F(12, 348)=0, 5$; $p=0, 30$), suggesting that participants did not recall better words when speech was

encoded with gestures than when encoded in audio only modality (see fig.2).

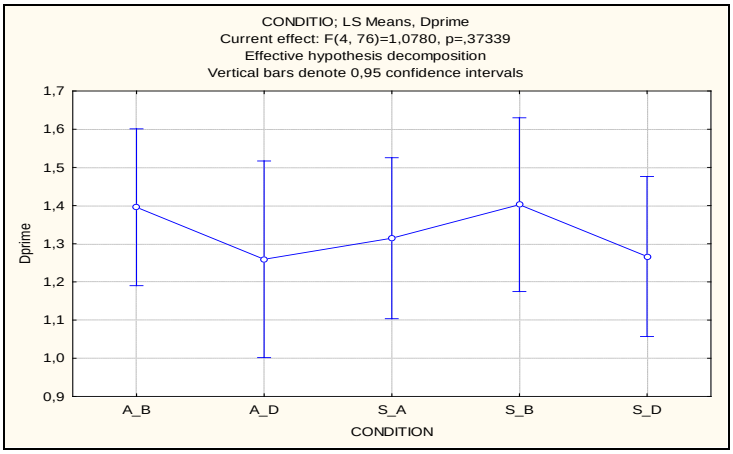

Fig.3: D-prime +/- std across the five conditions (Beat_async, Disc_async, Audio, Beat_sync and Disc_sync).

Results show that the d-prime values were not significantly different across conditions (F(4, 76)=1, 08; p=0, 37), suggesting that presence of additional visual information (beats or discs) or not (audio only modality) did not influence encoding during speech processing (see fig.3).

2) Word recall was affected only by the type of visual information accompanying continuous speech:

Results showed a significant effect of visual information (condition) on mnesic performances as the d-prime was higher when speech was accompanied by natural beat gestures as compared to moving discs (F(3,58)=4,07; p=0,04), irrespective from synchrony (see fig.4). In contrast, the asynchrony between speech and visual information did not affect the performances in both conditions

(F(1,58)<1; p=0,93). Finally, there was no interaction between Synchrony and Visual information (F(1,58)<1; p=0,82).

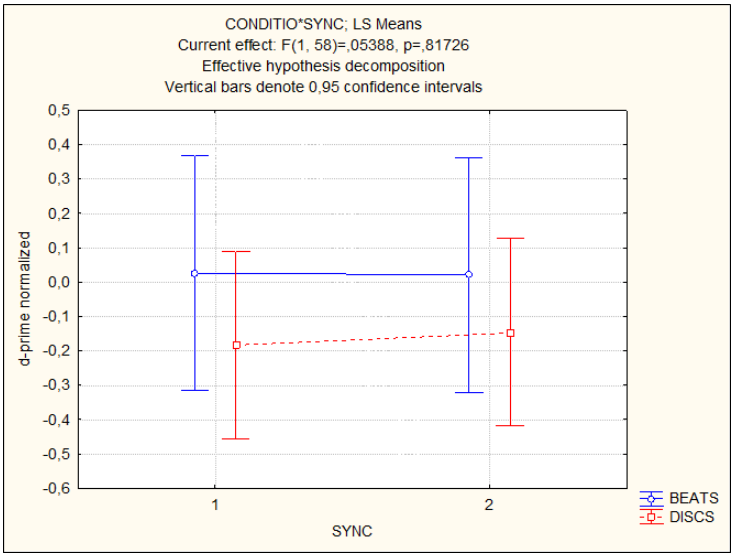

Fig.4: D-prime values according to the synchrony between audio and visual information (asynchronous 1, synchronous 2), and the type of visual information: beats (blue line) or discs (red line).

4. Discussion

In the present study, we wanted to investigate the possible effect of accompanying beat gestures on continuous speech encoding by mean of a word recall task. Through this behavioral task, we aimed at testing two hypotheses. First hypothesis: we hypothesized that mnesic performances may be improved for speech accompanied by natural beats compared to artificial moving dots. Beats, as part of speaker may probably convey additional communicative intention and engage cognitive processes of interpretation. Results showed that, independently from synchrony or asynchrony between audio

and video, word recall was greater when speech came with beats than discs. This suggests effectively that beat gestures were differently integrated with speech as compared to disc following the exact same spatiotemporal trajectories. Thus beats may engage additional cognitive process related to communicative posture interpretation or emotional for example. However, performances with beats were not significantly different from audio condition. Even if it is in line with previous reports (see for example Guellaï, Langus & Nespor, 2014), the present results do not allow concluding if listeners effectively relied on beats during speech perception as visual linguistic information, or discs significantly decreased mnesic performances respect to beat conditions. In this case, discs disturbed listeners that allocated too much attention on them instead of content during speech perception.

In the second hypothesis, we wanted to test if the possible effect of beat gestures on speech encoding was local or global. Results showed no effect of synchrony in performances between beat_sync and beat_async conditions. They suggested that the effect of gestures in this context is global and the simple fact to see someone gesturing during speech is enough to maintain the attention on speech content. Then, if congruent beats did not increase speech encoding respect to audio only modality, it was not surprising that incongruent beats did not affected neither speech encoding if the impact of gesture is global. Alternative, the asynchrony was large enough to be voluntary perceived but in passive speech perception, the brain was enough flexible to maintain the temporal relationship between a beat and its targeted word. That may explains why half of

participants did not actually reported asynchrony. Or else, in the beat_async condition, when we desynchronized the audio and the video, we actually targeted new words with a gesture (because of a sliding effect), and the general synchrony remains more or less the same as in the beat_sync condition.

Finally, the absence of effect between synchrony and asynchrony in the disc conditions may be explained by the fact that the potential disturbing effect of moving discs was already strong enough in the synchrony condition that it reached already its ceiling effect on speech modulation. Then a simple asynchrony did not brought significant additional effect on speech processing in the disc_asynch condition. In general, the present null results were not clear enough to conclude on the potential local or global effect of beat gestures on auditory speech processing. The difference between beats and discs also need further investigations as here, results suggest in the present experimental context only a marginal disturbing effect of discs rather than a facilitating effect of beats.

ANNEX 3.

The effect of auditory prosody (pauses) compared to visual prosody (beat gestures) on sentence disambiguation.

Lauren Fromont, Emmanuel Biau and Salvador Soto-Faraco.

(This study was part of the master' project of Lauren Fromont that I co-supervised with Salvador Soto-Faraco).

