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

Angèle Brunellière[1], Carolina Sánchez-García[2], Nara Ikumi[2] & Salvador Soto-Faraco[2,3]

[1] Unité de Recherche en Sciences Cognitives et Affectives, University of Lille 3, France

[2] Departament de Tecnologies de la Informació i les Comunicacions, Universitat Pompeu Fabra, Barcelona, Spain

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

Please address correspondence to:
Angèle Brunellière
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
angele.brunelliere@univ-lille3.fr

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 both over 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.

213 words

Keywords: Visual speech, semantic constraints, spoken-word recognition, event-related potentials

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 & Soto-Faraco, 2013; Sumby & Pollack, 1954; McGurk & 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 & 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 & Pollack, 1954, or while perceiving non-native languages, Navarra & 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 information, like sentence meaning, both have been suggested to constrain the processing of incoming speech input in a predictive coding framework (Pickering and Garrod, 2007; Stekelenburg & 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 & Hillyard, 1984). By far, the most studied electrophysiological component during word recognition in sentential context is the N400 wave (Kutas & 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 & Phillips, 1994; Connolly et al., 1992; Connolly et al., 1990; Kutas & 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 & Phillips, 1994; Connolly et al., 1992; Connolly et al., 1990; van den Brink et al., 2001; van den Brink & Hagoort, 2004). In a seminal study, Connolly and colleagues (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 & Connolly, 2009; van den Brink et al., 2001; van den Brink & 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 & Lyons, 1999; Marslen-Wilson & 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 audio-visually. 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.

<Insert Table 1 here>

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 & 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 N200 component, at pre-lexical stage of word recognition and, over the N400 component at the lexical stage. In that case, reductions in the N200 and/or N400 amplitude should be observed for the highly salient visual information with respect to when visual information is more ambiguous. Besides the purely semantic sentence-level effect and the visually-driven effect, this study focuses on the interaction between the two, that is, how visual saliency might modulate semantic constraints as indexed by the N200 and N400 components. As the N200 and N400 reflect the ease with which a word is processed at pre-lexical and lexical levels based on prior information from the preceding context, we reasoned that the N200 and N400

amplitude reduction elicited by the semantic constraints might be stronger for the highly salient visual information with respect to the ambiguous information.

With regard to Experiment 2, we directly compared audiovisually presented sentences with sentences presented in auditory modality alone, and expected that audio-visually presented target words would produce an amplitude reduction and an earlier peak of the auditory-evoked N100 component in line with past studies discussed earlier. Moreover, the temporal facilitation of N100 peak should depend on the saliency of visual articulatory constraints, that is, stronger for the highly salient visual information (/p/) with respect to the more ambiguous information (/k/). In addition to these early effects associated with the acoustic onset of speech sound, interactive effects between the general visual influence and the visual articulatory constraints also could be observed over later stages, including the N200 and N400 components, with a stronger amplitude reduction for the highly salient visual information if the spoken-word recognition is affected by the degree of visual articulatory constraints.

Experiment 1: Visual articulatory and semantic constraints on spoken-word recognition in audiovisual sentences

Methods

Participants

Twenty Spanish-speaking students from the Pompeu Fabra University, between 18 and 34 years old (12 female, mean age: 23.6, SD age: 4.5), were selected for Experiment 1. All reported normal audition and corrected-to-normal vision. They were right-handed as assessed

by the Edinburgh handedness inventory (Oldfield, 1971). They received monetary compensation for participation (10€/hr). Before the beginning of the experiment, participants gave their written informed consent.

Stimuli

The experimental stimuli consisted of 384 sentence frames, half of which were strong semantically constraining and the other half were weak semantically constraining. More specifically, 192 doublets of sentence frames were created such that one imposed a strong semantic constrain to a final target word and the other imposed a much weaker constrain over the same final target. The target words, selected using the B-Pal Spanish lexical database (Davis & Perea, 2005), began either with the high visually salient phoneme /p/ or with a low visually salient phoneme /k/ (see, Figure 1). Studies in visual confusions and identification tasks (e.g., Auer, Bernstein, Waldstein and Tucker, 1997; van Wassenhove et al., 2005; Walden, Prosek, Montgomery, Scherr & Jones, 1977) have indeed shown that visual saliency is stronger for the bilabial plosives (/p/) in comparison to the velar plosives (/k/). The two categories of target words were matched for different psycholinguistic variables, including lexical frequency, number of syllables, number of phonemes, stress pattern, number of phonological neighbors (see Table 2), syllabic structure, imageability and concreteness. The targets were bi- or trisyllabic words with a mean number of phonemes of 5.8 (range: 4-9) and a mean logarithmic frequency of 1 (range: 0.13-1.95). As seen in Table 2, the mean logarithmic frequency for the target words with a high salient visual cue onset and those with an ambiguous visual cue onset was equivalent (1.01 and 0.99, respectively). The two categories of target words were also matched for the visual saliency associated with the phonemes following the first one, according to an adaptation to Spanish of phoneme equivalence class established by Mattys, Bernstein and Auer (2002). As all target words of interest began with a plosive (/p/ or /k/), providing a clear physical marker on the

spectrogram, a good alignment at the onset of final auditory target words for the ERP recordings was possible.

