Contents lists available at SciVerse ScienceDirect

International Journal of Psychophysiology

journal homepage: www.elsevier.com/locate/ijpsycho

Visual information constrains early and late stages of spoken-word recognition in sentence context

Angèle Brunellière [a],*, Carolina Sánchez-García [b], Nara Ikumi [b], Salvador Soto-Faraco [b,c]

[a] Unité de Recherche en Sciences Cognitives et Affectives, University of Lille 3, France
[b] Departament de Tecnologies de la Informació i les Comunicacions, Universitat Pompeu Fabra, Barcelona, Spain
[c] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

ARTICLE INFO

Article history:
Received 20 January 2013
Received in revised form 30 May 2013
Accepted 14 June 2013
Available online 22 June 2013

Keywords:
Visual speech
Semantic constraints
Spoken-word recognition
Event-related potentials

ABSTRACT

Audiovisual speech perception has been frequently studied considering phoneme, syllable and word processing levels. Here, we examined the constraints that visual speech information might exert during the recognition of words embedded in a natural sentence context. We recorded event-related potentials (ERPs) to words that could be either strongly or weakly predictable on the basis of the prior semantic sentential context and, whose initial phoneme varied in the degree of visual saliency from lip movements. When the sentences were presented audio-visually (Experiment 1), words weakly predicted from semantic context elicited a larger long-lasting N400, compared to strongly predictable words. This semantic effect interacted with the degree of visual saliency over a late part of the N400. When comparing audio-visual versus auditory alone presentation (Experiment 2), the typical amplitude-reduction effect over the auditory-evoked N100 response was observed in the audiovisual modality. Interestingly, a specific benefit of high- versus low-visual saliency constraints occurred over the early N100 response and at the late N400 time window, confirming the result of Experiment 1. Taken together, our results indicate that the saliency of visual speech can exert an influence over both auditory processing and word recognition at relatively late stages, and thus suggest strong interactivity between audio-visual integration and other (arguably higher) stages of information processing during natural speech comprehension.

1. Introduction

In natural face-to-face communication, visual information from the speaker such as lip movements and hand gestures effectively contributes to speech processing (McNeill, 1992; Biau and Soto-Faraco, 2013; Sumby and Pollack, 1954; McGurk and Macdonald, 1976). Indeed, it has been well established that visual articulatory information is combined with auditory information during speech perception. For example, in the McGurk effect (McGurk and Macdonald, 1976), the perceptual fusion between incongruent auditory (i.e. /ba/) and visual (i.e. [ga]) information often produces the illusory perception of a new, intermediate sound (i.e. /da/). In normal, everyday life conditions, where auditory signals are strongly correlated with visual articulations, speech perception benefits from integrating cues across sensory modalities, especially when the processing of auditory information is difficult (such as in noisy contexts, Ma et al., 2009; Ross et al., 2007; Sumby and Pollack, 1954, or while perceiving non-native languages, Navarra and Soto-Faraco, 2007). In addition to behavioral evidence supporting a visual influence on auditory speech perception, electrophysiological studies have also suggested that viewing the speakers' lip movements elicits faster and more efficient processing of spoken cues (e.g., Besle et al., 2004; Klucharev et al., 2003; van Wassenhove et al., 2005). For instance, several authors (Besle et al., 2004; Klucharev et al., 2003; van Wassenhove et al., 2005) have reported a facilitation over early auditory event-related potentials (ERP) when the visual articulatory information was in accordance with the auditory information. In particular, it has been found that audio-visually congruent syllables elicited a reduced amplitude (Besle et al., 2004; Klucharev et al., 2003; van Wassenhove et al., 2005) and an earlier peak of the auditory N100 component (van Wassenhove et al., 2005), compared to auditory-alone stimulation.