1. Introduction

The observations made from the literature led us to two major premises. First, beats and prosodic are temporally related, and their congruency seems to lead to a better perception compared to a unimodal situation. Second, one function of prosody is to facilitate syntactic parsing. Given those, we wanted to assess the question of whether beats share a similar functional role with prosody. There might be no strong gain of beat gestures when prosodic information can be used. However, if prosodic cues are insufficient to resolve the ambiguity, may beat gestures compensate for them and maintain disambiguation? We hypothesized that gestures play a role in grouping of intonational phrases: we thus expected them to help perceivers to modulate their interpretation when the prosody was absent or conflicting.

To assess that question, we used structurally ambiguous sentences where prosody was sufficient to resolve the ambiguity. First we designed an Audio experiment to test if auditory prosody alone could disambiguate sentences and modulate listeners interpretation according to the placement of acoustic cues. Second, we assessed the role of beat gestures in a mirrored audiovisual experiment, allowing us to compare the influence of the beat placement in the sentence, with the influence of the acoustic prosodic cues. To do so, we removed prosodic cues from our auditory material and added beat gestures associated to the critical words for disambiguation of sentences. We expected gestures to compensate the lack of prosodic information and help disambiguate the sentences in a similar fashion. Both at behavioral and neural levels, we expected comparable modulations of acoustic prosodic cues and beat gestures.

2. Material and Method

2.1. Participants

40 native Spanish speakers (11 males; mean age: 24 ± 3, 4 years old) volunteered to the experiment after giving informed consent (20 in the A version and 20 in the AV version of the experiment). They received monetary compensation for their participation. All participants had normal or corrected-to-normal vision and no one reported known hearing deficit. One participant was excluded from analyses.

2.2. Material

2.2.1 Audio only modality

We first generated 100 experimental Spanish sentences containing closure-related ambiguity, based on de la Cruz Pavia (2010), with the following structure pattern:

(1) María encontró al amigo del niño que reía.
Maria met the friend of the child who was laughing.

In order to reverse the natural preference of Spanish listeners for high attachment, and enhance low attachment preference, length of the RCs was kept shorter than four syllables according to Fodor's Prosodic Hypothesis (Fodor, 1998). Nouns of phrases (NP) were between three and five syllables including the determinant to conserve a comparable rhythmicity across sentences. Additionally, we controlled for the frequency of NP1 and NP2 using *Busca palabra* (Davis & Perea, 2005). Second, we created a semantic context for each sentence to enhance the naturalness and implement a prosodic rhythm (breaks are marked as #).

(2) El jueves santo, María no quería salir, porque estaba lloviendo mucho! # Pero como no tenía comida, # no tuvo más remedio que ir al mercado a comprar jamón, naranjas y

cebollas. # Allí en el Mercado, **# María encontró al amigo del niño que reía**.

Last Thursday, María did not want to go out because it was raining. But, as she did not have anything more for cooking, she had to go to the market to buy some ham, oranges and onions. There in the market, ***# María met the friend of the kid who was laughing.***

Two versions of each sentence were recorded using a unidirectional microphone MK600, Sennheiser and the Audacity software (version 2.0.3), at a sample frequency of 24,000Hz. A female native speaker of standard Castellan was asked to read both the contexts sentences in a natural fashion, with a break after either NP1 or NP2, as shown in (2):

(3) María encontró al amigo # del niño que reía.
Condition NP1

(4) María encontró al amigo del niño # que reía.
Condition NP2

All the subsequent acoustic manipulations were made using the *Praat* software. Stimuli were examined acoustically and visually to insure there was no significant differences in intonation between sentences. Durations of all breaks were set constant at 200ms, both in contexts and experimental sentences. Additionally, we created a third condition where the prosody was non-informative. We applied the cross-

splicing method. The signal was cut during the transition between [l] from “del” and the first consonant of NP1. The first segment of sentence (4) was then cross-spliced with the second segment of sentence (3), generating a sentence with no prosodic break (*Condition NP*).

2.2.2 Audio only modality

To create the audiovisual version (AV) of the experiment, we used the same auditory material as in the A version, but we removed all the acoustic prosodic cues to generate sentences with no pause neither after NP1 or NP2. An actor was recorded while faking telling the prosody less version each of sentences. She was asked to produce spontaneous beats during the context of the clip, and to synchronize a beat with the critical NP1 or NP2, or stay still during the experimental final sentence of each story. After the recording session, all the apexes were manually adjusted with the pitch peaks of either NP1 or NP2 accented syllable to ensure the correct synchrony between auditory speech envelope and beats. In total, the AV version of the experiment contained 3 conditions equivalent to the one of the A only version: NP1 (beat synchronized with the first noun of the experimental sentence), NP2 (beat synchronized with the second noun of the experimental sentence) or NP (no gesture at all). Comparisons are summarized in the following:

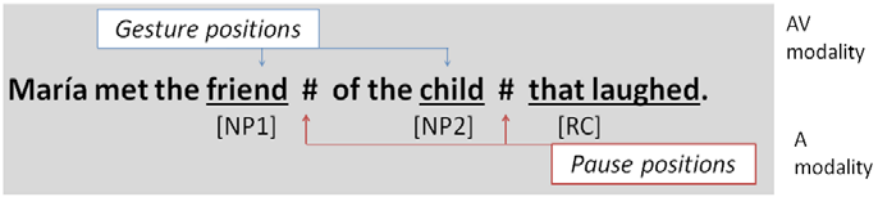

Fig.1: Equivalence of NP1, NP2 and NP conditions in both the A and AV versions of the experiments.

The A and AV versions of the experiment were presented using E-Prim2 pro software. In total, a procedure contained 4 blocks of 25 ambiguous sentences (33 from each NP1, NP2 and NP conditions), separated by 5-minutes resting breaks. Finally, we evaluated the listener's interpretation of ambiguous sentences by reporting the proportion of low attachment across conditions.

2.3. Procedure

Participants were comfortably installed in a sound attenuated room, sitting approximately at 60cm from the screen. Each trial began with a 500 ms white fixation cross displayed on a black screen. The cross would turn read when the audio stimulus started. The participants were presented with the lead-in context followed by the experimental sentence. When the audio ended, participants were asked to decide between two possible interpretations of the last sentence through a 2 forced-choice question (which name the RC referred to). Two different words corresponding to NP1 and NP2 were displayed on the screen and participants had to respond by mean of the keyboard. In order to check whether participants correctly attended to all stimuli, an additional 2-alternative forced

choice comprehension question was asked at the very end of 20% of the trials (the general structure of a trial is described in the figure 2).