<Insert Figure 1 & Table 2 here>

The selection of the 192 doublets of strong and weak semantically constraining sentence frames resulted from the classical cloze procedure during which participants were asked to complete a given sentence fragment with the first word which comes to their mind. More specifically, seven lists of at least 100 sentence fragments were constructed and each one was completed by fifteen participants. The strong semantically constraining sentence frames were ended with the most expected word and had a cloze probability of at least 0.50 (mean: 0.84; range: 0.50-1, as measured by the questionnaire). The weak semantically constraining sentence frames had a cloze probability less than 0.33 (mean: 0.18; range: 0.07-0.33) and ended with either the expected word or an unexpected but semantically plausible word. The manipulation of cloze probability was independent of the visual saliency of the onset of spoken target word, leading to four experimental conditions (see Table 1): High visual saliency and high semantic constraints (HV-HS), High visual saliency and low semantic constraints (HV-LS); Low visual saliency and high semantic constraints (LV-HS); Low visual saliency and low semantic constraints (LV-LS). The sentence frames did not differ across the four experimental conditions in mean number of words (HV-HS: 10.07; HV-LS: 9.76; LV-HS: 9.75; LV-LS: 9.36). Two experimental lists of 48 trials per condition were constructed in order to expose participants to all experimental conditions without repeated targets. Thus, only one sentence in each doublet was exposed to any given participant. We also checked that none of the final words was repeated during the sentential context. In addition to the experimental stimuli, one hundred ninety-two congruent filler sentences were

created to prevent participants to develop strategies based on the first phoneme of final words (/p/ or /k/). The final words embedded in these filler sentences could begin with any phoneme (consonant or vowel) apart from /p/ and /k/, i.e., the initial phonemes of experimental stimuli. During the experiment, half of the presented stimuli were experimental stimuli and the rest were fillers.

For the recording of the stimuli, a Spanish-speaking female speaker was asked to pronounce the sentences several times with natural prosody at normal speaking rate. To make sure that intonation and speaking rate were kept as constant as possible, sentences within the same doublet (i.e., bearing the same target word embedded in different frames) were recorded back to back (the order of strong and weak constraining frames was counterbalanced). The video recording featured a full face frontal view of the speaker recorded simultaneously with the auditory stream during the production of sentences. The selection of the final materials was based on natural intonation and speaking rate. The selected audiovisual sentences were then clipped out (using Adobe Premiere Pro 1.5), and enveloped with a 560 ms linear fade-in ramp and a 360 ms linear fade-out ramp. The total duration of the audiovisual sentence context (up to the onset of the final auditory word) and that of the final auditory word was equivalent across experimental conditions (see Table 1). T-tests comparisons between each experimental condition revealed no significant difference in duration (see Table 3). In accordance with Chandrasekaran et al. (2009), visual motion onset preceded the auditory burst in plosives, with at least 50 ms lead. The visual motion onset was estimated by visual inspection of the video clips targeting the onset of relevant lip articulation and tongue position. The precocity of visual information in comparison to auditory information makes it logically possible that visual information might constrain the processing of incoming speech sounds.

Experimental procedure

Each trial started with a red fixation cross for 500 ms in the middle of the screen followed by the presentation of a sentence. Sounds were played binaurally at a comfortable sound pressure level via headphones and the video was played from a computer monitor placed 100 cm away from the participant. The speaker's head subtended a visual angle of 6.9° and 7.5°, respectively, for the horizontal and vertical dimension. The next trial began 3000 ms after the previous one ended. To minimize muscular artefacts, participants were asked to not move their eyes during the presentation of audiovisual sentences; when a white cross fixation appeared after the audiovisual sentence, they were free to make any movements for comfort. Participants received 24 practice sentences prior to the 6 blocks of 64 trials in the experiment, consisting of sentences from all conditions plus the corresponding fillers, presented randomly within each block. Each block was approximately 8 min long. During the experiment, participants were instructed to attentively listen to the sentences and look at the speaker's face on the screen.