Although most electrophysiological studies presenting isolated syllables or vowel segments show early effects at pre-lexical level, little is known about the impact of visual articulatory information on word recognition in more complex spoken contexts. Indeed, the few ERP studies investigating the influence of visual articulatory cues during spoken word perception (Mengin et al., 2012; Shahin et al., 2012) have mainly examined early electrophysiological components (i.e. the N100/P200), mostly overlooking later components associated with the process of word recognition and, without manipulating linguistic variables involved in spoken word recognition. Moreover, to our knowledge, no ERP studies have yet explored the influence of visual articulatory cues in sentence context. This is surprising, not only because speech is most often experienced in sentential context, but also because visual speech

* Corresponding author at: Unité de Recherche en Sciences Cognitives et Affectives, Université Charles-de-Gaulle Lille3, Domaine universitaire du Pont de Bois, BP 149, 59653 Villeneuve d'Ascq Cedex, France. Tel.: +33 3 20 41 72 04.
E-mail address: angele.brunelliere@univ-lille3.fr (A. Brunellière).

0167-8760/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ijpsycho.2013.06.016

information, like sentence meaning, has been suggested to constrain the processing of incoming speech input in a predictive coding framework (Pickering and Garrod, 2007; Stekelenburg and Vroomen, 2007; van Wassenhove et al., 2005). Thus, one might argue that these two different constraining sources of the speech input (sentence context and visual information) exert an interactive influence on word recognition. In order to start investigating such interaction, the present ERP study sets out to explore the contribution of visual articulatory information during spoken-word recognition in the context of sentences varying in semantic constraints.

Research on auditory word recognition in sentence context has described three main ERP components, the N100, N200 and N400. While the early auditory-evoked N100 component is triggered by the onset of auditory events including speech sounds and reflects auditory sensory cortex activity, the two other negative-polarity components called N200 and N400 are thought to reflect various specific stages of spoken word processing (e.g. Connolly et al., 1990; Kutas and Hillyard, 1984). By far, the most studied electrophysiological component during word recognition in sentential context is the N400 wave (Kutas and Hillyard, 1984). This negativity peaks maximally around 400 ms after word onset and usually has a centro-parietal scalp distribution. The fluctuation of the N400 wave reflects the ease with which a word is processed at a lexico-semantic level. In written or spoken sentential contexts, the amplitude of the N400 is larger in response to words that do not fit well with the preceding context, compared to words which are highly expected (Connolly and Phillips, 1994; Connolly et al., 1992; Connolly et al., 1990; Kutas and Hillyard, 1984; van Berkum et al., 2005). In addition to the semantically-related N400 wave, an earlier electrophysiological component, peaking around 200 ms after word onset (N200), has been reported in sentential contexts when words are presented in the auditory modality (Connolly and Phillips, 1994; Connolly et al., 1992; Connolly et al., 1990; van den Brink et al., 2001; van den Brink and Hagoort, 2004). In a seminal study, Connolly et al. (1992) observed an amplitude reduction of the N200 wave for words presented in strongly constraining sentences relative to words in weakly constraining sentences. Connolly et al. (1992) interpreted that the N200 reflects the phonological processing of incoming words and the ease with which a word is processed at a phonological pre-lexical level from the preceding context (see also, Newman and Connolly, 2009; van den Brink et al., 2001; van den Brink and Hagoort, 2004).

Besides the semantic context in natural speech comprehension, visual articulatory cues can constrain the processing of ensuing speech sounds (van Wassenhove et al., 2005; Skipper et al., 2005; Sánchez-García et al., 2011, 2013). This possibility is particularly supported by the fact that visible articulations are often available temporally in advance, by tenths or even hundredths of milliseconds, of the corresponding speech sound in production (e.g., Chandrasekaran et al., 2009). For instance, van Wassenhove et al. (2005) investigated the influence of the saliency of visual articulatory cues on the processing of spoken syllables (/pa/, /ta/, /ka/) by measuring event-related potentials. van Wassenhove et al. (2005) reported a reduction in amplitude and a latency shortening of the N1/P2 complex when the auditory syllable was accompanied by the sight of the corresponding visual articulatory information. Interestingly, the size of latency shift depended on the degree of visual saliency related to the phoneme. Compared to audio-alone presentation, the audiovisual syllable /pa/ (for which the initial phoneme is highly visually salient) elicited a larger latency shift of the N1/P2 response than the audiovisual syllable /ka/ (for which the initial phoneme provides visually more ambiguous, and less salient, information).