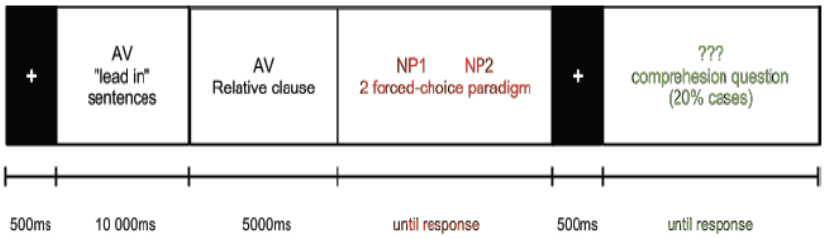

Fig.2: General structure of the experimental procedure (A and AV versions).

2.4 ERPs recording and analyses

While participants run the experiment, we recorded their EEG signal to perform ERPs analysis. ERPs were time-locked to the onsets of NP1, NP2 and Relative clause.

3. Analyses and results

3.1. Behavioral results

3.1.1. Proportions of High and Low Attachment preference in the A version.

Responses were then classified into two categories High attachment (with NP1, or High Attach) *vs* Low attachment (with NP2, or Low Attach) interpretations**.** For each of the three conditions, we calculated the proportion of How Attachment (in percentage) across

sentences and participants. We applied repeated measures ANOVAs with prosody as a three-level factor (NP1, NP2 and NP). Mauchly's test was not significant, therefore sphericity can be assumed. The analyses of variance evidenced a significant effect of condition for attachment preference ($F(2, 51)= 44,12$; $p< 0.001$).

Post-hoc paired t-test using Bonferroni correction revealed that there is an effect of the locus of the pause (between conditions NP1-NP2: $t= 59.505$, $p< 0.001$). The effect of prosodic cue placement also proved to be significant (NP-NP1: $t= 26.266$, $p< 0.001$) NP-NP2: $t= -33.239$, $p< 0.001$). These results are illustrated in the fig.3.

3.1.2. Proportions of High and Low Attachment preference in the AV version.

We performed the exact same analyses with the AV version data. Results showed no difference of Low attachment preference between the three conditions $F(2, 40)= 0, 96$; $p= 0,39$). These results are illustrated in the fig.3, and suggest that the placement of the beat gesture on NP1, NP2 or its absence (NP) did not modulated participants interpretation of the ambiguous sentences.

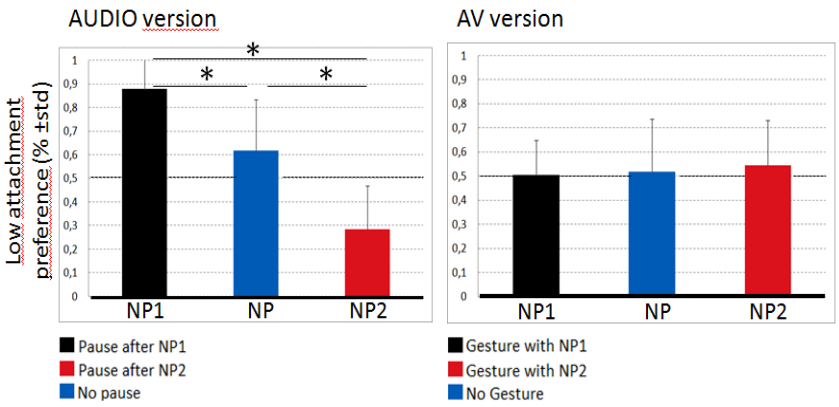

Fig.3: Mean low attachment preferences rates (% +/- std) per condition in the A only (left graph) and AV (right graph) modalities : NP1 (prosodic cue associated to NP1, black column), NP2 (prosodic cue associated to NP2, red column) and NP (no prosodic cue, blue column).

3.2. ERP results

3.2.1. Audio only condition

The ERPs time-locked to the NP1 onset revealed no differences of signal across conditions. Similarly, the ERPs time-locked to the NP2 onset revealed no significant time window of interest between the three conditions. In contrast, the ERPs time-locked to the relative clause onset revealed a time window of interest at 0-150ms, for which we found an effect of condition ($F(2,19)=4,77$; $p=0,009$). Posthoc analyses showed that the signal in the condition NP2 was significantly more negative than NP1 and NP conditions, which were not different with each others (NP2 vs NP1: pvalue=0,027; NP2 vs NP: pvalue=0,020; NP1 vs NP: pvalue>0,5). This last result may suggest that, as the acoustic prosody associated with NP1

reinforced the preference for low attachment naturally preferred if guessing (in NP condition), the relative clause was processed the same in NP1 and NP conditions. In contrast, as the prosody associated with NP2 favored the high attachment preference, this may explain the difference of processing as compared to NP1/NP conditions.

3.2.2. AV condition

The ERPs time-locked to the NP1 onset revealed a relevant temporal window of interest corresponding to the N100/P200 component. Within the 60-120 ms (N100 component), there was a significant effect of condition ($F(2,18)=6,45$; $p=0,004$). Posthoc analyses showed that the signal in the NP1 condition (black line) was more positive than both the NP2 and PN conditions (respectively red and blue lines), which were not different with each other (NP1 vs NP2: pvalue<0,005; NP1 vs NP: pvalue<0,005; NP2 vs NP: pvalue>0,5). In the 170-240 ms time window (P200 component), there was a significant effect of condition ($F(2,18)=7,40$; $p=0,002$). Posthoc analyses showed that the signal in the NP1 condition containing the beat gesture (black line) was more positive than both the NP2 and PN conditions (respectively red and blue lines), which were not different with each other (NP1 vs NP2: pvalue<0,009; NP1 vs NP: pvalue<0,009; NP2 vs NP: pvalue>0,5).
Similarly, the ERPs time-locked to the NP2 onset revealed a relevant temporal window of interest corresponding to the

N100/P200 component. Within the 60-120 ms (N100 component), there was a significant effect of condition (F(2,18)=9,46; p<0,001). Posthoc analyses showed that the signal in the NP2 condition (red line) was more positive than both the NP1 and PN conditions (respectively black and blue lines), which were not different with each other (NP2 vs NP1: pvalue<0,005; NP2 vs NP: pvalue<0,05; NP1 vs NP: pvalue=0,85). In the 170-240 ms time window (P200 component), there was a significant effect of condition (F(2,18)=6,7; p=0,004). Posthoc analyses showed that the signal in the NP1 condition containing the beat gesture (black line) was more positive than both the NP2 and PN conditions (respectively red and blue lines), which were not different with each other (NP2 vs NP1: pvalue=0,03; NP2 vs NP: pvalue<0,005; NP1 vs NP: pvalue>0,5). Finally, results showed no differences of signal across conditions for the ERPs time-locked to the onset of the relative clause F(2,18)=0,71; p=0,5.