EEG recording

The EEG signal was recorded in a silent room during the presentation of audiovisual sentences from 34 passive channels mounted in an elastic cap (Fp1, Fp2, Fpz, Fz, FCz, Cz, CPz, Pz, POz, Oz, F7, F3, F4, F8, FT7, FC3, FC4, FT8, T7, C3, C4, T8, TP7, CP3, CP4, TP8, P7, P3, P4, P8, PO1, PO2, O1 and O2). The channels were distributed over the head surface according to the 10% standard system of the American Electroencephalographic Society (see Figure 2). Eye movements were monitored with two channels placed close to the right eye. The on-line reference electrode was attached to the tip of the nose. The activity over right and left mastoids was also measured by two other electrodes. Electrode impedance was kept below 5kΩ during the recording. The EEG signal was filtered on-line with a bandpass of 0.1-

100Hz and digitized at 500Hz. Since we want to explore the visual constraint effects on spoken-word recognition embedded in sentence context, the ERPs were time-locked to the auditory burst of target word onset. EEG epochs starting 100 ms before the onset of the final word and ending 600 ms after were extracted from the EEG signal and averaged for each participant, condition and electrode. Prior to averaging, EEG epochs containing eye blinks and other artefacts were filtered out under an artefact rejection criterion of -/+70 μV at any channels. The average number of accepted EEG epochs did not differ across experimental conditions (HV-HS, 45.5; HV-LS, 45.4; LV-HS, 45.7; LV-LS, 45.2). The signal was filtered offline with a 1-30Hz bandpass and a notch filter at 50 Hz. A 100 ms pre-stimulus baseline correction was also applied. For each participant, bad channels were interpolated (Perrin et al., 1987) and the initial reference was changed offline to the average mastoid reference (left and right).

<Insert Figure 2 here>

ERP Data Analyses

The analyses focused on three ERP components commonly observed during spoken word recognition (N100, N200 and N400). Based on previous studies and visual inspection of ERP data, the mean amplitude of each component was extracted across participants within three time windows as follows: 110-160 ms (N100); 180-230 ms (N200); 250-500 ms (N400). While the latency window of N100 corresponds to the onset of the ascending flank and the offset of descending flank observed at Fz for the N100, that of the N200 is based on electrical fluctuations over FCz. For the N400, the window encompassed a large time range as is normally used to study N400 effects that contained the maximum peak amplitude at around 300 ms as measured in Pz (e.g., Jiang & Zhou, 2012; Ledoux et al., 2007). As Descroches et

al. (2009) and Dufour et al. (2013) highlighted the sensitivity of a late period of N400 to phonological manipulations, a late time window (520-600 ms), based on the visual inspection of ERP data, was also included to explore late effects over the N400 component. Statistical analyses were performed using 12 scalp sites (F3, FC3, Fz, FCz, F4, FC4, CP3, P3, CPz, Pz, CP4, P4) to cover the topography of the components of interest. A four-way repeated-measures ANOVA was conducted on each time window including the following within-participant variables (2x2x2x3 design): Semantic constraints (low vs. high semantic constraints), visual articulatory constraints (low vs. high visual saliency), electrode site (frontocentral vs. centroparietal) and electrode laterality (left, midline, and right). When there was more than one degree of freedom in the numerator, the Greenhouse-Geisser correction was applied (Greenhouse & Geisser, 1959) and the corrected p-values are reported.

Results

<Insert Figures 3&4 here>

Grand-average waveforms for the target words embedded in high- vs. low-semantic constraint targets are shown separately at each level of visual articulatory constraint (high, in Figure 3 and low, in Figure 4). Interestingly, there was a main effect of visual articulatory constraints ($F(1,19)=9.10$, $p<0.01$) with a greater amplitude of the N100 for the high visual saliency compared to the low visual saliency targets. This effect was mainly localized over frontocentral sites as suggested by a significant visual articulatory constraints × electrode site interaction ($F(1,19)=5.41$, $p<0.05$). Additionally, a significant visual articulatory constraints × electrode site × electrode laterality interaction ($F(2,38)=5.93$, $p<0.01$) revealed that the visual articulatory constraints × electrode site interaction occurred mainly over midline and right

hemiscalp. The main effect of semantic constraints as well as the interaction between semantic and visual articulatory constraints were not significant in this analysis (respectively, $F(1,19)=1.98$, $p>0.2$; $F(1,19)=0.01$, $p>0.2$). We estimated the N100 peak latency for each participant using the traditional approach (e.g., Miller, Ulrich, & Schwarz, 2009). An ANOVA on these latencies, over Fz and CPz, revealed no effects, including those involving the variable, visual articulatory constraints (main effect, Fz $F(1,19)=0.05$, $p>0.2$; CPz $F(1,19)=2.05$, $p=0.17$; interactive effect, Fz $F(1,19)=0.66$, $p>0.2$; CPz $F(1,19)=0.04$, $p>0.2$). In addition, peak latencies extracted using the jackknife procedure (Miller, Patterson, & Ulrich, 1998; Miller, Ulrich, & Schwarz, 2009), based on the estimation on the standard error of the component onset time, showed no significant difference over Fz and CPz, in N100 latency between the high and low visual articulatory constraints at each level of semantic constraints (high semantic constraints, Fz $t(19)=-0.85$, $p>0.2$, CPz $t(19)=1.46$, $p>0.2$; low semantic constraints, Fz $t(19)=1.09$, $p>0.2$, CPz $t(19)=0.48$, $p>0.2$).