These demonstrations thus strongly suggest that viewing speakers' lip movements can exert a facilitative influence in speech processing, and that this influence possibly expresses at a pre-lexical level by constraining speech parsing at a phonological or even pre-phonological stages (see Sánchez-García et al., 2011, 2013, for behavioral evidence). However, an intriguing question is how this visual facilitation carries over to ensuing processing stages such as for example lexical access, when speech is processed in its natural, sentential, context. Past studies have proposed that the initial portion of a spoken word determines the set of activated lexical candidates from the auditory input (Marslen-Wilson, 1987, 1990; see also for experimental evidence, Luce and Lyons, 1999; Marslen-Wilson and Zwitserlood, 1989; Spinelli et al., 2001). Therefore, visual articulatory information might also exert an influence in the generation of the set of activated lexical candidates matching with the initial portion of spoken words, leading, when highly predictable/salient, to a facilitation of word recognition. Recently, behavioral priming studies have examined this possibility by addressing whether the visual articulatory information facilitates spoken-word recognition (Buchwald et al., 2009; Kim et al., 2004). Taken together, these studies suggested that the visual articulatory information might contribute to lexical recognition (see also, Jesse and Massaro, 2010).

To address the contribution of visual speech information during word recognition, the present ERP study examines the effect of visual articulatory constraints on the processing of spoken words embedded in sentences exerting various levels of semantic constraints. To do so, we used strong and weak semantically constraining sentences whose ending was either a target word beginning with a salient visual articulatory cue, corresponding to the phoneme /p/, or a target word beginning with an ambiguous visual articulatory cue, corresponding to the phoneme /k/. Examples of the experimental stimuli are displayed in Table 1. In Experiment 1, all sentences were presented audiovisually. This design made it possible to probe interactions between sentence-level constrains and visually-driven constrains. In Experiment 2 wherein the sentences were presented in audiovisual vs. auditory-alone modality, we examined the visual influence in natural sentence contexts and any interaction between the general visual influence and the degree of visual articulatory constraints.

When presenting sentences audio-visually (Experiment 1), we expected that high semantic constraints would produce a reduction in amplitude of the N200 and N400 components as compared to low semantic constraints, an effect that previous studies have associated with pre-lexical and lexical stages of word recognition, respectively (Kutas and Hillyard, 1984: Connolly et al., 1992). In addition, if highly salient visual information (i.e. /p/) helps pre-activating a set of lexical candidates matching with the beginning of the word more efficiently, then an effect of visual articulatory constraints could be seen over the

Table 1
Examples and mean duration (in ms) of experimental conditions.

Experimental conditions	Examples	AV context duration	Target duration
High visual saliency and high semantic constraints (HV-HS)	Gritó que iba a atracar el banco, y sacó una pistola. *He shouted that he was going to rob the bank, and pulled a gun.*	2948	386
High visual saliency and low semantic constraints (HV-LS)	Como no sabía lo que podía pasar, siempre llevaba una pistola. *Not knowing what might happen, he always carried a gun.*	2830	391
Low visual saliency and high semantic constraints (LV-HS)	No hay agua caliente, creo que se ha estropeado la caldera. *There is not any hot water, I think that the boiler is damaged.*	2941	389
Low visual saliency and low semantic constraints (LV-LS)	A las ocho de la mañana, vino el técnico para intentar arreglar la caldera. *At eight o'clock morning, the technician came to try to fix the boiler.*	2891	392

AV: Audiovisual.