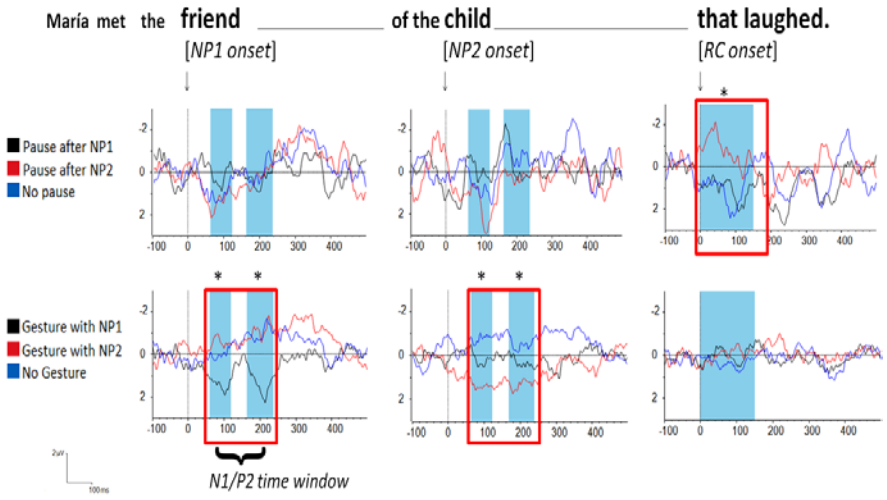

Fig.4: ERPs time-locked to the first noun (NP1), the second noun (NP2) and the relative clause onsets per condition in the A only (top) and AV (bottom) modalities: NP1 (prosodic cue associated to NP1, black column), NP2 (prosodic cue associated to NP2, red column) and NP (no prosodic cue, blue column).

4. Discussion

This study aimed to assess the prosodic role of gestures in a context of ambiguity. Based on the observation that beats share some features with prosody, we suggested that they might share functional characteristics as well. In order to demonstrate the analogy, we needed to provide evidence that prosody alone plays a role in syntactic comprehension and compare its potential effects on sentence interpretation, with beats. More specifically, we addressed the question of whether intonational boundaries, such as pauses, could modulate the perceiver's interpretation on ambiguous relative clauses (A only version of the experiment). Our results offered a congruent and complementary view on the topic. In contrast with the acceptability judgment task previously used, the data we gathered is a direct comprehension task, which provided direct access to the listener's interpretation. The behavioral results of the A only version of the experiment confirmed the role of prosody in syntactic parsing, showing that almost perfectly balanced ambiguity can be resolved by the use of prosodic cues. The ERPs results of the A version were less clearer as the modulation found at the onset of the relative clause between NP2 vs NP1/NP may be explained by the simple pause before the

RC onset in the NP2 condition, as compared to NP1/NP condition. Further investigations are needed to maybe set a more adapted contrast for the ERPs analysis. In contrast, the parallel study assessing beat gestures has proved to be more challenging (AV version of the experiment). Gestures seemed to be subject to more inter-individual variability than prosody. One explanation may reside in individual characteristics: we do rely differently on visual information when perceiving speech. It has also been shown in other audiovisual studies with the McGurk effect which does not work for everyone. Another explanation is methodological: the videos were created respecting to opposing constraints. On the one hand, we had to control our stimuli to unable comparisons between sentences; on the other, ecological validity should be maintained. Respecting the former meant weakening the latter, and vice versa. Alternatively, we may reach a situation where gestures did not seem trivial, but where the videos were not quite natural either. That may affect the participants' judgment. In any case, AV results suggested that in our experimental conditions, beats did not help disambiguate sentences, whereas acoustic prosody did. ERP results were also unclear as in the AV experiment. Gesture synchronized either with the first noun (NP1 condition) or the second one (NP2 condition) affected the signal of auditory integration at N100/P200 component time windows, respect to the no gesture condition (NP condition) or if the gesture occurs before the word (NP1 respect to word NP2) or late after (NP2 respect to word NP1). At relative closure (RC), the presence of a previous gesture (NP1 or NP2) did not affect the

signal, ssuggesting that the gesture did not modulate the interpretation of the last part of the experimental sentence, nor helped respect to no gesture condition (NP). This is in line with behaviour data as we did not modulate attachment preference in the gesture version of the experiment. Both behavioral and ERPs data suggested that gestures were perceived as simple movements in this particular experimental conditions, maybe affecting local attentional processes of AV integration (as reflected at the early N100/P200 time window modulations) but no later syntactic processes of speech (as they did no modulate the participants preference for the attachment neither). This is in line with our ERPs and oscillations studies that suggested a local attentional effect of beat gestures on the following associated word (section 4.1 and 4.2 in the present thesis). In this case, it is somehow not surprising to find an early effect of the co-occurring beat on the auditory signal of the associated word, without modulating later higher syntactic analysis processing. However, the present results are not conclusive and further investigations are needed to find a finer behavioral measure or adapt the experimental procedure to isolate the impact of beats on syntactic parsing.

ANNEX 4.

Synchronization by the hand: The sight of gestures modulates low-frequency activity in brain responses to continuous speech

Emmanuel Biau [a*] and Salvador Soto-Faraco [a, b]

[a] Multisensory Research Group, Center for Brain and Cognition, Universitat Pompeu Fabra, Barcelona, Spain.

[b] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain.

Correspondence:

Emmanuel Biau
Center for Brain and Cognition
Universitat Pompeu Fabra
Roc Boronat, 138
08018 Barcelona, Spain
emmanuel.biau@free.fr

Total number of words: 2795

Total number of figures: 2

Abstract

During social interactions, speakers often produce spontaneous gestures to accompany their speech. These coordinated body movements convey communicative intentions, and modulate how listeners perceive the message in a subtle, but important way. In the present perspective, we put the focus on the role that congruent non-verbal information from beat gestures may play in the neural responses to speech. Whilst delta-theta oscillatory brain responses reflect the time-frequency structure of the speech signal, we argue that beat gestures promote phase resetting at relevant word onsets. This mechanism may facilitate the anticipation of associated acoustic cues relevant for prosodic/syllabic-based segmentation in speech perception. We report recently published data supporting this hypothesis, and discuss the potential of beats (and gestures in general) for further studies investigating continuous AV speech processing through low-frequency oscillations.