Regarding the N200 time window, a main effect of semantic constraints was found ($F(1, 19)=12.52$, $p<0.01$) with a greater amplitude of the N200 for the low compared to high semantic constraints. An increased N200 was also observed for the high visual saliency relative to the low visual saliency targets, though mainly at the right centroparietal sites, as shown by the significant interaction involving visual articulatory constraints × electrode site × electrode laterality ($F(2,38)=3.71$, $p<0.05$). This pattern was confirmed by simple-effect tests indicating a significant effect of visual articulatory constraints (high vs. low) over right centroparietal recording sites only ($p<0.05$), whereas over left centroparietal sites the effect only approached significance ($p=0.07$) and it was not found at all over the other sites. No other interactions, including the interaction between semantic constraints and visual constraints ($F(1,19)=0.04$, $p>0.2$) and that between semantic constraints, electrode site and visual constraints ($F(1,19)=0.13$, $p>0.2$), were significant in this analysis.

In the N400 time window, there was a main effect of semantic constraints ($F(1,19)=57.61$, $p<0.001$), which followed the usual pattern (larger amplitude for low vs. highly constrained targets). As is commonly observed, the semantic constraint effect was greater at centroparietal sites as shown by a significant interaction between semantic constraints × electrode site ($F(1,19)=45.05$, $p<0.001$). There was also a significant interaction between visual articulatory constraints × electrode laterality ($F(2,38)=3.91$, $p<0.04$). This interaction indicated that the N400 deflection was stronger over midline and right hemiscalp ($p<0.05$) for the high visual saliency targets, whereas this difference in scalp distribution was not observed for the low visual saliency targets. Besides the interactions reported above, the analysis did not show an interaction between semantic constraints and visual articulatory constraints ($F(1,19)=0.12$, $p>0.2$) or between semantic constraints, electrode site and visual constraints ($F(1,19)=1.22$, $p>0.2$).

Finally, analyses on the late part of the N400 revealed main effects of semantic constraints ($F(1,19)=5.36$, $p<0.05$) and of visual articulatory constraints ($F(1,19)=6.91$, $p<0.05$). Both effects followed the main tendency detected in the earlier N400 window, namely: The late part of the N400 was larger (more negative) for low with respect to high semantic constraints, and larger for the high with respect to low visual saliency words. Note that the semantic constraints effect was limited to specific sites as suggested by a significant semantic constraints × electrode site interaction ($F(1,19)=4.14$, $p<0.05$). Indeed, this semantic effect was only observed at centroparietal sites ($p<0.01$), while no effect of semantic constraints was found at frontocentral sites ($p>0.2$). A significant semantic constraints × electrode site × electrode laterality interaction ($F(2,38)=3.93$, $p<0.05$) was also found. This was due to a semantic constraints × electrode site interaction observed only over midline and right hemiscalp (Midline, $F(1,19)=5.76$, $p<0.05$; Right, $F(1,19)=4.03$, $p=0.05$). Interestingly, a significant semantic constraints × electrode site × visual articulatory constraints interaction

was observed ($F(2,38)=4.68$, $p<0.05$). In particular, the significant semantic constraints × electrode site interaction, revealing that the semantic effect limited to the centroparietal sites, was observed for the low visual saliency targets ($F(1,19)=13.45$, $p<0.01$) but not found for the high visual saliency ones ($F(1,19)=0.26$, $p>0.2$). This interaction is illustrated in Figure 5. A negative ERP difference between low vs. high semantic constraints was only seen over centroparietal sites for the low visual saliency targets. Contrary to what was observed for the low visual saliency targets, the negative ERP difference between low vs. high semantic constraints was similar in size over the two electrode sites for the high visual saliency targets. As it can be seen in Figure 6, the ERP difference between the two levels of semantic constraints presented a negative predominance over centroparietal sites for the low visual saliency, while there was an equal distribution across the scalp for the high visual saliency targets. It seems that differently to the predicted pattern, this interaction was not explained by amplitude changes over limited sites but more globally by differential scalp distributions. Hence, the scalp distribution elicited by the semantic constraints between low and high strength depended upon the visual constraints (see, Figure 6).