Keywords: audiovisual speech, gestures, beats, low-frequency oscillations, EEG

Speakers spontaneously gesture to accompany their speech, and listeners definitely seem to take advantage of this source of complementary information from the visual modality (Goldin-Meadow, 1999). The aim of the present perspective is to bring attention to the relevance of this visual concomitant information when investigating continuous speech. Here we argue that part of this explanation may have to do with the modulations that speaker's gestures impose on low-frequency oscillatory activity related to speech segmentation in the listener's brain. The speaker modulates the amplitude envelope of the utterance (i.e. the summed acoustic power across all frequency ranges for each time point of the signal) in a regular manner, providing quasi-rhythmic acoustic cues in at least two low-frequency ranges. First, speech syllables are produced rhythmically at frequency of 4-7Hz, corresponding to a theta rate imposed by voicing after breath taking and jaw aperture (Peelle & Davis, 2012). Second, the speaker modulates pitch accents in her/his vocalization to convey particular speech acts (e.g. declarative or ironic), and emphasize relevant information to convey communicative intentions. These pitch peaks also occur with a quasi-rhythmic rate of 1-3Hz corresponding to a delta frequency and constituting part of prosody (Park et al., 2015; Munhall et al., 2004). Recently, Electroencephalography (EEG) and Magnetoencephalography (MEG) studies investigated auditory speech segmentation mechanisms, taking advantage of time-frequency analyses to look at brain activities that are not time-locked to stimuli onsets, and measure the amount of activity in frequency bands of interest (typically missing in the classic Event-Related Potential (ERPs) averages). These studies reported that spontaneous delta-theta activities in the auditory cortex reset their phase to organize in structured patterns, highly similar to the spectro-temporal architecture of the auditory speech envelope, reflecting entrainment mechanism (Gross et al., 2015; Park et al., 2015; Zoefel & VanRullen, 2015; Giraud & Poeppel, 2012; Nourski et al., 2009; Abrams et al., 2008; Luo & Poeppel, 2007; Ahissar et al., 2001). Then, delta-theta periodicity seems to constitute a fundamental window of compatibility between brain's activity and speech segmentation (Ghitza & Greenberg, 2009; Peelle & Davis, 2012). Thus, when the natural delta-theta periodicity in the auditory signal is affected by time compression, speech comprehension worsens significantly. But more interestingly, the degradation of the delta-theta rhythms also decreases the spectro-temporal similarity between the speech envelope and the low-frequency activities in the auditory cortex (Ahissar et al., 2001). These important

spectro-temporal features of the acoustic signal seem to be, therefore, important in determining brain responses to speech.

Yet, the acoustic signal is not the only communicative cue between speaker and listener. Coherent face and body movements often accompany verbalization. Before placing the focus on the speaker's hand gestures, it is important to note that the relevance of non-verbal information has been first established regarding the speaker's face (van Wassenhove, Grants & Poeppel, 2005). Corresponding lip movements have been long shown to facilitate comprehension in noisy conditions (Sumby & Pollack, 1954), or in contrast, affect speech processing when incongruent with utterance, e.g. the famous McGurk illusion (McGurk & McDonald, 1976). More recently, visual speech information has been proposed to play a role in the extraction of the aforementioned rhythmic aspects of the speech signal (van Wassenhove, Grants & Poeppel, 2005). Due to the natural precedence of visual speech cues over their auditory counterparts in natural situations (i.e. the sight of articulation often precedes its auditory consequence; see Sánchez-García et al., 2011), it has been hypothesized that visual information conveys predictive information about the timing and contents of corresponding auditory information, facilitating its anticipation (Vroomen & Stekelenburg, 2010; Stekelenburg & Vroomen, 2007; van Wassenhove, Grants & Poeppel, 2005). For example, van Wassenhove, Grant and Poeppel (2005) presented isolated consonant-vowel syllables in audio, visual or audiovisual modalities. They showed that the N1-P2 component in the auditory evoked responses time-locked to the phoneme onset were significantly reduced in amplitude and speeded up in time in the AV modality, compared to the responses to auditory syllables. In the time-frequency dimension, delta-theta entrainment has been proposed to underlie predictive coding mechanism based on the temporal correlation between audio-visual speech cues (Arnal & Giraud, 2012; Lakatos, Karmos, Mehta, Ulbert, & Schroeder, 2008; Schroeder & Lakatos, 2009; Schroeder, Lakatos, Kajikawa, Partan, & Puce, 2008). Thus, Arnal and Giraud (2012) hypothesized that visual information provided by lip movements increases delta-theta phase resetting at relevant associated acoustic cue onsets (word onsets), reflecting predictive coding mechanisms that minimize the uncertainty about when regular event are likely to occur, and a better speech segmentation.

Along these lines, one could ask whether other speech-related visible body movements of the speaker may also bear predictive information and have an impact on low-frequency neural activity in the listeners' brain. In continuous speech production, which movements may be correlated with delta-theta acoustic cues in the auditory signal? Head movements for example, were shown to be highly correlated with pitch peaks and facilitate comprehension of speech perception in noisy conditions (Munhall et al., 2004). Looking at public addressees, and in particular political discourses, we observed that speakers almost all the time accompany their speech with spontaneous hand gestures called "beats" (McNeill, 1992). Beats are simple and biphasic arm/hand movements that often bear no semantic content in their shape produced by speakers when they want to emphasize relevant information or develop an argument with successive related points. They belong to what could be considered as visual prosody, as they are temporally aligned with the prosodic structure of the verbal utterance, just like eyebrow, shoulders and head nods (Leonard & Cummins, 2012; Krahmer & Swers, 2007; McNeill, 1992). Yasinnik, Renwick and Shattuck-Hufnagel (2004) showed that beats' apexes (i.e. the maximum extension point of the arm before retraction, corresponding to the functional phase of the gesture) align quite precisely with pitch-accented syllables (peaks of the F0 fundamental frequency). In other words, the kinematics of beats match with spectro-temporal modulation of auditory speech envelope and are thought to modulate both the acoustic properties and the perceived saliency of the affiliated utterance (Krahmer & Swers, 2007; Munhall et al., 2004). Albeit simple, beats have been found to modulate syntactic parsing (Henning et al., 2012; Guellaï, Langus & Nespor, 2014), semantic processing (Wang & Chu, 2013) and encoding (So, Chen-Hui & Wei-Shan, 2012) during audiovisual speech perception. In a previous ERP study, we showed that the sight of beats modulate the ERPs produced by the corresponding spoken words at early phonological stages by reducing negativity of the waveform within the 200-300 ms time window (Biau & Soto-Faraco, 2013). Since the onsets of the beats systematically preceded affiliated words onsets by around 200 ms, we concluded that the order of perception and congruence between pitch accents and apexes attracted the focus of local attention on relevant acoustic cues in the signal (i.e. words onsets), possibly modulating speech processing from early stages.

Based on these previous studies and the stable spatio-temporal relationship between beats and auditory prosody, we argued that continuous speech segmentation

should not be limited to the auditory modality, but also take into account visual congruent information both from lip movements and the rest of the body. Recently, Skipper (2014) proposed that listeners use the visual context provided by gestures as predictive information because of learned preceding timing with associated auditory information. Gestures may pre-activate words associated with their kinematics, to process inferences that are compared with following auditory information. In the present context, the idea behind was that if gestures provide robust prosodic information that listeners can use to anticipate associated speech segments, then beats may have an impact on the entrainment mechanisms capitalizing on rhythmic aspects of speech, discussed above (Arnal & Giraud, 2012; Giraud & Poeppel, 2012; Peelle & Davis, 2012). More precisely, we expected that if gestures provide a useful anticipatory signal for particular words in the sentence, this might reflect in phase synchronization of low frequency at relevant moments in the signal, coinciding with the acoustic onsets of the associated words (see figure 1). This is exactly what we have tested in a recent EEG study, by presenting a naturally spoken, continuous AV speech in which the speaker spontaneously produced beats while addressing the audience (Biau et al., 2015). We recorded the EEG signal of participants during AV speech perception, and compared the phase-locking value (PLV) of low-frequency activity at the onset of words pronounced with or without a beat gesture (see figure 1). The PLV analysis revealed strong phase synchronization in the theta 5-6 Hz range with a concomitant desynchronization in the alpha 8-10 Hz range, mainly at left fronto-temporal sites (see figure 2). The gesture-induced synchronization in theta started to increase around 100 ms before the onset of the corresponding affiliate word, and was maintained for around 60 ms thereafter. Given that gestures initiated approximately 200 ± 100 ms before word onsets, we thought that this delay was enough for beat to effectively engage the oscillation-based temporal prediction of speech in preparation for the upcoming word onset (Arnal & Giraud, 2012). Crucially, when visual information was removed (that is, speech was presented in audio modality only), our results showed no difference in PLV or amplitude between words that had been pronounced with or without a beat gesture in the original discourse. Such pattern suggested that the effects observed in the AV modality could be attributed to the sight of gestures, and not just acoustic differences between gesture and no gesture words in the continuous speech. We interpreted these results within the following framework: Beats are probably perceived as communicative rather than simple body movements disconnected from the message (Hubbard et al., 2009; McNeill, 1992).

Through daily social experience, listeners learn to attribute linguistic relevance to beats because they gesture when they speak (So et al., 2012; McNeill, 1992), and seem to have an understanding of the sense of a beat at a precise moment. Consequently, listeners may rely on beats to anticipate associated speech segmentation that is reflected through an increase of low-frequency phase resetting at relevant onsets of accompanied words. In addition, it is possible that this prediction engages local attentional mechanisms, reflected by early ERP effects and the alpha activity reduction seen around word onsets with gesture. As far as we know, Biau et al. 2015 was the first study investigating the impact of spontaneous hand gestures on speech processing through low-frequency oscillatory activities in a close-to-natural approach. Further investigations are definitely needed to increase data and set new experimental procedures combining behavioural measures with EEG analyses.

Figures 1 & 2

A recent study by He and others (2015) has investigated AV speech processing through low-frequency activity, albeit with a very different category of speech gestures. He et al. used intrinsically-meaningful gestures (IMG) conveying semantic content, such as when the speaker makes a "thumbs-up" gesture while uttering "the actor did a good job". The authors investigated the oscillatory signature of gesture-speech integration by manipulating the relationship between gesture and auditory speech modalities: AV integration (IMG produced in the context of an understandable sentence in the listener's native language), V (IMG produced in the context of a sentence in a foreign language incomprehensible for the listener) and A (an understandable sentence in the listener's native language without gestures). The results of a conjunction analysis showed that the AV condition induced a significant centrally-distributed power decrease in the alpha band (7-13Hz; from 700 to 1400 ms after the onset of the critical word associated with the gesture in the sentence), as compared to the V and A conditions that contained only semantic inputs from one modality (respectively: in the V condition only the gesture was understandable and in the A condition only the utterance was understandable). The authors concluded that the alpha power decrease reflected an oscillatory correlate of the meaningful gesture–speech integration process.

Investigations on the neural dynamics of hand gesture-speech integration during continuous AV speech perception have just begun but the results reported in both studies (He et al., 2015; Biau et al., 2015) already suggest two important conclusions for the present perspective. First, whereas auditory speech seems at first glance to attract all the listeners' attention, hand gestures count as well, and may definitely be considered as visual linguistic information for online AV speech segmentation. If the delta-theta rhythmic aspects in the auditory signal can play the role of anchors for predictive coding during speech segmentation (Park et al., 2015, Arnal & Giraud, 2012; Peelle & Davis, 2012), then preceding visual gestural information, naturally present in face to face conversations, may convey very useful information for decoding the signal and thus, be taken into account. For instance, beats are not only exquisitely tuned to the prosodic aspects of the auditory spectro-temporal structure, but also engage language-related brain areas during continuous AV speech perception (Hubbard et al., 2009). This idea is in line with earlier arguments considering auditory speech and gestures as two sides of the same common language system (Kelly, Creigh & Bartolotti, 2009; McNeill, 1992 for some examples). Gestures may constitute a good candidate to investigate the multisensory integration between natural auditory speech and social postures. For example, Mitchel and Weiss (2014) showed that the simple temporal alignment between V and A information did not fully explain the AV benefit (i.e. multisensory integration) in a segmentation task of artificial speech. Indeed, segmentation was significantly better when visual information came from a speaker that was previously exposed to the words he had to pronounce during the stimuli recording (then, knowing the prosodic contours of words, i.e. boundaries), compared from a speaker that was unaware of word boundaries when recording. These results suggested that facial movements conveyed helpful visual prosodic contours if the speakers was aware of them. The same conclusion may apply to beat gestures as they synchronize with auditory prosody in communicative intent (and the speaker knows the prosodic contours of her/his own discourse). For example, it may be interesting to compare delta-theta activity patterns between gestures conveying the proper communicative prosody and simple synchronized hand movements without the adequate prosodic kinematics.