<Insert Figures 5 and 6 here>

To sum-up, the manipulation of semantic constraints affected both the N200 and N400 components with greater amplitudes for the low semantically constraining sentences as traditionally observed in the literature. Thus, like previous studies, we interpret these results to show that the semantic constraints exert an influence on pre-lexical and lexical stages of word recognition. Most relevant for the purposes of this study, the effects of semantic and visual articulatory constraints interacted. This interaction was detected at the late period of the

N400 component, suggesting that visual articulatory information can constrain upcoming lexical processing by modulating semantic constraint effects. More than a direct influence over lexical stages, the visual articulatory constraints actually elicited an increase in amplitude for high visual saliency relative to the low one over all ERP components found (i.e. N100, N200, N400, lateN400). This could indicate a modulatory influence of visual articulatory constraints across the various stages of word processing. However, a main effect of visual saliency is difficult to interpret on its own in this particular experiment, because the direct comparison of audiovisual presentations of words starting with different phonemes, /p/ and /k/ may present effects due to low-level acoustic or linguistic properties independently of the visual information. By comparing between the sentences presented in audiovisual modality and auditory baseline, Experiment 2 made it possible to provide a more conclusive interpretation of the visual saliency effects reported in Experiment 1.

Experiment 2: Visual articulatory constraints on spoken-word recognition in auditory and audiovisual modalities

In Experiment 1 we found an interaction between the visual saliency and the constraints imposed by semantics, which expressed at a relatively late latency. However, the conditions included in Experiment 1 did not make it possible to ascertain the visual contribution directly as compared to purely auditory processing. In Experiment 2, we examined ERP modulations elicited by audiovisual presentations relative to the auditory-alone modality and addressed if these ERP modulations depended on the saliency of the visual articulatory constraints, so that we can further ensure the interpretation of the modulatory pattern found in Experiment 1.

Methods

Eighteen newly selected participants, following the same criteria as the ones tested in Experiment 1 participated in Experiment 2. The experimental stimuli and design was adapted from Experiment 1, only that here we did not manipulate the semantic constraints. Instead, in addition to the manipulation of visual articulatory constraints in an audiovisual condition, we included a matching auditory-only condition for comparison. The experimental blocks with auditory and audiovisual modalities of presentation were counterbalanced. While participants were asked to look at the speaker's face on the screen in audiovisual presentations, they were asked to look at a red fixation cross to avoid any eye movements in the auditory-only modality. The EEG set-up in Experiment 2 was identical to Experiment 1.

Results and discussion

In Experiment 2, the analyses on the components of interest (N100, N200, N400 and late N400) were conducted on the ERP obtained in the auditory-only and audiovisual modalities. The time windows to examine the components of interest in Experiment 2 were identical to Experiment 1. A four-way repeated-measures ANOVA was performed on the amplitude of components in each critical time window with the following factors in a 2x2x2x3 design: modality (auditory vs. audiovisual), visual articulatory constraints (low vs. high visual saliency), electrode site (frontocentral, centroparietal) and electrode laterality (left, midline and, right). Over the time window of the N100, an effect of modality only approached significance ($F(1,17)=3.35$, $p=0.08$). Interestingly, when we estimated the N100 peak latency using the traditional approach, the ANOVA showed a significantly earlier latency in the audiovisual modality relative to auditory-only modality over CPz and Fz ($F(1,17)=6.84$, $p<0.05$, $F(1,17)=5.23$, $p<0.05$, respectively). Similarly to the study of van Wassenhove et al. (2005), there was a significant interaction between modality and visual articulatory

constraints on the latency of auditory-evoked N100 response over CPz ($F(1,17)=5.37$, $p<0.05$). In particular, our data showed a temporal-facilitation effect in the audiovisual modality compared to the auditory-alone modality for the high visual saliency target words ($p<0.01$), while there was no significant speed-up of the N100 response in audiovisual modality for the low visual saliency target words ($p>0.2$). When the jackknife procedure was applied, the pattern in peak latency was replicated (CPz, high visual saliency target words, $t(17)=12.50$, $p<0.05$, low visual saliency target words, $t(17)=0.25$, $p>0.2$). This latency pattern is illustrated in Figure 7.

<Insert Figure 7 here>

As a latency effect was observed, we conducted again the four-way ANOVA of N100 amplitude but this time based on 40-ms-wide windows placed around the maximum amplitude peak over CPz. A significant effect of modality was observed ($F(1,17)=4.06$, $p<0.05$), indicating a N100 amplitude reduction in the audiovisual modality with respect to the auditory-only modality (as described in prior literature, Besle et al., 2004; Klucharev et al., 2003; van Wassenhove et al., 2005). Additionally, a significant visual articulatory constraints × electrode laterality interaction was observed ($F(2,34)=4.23$, $p<0.05$), revealing a greater amplitude of the N100 for the high visual saliency compared to the low visual saliency over left and midline hemiscalp ($p<0.05$). A supplementary analysis of N100 amplitude performed only over CPz, where the temporal-facilitation effect triggered by the audiovisual modality was observed, showed a significant effect of modality ($F(1,17)=5.83$, $p<0.05$), but did not reveal a significant interaction between the modality and the visual articulatory constraints ($F(1,17)=0.07$, $p>0.2$).