A second interim conclusion from the few current studies addressing the oscillatory correlates of gestures is that low-frequency brain activity appears to be a successful neural marker to investigate gesture-speech integration and, continuous AV speech processing in general. Based on the results reported in these two pioneer studies, low

frequency activity seemed sensitive to the type of gesture (intrinsically meaningful gestures in He et al., 2015 and beats in Biau et al., 2015). Both studies analysed a contrast, comparing the low-frequency activity modulations between an AV gesture condition (i.e. words were accompanied with a gesture) and an AV no gesture condition (i.e. words were pronounced without gesture, but the speaker was visible). He and colleagues reported a decrease of alpha power (from 400ms to 1400ms) and a beta power decrease (from 200 to 1200ms) after the critical word onset, whilst Biau et al. reported a theta synchronization with a concomitant alpha desynchronization temporally centred on the affiliate word onset (note that the alpha activity modulation was found in both studies). Even if the experimental procedures and stimuli were not the same (in He et al. the speaker was still in the no gesture condition, whereas moving in Biau & Soto-Faraco), the distinct patterns of low-frequency modulations in the gesture-no gesture contrasts suggested that different kind of gestures may be associated to different aspects of the verbalization, modulating speech processing diversely. Indeed, IMGs describe a conventionally established meaning and can be understood silently whereas beats do not and need to be contextualized by speech to become functional. This might explain why the timing of modulations in He et al. was quite different respect to Biau et al. Then, oscillations may constitute an excellent tool for further investigations on neural correlate of AV speech perception and associated social cues with different communicative purposes (IMG vs. beats).

Speech is an intrinsically multisensory object of perception, as the act of speaking produces correlates to the ear and to the eye of the listener. The aim of the present short perspective was to bring attention to the fact that conversations engage a whole set of coordinated body movements. Furthermore, we argue that considering the oscillatory brain responses to natural speech may capture an important aspect of how the listeners' perceptual system integrates back all the different aspects of the communicative production from the talker. Future studies may investigate more precisely how this integration occurs, and what is the role of synchronization and desynchronization patterns that we have tentatively interpreted here.

FUNDING

This research was supported by the Spanish Ministry of Science and Innovation (PSI2013-42626-P), AGAUR Generalitat de Catalunya (2014SGR856) and, the European Research Council (StG-2010 263145).

AKNOWLEDGMENTS

We would like to thank Mireia Torralba, Ruth de Diego Balaguer and Lluis Fuentemilla who took part in the project reported in the present perspective (Biau et al., 2015).

REFERENCES

Abrams, D. A., Nicol, T., Zecker, S., & Kraus, N. (2008). Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *28*(15), 3958–65.

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(23), 13367–72.

Arnal, L. H., & Giraud, A. L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences, 16*(7), 390-398.

Biau, E., & Soto-Faraco, S. (2013). Beat gestures modulate auditory integration in speech perception. *Brain and Language, 124*(2), 143-152.

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.

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, *11*(4), 796–804.

Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica, 66*(1-2), 113-126.

Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience, 15*(4), 511-517.

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Sciences*, *3*(11), 419–429.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biology*, *11*(12), e1001752.

Guellaï, B., Langus, A., & Nespor, M. (2014). Prosody in the hands of the speaker. *Frontiers in Psychology*, *5*, 700.

He, Y., Gebhardt, H., Steines, M., Sammer, G., Kircher, T., Nagels, A., & Straube, B. (2015). The EEG and fMRI signatures of neural integration: An investigation of meaningful gestures and corresponding speech. *Neuropsychologia*, *72*, 27–42.

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.

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*(3), 1028-1037.

Kelly, S. D., Creigh, P., & Bartolotti, J. (2010). Integrating speech and iconic gestures in a stroop-like task: Evidence for automatic processing. *Journal of Cognitive Neuroscience, 22*(4), 683-694.

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*(3), 396-414.

Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science (New York, N.Y.), 320*(5872), 110-113.

Leonard, T., Cummins, F. The temporal relation between beat gestures and speech. (2011). *Language and Cognitive Processes*, *26*, 10.

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron, 54*(6), 1001-1010.

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

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

Mitchel, A. D., & Weiss, D. J. (2014). Visual speech segmentation: using facial cues to locate word boundaries in continuous speech. *Language Cognitive Processes*, *29*(7), 771–780.

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*(2), 133-137.

Nourski, K. V, Reale, R. A., Oya, H., Kawasaki, H., Kovach, C. K., Chen, H., … Brugge, J. F. (2009). Temporal envelope of time-compressed speech represented in the human auditory cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *29*(49), 15564–74.

Park, H., Ince, R. A. A., Schyns, P. G., Thut, G., & Gross, J. (2015). Frontal Top-Down Signals Increase Coupling of Auditory Low-Frequency Oscillations to Continuous Speech in Human Listeners. *Current Biology*, *25*(12), 1649–53.

Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology, 3*, 320.

Sánchez-García, C., Alsius, A., Enns, J. T., & Soto-Faraco, S. (2011). Cross-modal prediction in speech perception. *PloS one*, *6*(10), e25198.

Schroeder, C. E., & Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in Neurosciences, 32*(1), 9-18.

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., & Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Sciences, 12*(3), 106-113.

Skipper, J. I. (2014). Echoes of the spoken past: how auditory cortex hears context during speech perception. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *369*(1651), 20130297.

So, W. C., Chen-Hui, C. S., Wei-Shan, J. L. (2012). Mnemonic effect of iconic gesture and beatgesture in adults and children: Is meaning in gesture important for memory recall? *Language and Cognitive Processes*, 27(5), 665-681.

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience, 19*(12), 1964-1973.

Sumby, W., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*(2), 212-215.

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*(4), 1181-1186.

Vroomen, J., & Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *Journal of Cognitive Neuroscience*, *22*(7), 1583–96.

Wang, L., & Chu, M. (2013). The role of beat gesture and pitch accent in semantic processing: An ERP study. *Neuropsychologia, 51*(13), 2847-2855.

Yasinnik, Y. (2004). The timing of speech-accompanying gestures with respect to prosody. *Proceedings of From Sound to Sense,MIT*. MIT.