Similarly to Experiment 1, the N200 time window showed a main effect of visual articulatory constraints ($F(1,17)=6.30$, $p<0.05$), with a greater amplitude for the high vs. low visual saliency word targets. The significant interaction involving the factors, visual articulatory constraints × electrode site × electrode laterality was also found ($F(2,34)=4.34$, $p<0.05$).

In regard to the N400, the analysis revealed a significant interaction with the factors, visual articulatory constraints × electrode site × electrode laterality ($F(2,34)=4.60$, $p<0.05$). This interaction indicated that the N400 deflection was stronger over frontocentral sites at midline and left hemiscalp ($p<0.05$) for the low relative to high visual saliency targets, whereas this effect was not found over any centroparietal sites. Additionally, a significant modality × electrode site × electrode laterality interaction ($F(2,34)=4.60$, $p<0.05$) showed an amplitude-increase effect triggered by the audiovisual modality mainly over right frontocentral sites ($p<0.05$).

Over the late part of the N400, there was a significant interaction between the modality and visual articulatory constraints ($F(1,17)=9.76$, $p<0.01$). This interaction indicated that an amplitude-increase effect triggered by the audiovisual modality compared to the auditory alone modality was only found for the high visual salient targets ($p<0.05$). Indeed, this effect did not occur for the low visual saliency word targets ($p>0.2$). This amplitude pattern driven by the visual benefit of articulatory constraints is illustrated in the grand-average waveforms of the ERP subtraction between audiovisual and auditory modalities for the high- and low-visual saliency in Figure 8. A main effect of electrode site ($F(1,17)=38.34$, $p<0.001$) and a significant visual articulatory constraints × electrode site interaction ($F(1,17)=4.55$, $p<0.05$) were also found. The negativity was stronger over frontocentral sites than centroparietal sites and this effect seemed to be higher for the high visual saliency word targets. A significant modality × electrode site × electrode laterality interaction was found ($F(2,34)=3.82$, $p<0.05$),

showing that the audiovisual modality produced a greater negativity with respect to the auditory modality only at left and right hemiscalp over frontocentral sites ($p<0.05$). Finally, there was a significant articulatory constraints × electrode site × electrode laterality interaction ($F(2,34)=3.34$, $p<0.05$). The high visual saliency targets elicited a greater negative amplitude than low saliency ones at the right hemiscalp only over frontocentral sites ($p<0.05$).

<Insert Figure 8 here>

In summary, the results of Experiment 2 showed an earlier latency of peak amplitude and amplitude reduction of the auditory-evoked N100 response triggered by the audiovisual modality as already observed in prior literature. Additionally, the audiovisual modality modulated late processing stages over the N400 and its late part, producing an amplitude-increase effect. Interestingly, the effect of audiovisual modality depended on the degree of visual saliency, affecting the early and late processing stages. First, visual saliency was associated with a temporal facilitation in the auditory processing of spoken word targets. Then, visual saliency also affected word recognition processes at late stages, where the high informative visual cues produced an increased negativity in audiovisual modality when compared to auditory speech alone (see Figure 6). This latter result clarifies that the effect of visual articulatory constraints found at the late N400 in Experiment 1 could not be attributed to differences in auditory information (such as, for example, acoustic differences between /p/ and /k/). According to the results of Experiment 2, this effect detected in the late part of the N400 cannot be explained just by differences between target words, because this effect remained when the auditory-only ERP was subtracted from the audio-visual ERP. That is, when we conducted the ERP analyses on the audiovisual-auditory subtraction, we replicated

the significant difference between the high and low visual saliency target words over the late period of the N400 ($p<0.05$).

General Discussion

The present study investigated the time course of visual articulatory and semantic constraints during spoken word recognition embedded in sentential context. We recorded ERPs to determine the interplay between these two sources of information during spoken word recognition. To do so, we manipulated the strength of the semantic constraints provided by the context within the carrier sentences in which the target words were embedded, as well as the degree of visual saliency related to the first phoneme of the target words. In agreement with previous studies manipulating the semantic constraints in the auditory modality (e.g., Connolly et al., 1992), we reported, in Experiment 1, larger amplitudes of the N200 and N400 waves for words preceded by a weak semantically constraining context in comparison to words preceded by a strong semantically constraining context. From visual inspection, similarly to van den Brink et al. (2001), the observed N200 effects seemed to present a flat scalp distribution, whereas the N400 effects seemed to have a clear posterior spatial distribution. However, topographical analysis revealed that the semantic effects found in the N200 and N400 latency time windows did not have a significantly different spatial distribution (after global field power data normalization, $F(1,19)=2.69$, $p=0.15$). Thus, contrary to the visual impression, there was no conclusive evidence that the N200 effect differs topographically from the N400 effect (i.e. the activated neuronal pathways in the N200 and N400 latency time windows do not appear to be different). Such difficulties to separate the early N200 negativity from the classical N400 effects have been already reported in auditory-alone experiments using sentences by van den Brink and Hagoort (2004). In line