Zoefel, B., & VanRullen, R. (2015). Selective perceptual phase entrainment to speech rhythm in the absence of spectral energy fluctuations. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *35*(5), 1954–64.

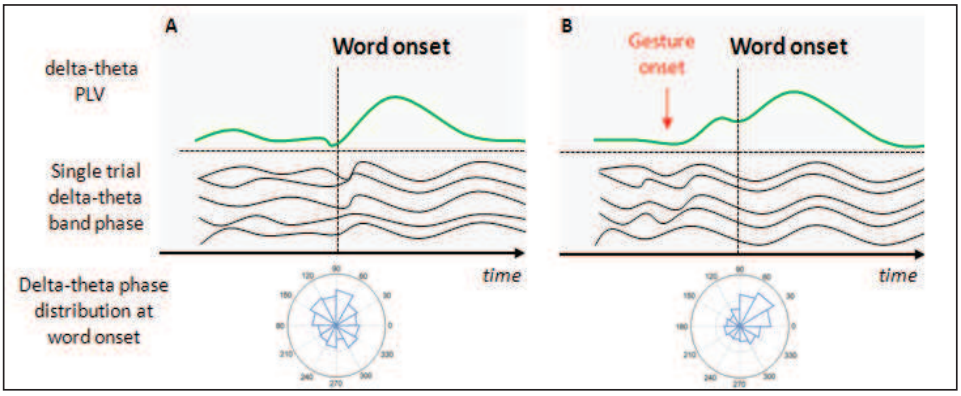

Figure 1. Illustration of the potential effect of beat gestures on the delta-theta phase resetting. **(A)** At the beginning of speech, neural populations in the auditory cortex spontaneously discharge at delta-theta rates but not at the same phase for a given time point (this is illustrated by the single trial delta-theta band phase before the word onset). At the first word onset, a phase distribution in the auditory sensors shows no preferred angle in the delta-theta band. In consequence, the delta-theta phase locking value (PLV) at the first word onset is weak. With progressive entrainment, delta-theta phase synchronizes, increasing PLV with a preferred angle at relevant syllable/word onsets. **(B)** Beat onsets systematically precede word onsets and potentially increase the delta-theta entrainment before the arriving word onset. When the relevant gesture onset occurs, delta-theta activity synchronizes with a preferred angle in the phase, increasing PLV before the associated word onset arrives to anticipate its processing.

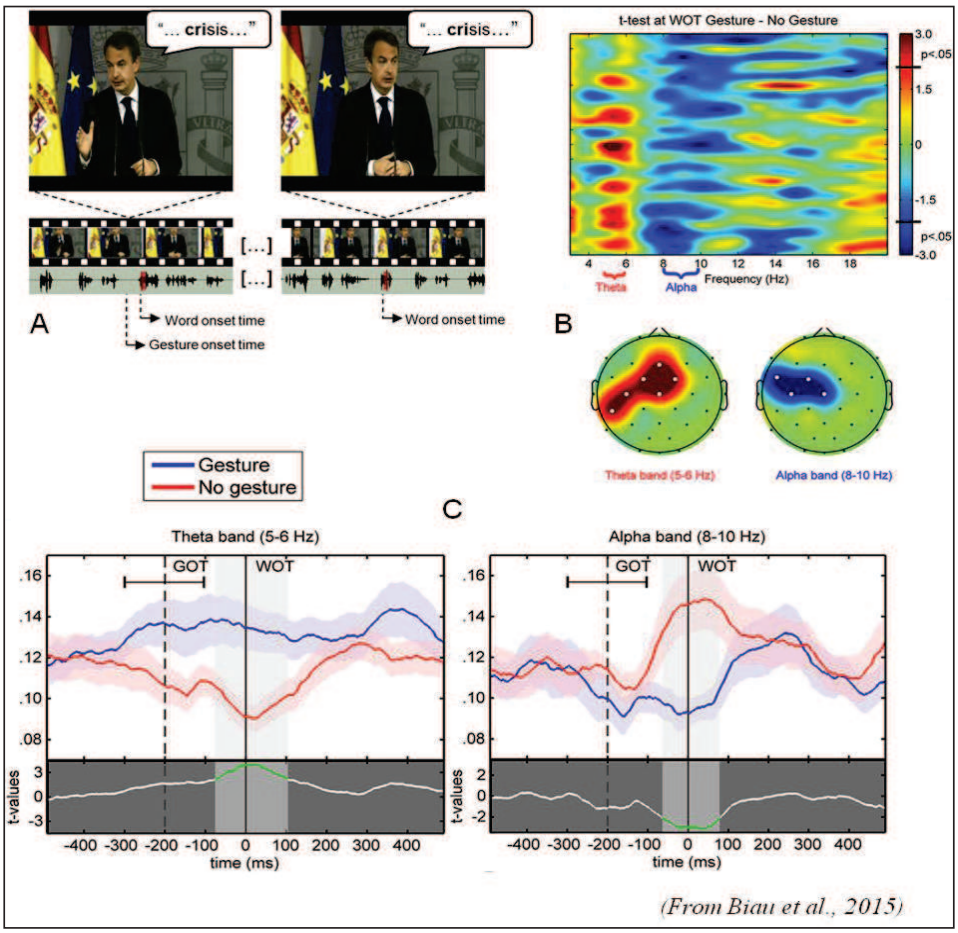

Figure 2. (A) Example of video-frames for the gesture (left) and no gesture (right) conditions associated to the same stimulus word "crisis". The speaker is the former Spanish President Luis Rodríguez Zapatero, recorded at the palace of La Moncloa, and the video is freely available on the official website (Balance de la acción de Gobierno en 2010, 12-30-2010; http://www.lamoncloa.gob.es). Below, the oscillogram of corresponding audio track fragments (section corresponding to the target word shaded in red). The onsets of both the gesture and corresponding word (gesture condition) are marked. **(B)** (top B) Representation of paired t-test values for the comparison between PLV at word onset in the gesture and no gesture conditions with frequency bands of interest labeled in the x axis. (bottom B) Topographic representation of the significant clusters (significant electrodes marked with white dots) for the t-tests within the theta and alpha bands. **(C)** PLV time course in 5-6Hz theta (left) and 8-10Hz alpha (right) frequency bands at Cz electrode for the gesture (blue line) and no gesture (red line)

conditions. The mean average ± standard deviation of gesture onset time (GOT) is represented respect to word onset time (WOT). The lower part of each plot displays the paired t-test values between gesture and no gesture conditions. The shaded bands indicate significant time intervals (highlighted in green in the t-test line).