with these findings, Van Petten and colleagues (1999) have shown that the onset of N400 effects is related to the moment at which there is sufficient incoming signal during spoken word recognition to register a mismatch between the expected word and the incoming word.

Regarding visual articulatory constraints, in Experiment 2 we found an earlier peak of the auditory-evoked N100 response for the highly salient visual cue (/p/) in the audiovisual modality when compared to auditory-only modality. In accordance with van Wassenhove et al., (2005), the degree of visual saliency affected the time processing of auditory speech. Moreover, the N100 amplitude-reduction triggered by the audiovisual modality with respect to the auditory-only modality was independent of the degree of saliency on the basis of the incoming first phoneme of the target words. Hence, we reproduced, in a sentence context, prior findings of amplitude-reduction in audiovisual modality over the N100 auditory-evoked response commonly observed with isolated syllable or vowel presentations (Besle et al., 2004; Klucharev et al., 2003; Stekelenburg and Vroomen, 2007; van Wassenhove et al., 2005). A N100 amplitude-reduction triggered by the audiovisual modality compared to the auditory-only modality has been already described in isolated spoken word presentations by the recent ERP study of Mengin et al., (2012), but the present study is, to the best of our knowledge, the first to report evidence for a N100 amplitude-reduction and temporal facilitation triggered by audiovisual presentation in sentence context. Moreover, while visual articulatory speech is known to provide informative cues at the pre-lexical level over phonological or even pre-phonological stages (Bernstein et al. 1998; Soto-Faraco et al. 2007; Altieri et al. 2011; Sánchez-García et al., 2011), another novelty of the present results is that the ERP evidence consistent with the idea that visual speech might also have an impact during lexical processing, as initially proposed in some behavioral studies (Buchwald et al., 2009; Kim et al., 2004). In Experiment 2, we observed an amplitude-increase effect triggered by the audiovisual modality compared to the auditory alone modality over the late part of N400 time

window, whereby high visual saliency led to an increased negativity in the audiovisual modality. As the N400 is known to reflect word processing at a lexical level, this result could thus suggest that visual speech can modulate word recognition at late processing stages. Complementing this finding, evidence from Experiment 1 confirms that this visual saliency effect exerts a modulation during a stage of processing overlapping with semantic integration. Despite this argument is based on the separate findings of Experiment 1 and 2 (we did not manipulate together the semantic constraints and modality factors within a same experiment), the fact is that in both experiments we replicate an ERP effect due to the visual saliency over the same late time window usually associated with lexical processing. This suggests that visual speech information can influence word processing in audiovisual speech circumstances.

Despite we replicated the usual audiovisual effects over the early N100 component when compared to the auditory alone, the audiovisual effects over the N400 time window were perhaps less expected. In particular, over the late part of N400 wave, high visual constraints involved an amplitude increased, instead of a reduction, when comparing audiovisual with auditory alone conditions. We contend that this result might relate to recent studies showing that the late part of the N400 component is sensitive to phonological manipulations (Descroches et al., 2009; Dufour et al., 2013). During isolated spoken word recognition, Dufour et al. (2013) showed amplitude modulations over a late part of the N400 component occurring around 600 ms after word onset, coinciding with the time window of the modulation in our study. Dufour et al. (2013) found greater amplitude of the late N400 component for words residing in dense phonological neighborhoods which encounter more intense competition from activated lexical candidates with respect to words residing in sparse neighborhoods. By using a picture-word matching task during which participants had to judge whether a picture and an auditory word were identical, Descroches and colleagues (2009) had previously investigated the influence of phonological competitors on spoken word

recognition. Again, and in line with the findings of Dufour et al. (2013) the late part of N400 component around 600 ms was modulated by phonological competition based on similarity during the lexical selection of target word. Taken together, these studies thus show that the late N400 reflects the processing of phonological cues at lexical level during the selection of target word. We argue that our current modulation of visual saliency over the N400 might reflect similar phonological effects expressing at a lexical processing stage.

Interestingly, in contextual situations with a picture, Descroches and colleagues (2009) reported that the amplitude of the late N400 component was significantly increased when the auditory word target (i.e., *comb*) shared initial phonological cues but not the ending with the picture prime (i.e., a CONE), relative to a related picture prime-auditory target condition. This finding suggests that the increase of late N400 amplitude is related to the moment at which there is sufficient evidence in the spoken signal to reject the alternative lexical candidate initially activated by the picture prime. Following this view, the increase of late N400 amplitude for the highly salient visual cue (/p/) with respect to the less salient visual cue (/k/) in our study could be explained by the stronger basis for rejection of inadequate lexical candidates in word targets with highly salient visual cues. Overall, our findings seem to indicate that the visual articulatory constraints could contribute to spoken word recognition, possibly by modulating lexical selection of the target word. Note that since the early part of N400 is proposed to reflect the lexical activation by Descroches et al. (2009), it may exclude that the visual articulatory constraints might act during the stage processing by which the lexical candidates are activated. However, further studies would be necessary to attest that visual speech operates exclusively during the lexical selection.

The main findings of the present study are that the visual-speech effect expressed at late processing stage and the late interactive effects between the visual articulatory and semantic constraints. This type of modulation could only be detected because we measured

the influence of visual cues on speech processing in the context of sentence-level processing. This could explain the singularity of visual-speech effects occurring over late stages during word recognition. In particular, we believe that contrary to past audio-visual studies often restricted to single syllables or words, the fact that visual speech was presented in sentential context may have induced the processing of visual speech to permeate higher linguistic levels of processing (such as lexical selection). Thus, for now, our study makes a point proving the existence of late visual contribution during spoken word recognition, thought it would be interesting that future studies might explore in more detail whether the impact of visual speech differs as a function of linguistic level of stimuli (that is, isolated syllable or word, sentences). This, however, is beyond the scope of the present study.

In sum, visual articulatory constraints, like semantic constraints, can exert an influence on word recognition. Beyond the well-known early effects during auditory processing, our findings thus highlight a role of visually salient cues at the moment of word retrieval from the lexicon in natural speech comprehension.

Acknowledgements: This research was supported by the Spanish Ministry of Science and Innovation (PSI2010-15426 and Consolider INGENIO CSD2007-00012), Comissionat per a Universitats i Recerca del DIUE-Generalitat de Catalunya (SGR2009-092), and the European Research Council (StG-2010263145). ERP analyses were performed with the Cartool software (supported by the Center for Biomedical Imaging of Geneva and Lausanne). We are grateful to the anonymous reviewers for their helpful comments.

References:

Altieri, N., Pisoni, D.B., Townsend, J.T., 2011. Some behavioral and neurobiological constraints on theories of audiovisual speech integration: a review and suggestions for new directions. Seeing Perceiving 24, 513-539

Auer, E.T., Bernstein, L.E., Waldstein, R.S., Tucker, P.E, 1997. Effects of phonetic variation and the structure of the lexicon on the uniqueness of words. In C. Benoît & R. Campbell (Eds.) Proceedings of the ESCA/ESCOP workshop on audio-visual speech processing (pp21-24). Rhodes, Greece.

Bernstein, L.E., Demorest, M.E., Tucker, P.E., 1998. What makes a good speechreader? First you have to find one. In R. Campbell, B. Dodd, & D. Burnham (Eds.), Hearing by eye II: Advances in the psychology of speechreading and auditory–visual speech. Hove, U.K.: Psychology Press, pp. 211-228.

Besle, J., Fort, A., Delpuech, C., Giard, M.H., 2004. Bimodal speech: Early suppressive visual effects in human auditory cortex. Eur. J. Neurosci 20, 2225-2234

Biau, E., Soto-Faraco, S., 2013. Beat gestures modulate auditory integration in speech perception. Brain Lang 124, 143-152

Buchwald, A.B., Winters, S.J., Pisoni, D.B., 2009. Visual speech primes open-set recognition of spoken words. Lang. Cogn. Proc 24, 580-610

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., Ghazanfar, A.A., 2009. The natural statistics of audiovisual speech. PLoS Comput. Biol 5, e1000436

Connolly, J.F., Phillips, N.A., 1994. Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. J. Cogn. Neurosci 6, 256-266

Connolly, J.F., Phillips, N.A., Stewart, S.H., Brake, W.G., 1992. Event-related potential sensitivity to acoustic and semantic properties of terminal words in sentences. Brain Lang 43, 1-18

Connolly, J.F., Stewart, S.H., Phillips, N.A., 1990. The effects of processing requirements on neurophysiological responses to spoken sentences. Brain Lang 39, 302-318

Davis, C.J., Perea, M., 2005. BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. Behav. Res. Methods 37, 665-671

Desroches, A.S., Newman, R.L., Joanisse, M.F., 2009. Investigating the time course of spoken word recognition: Electrophysiological evidence for the influences of phonological similarity. J. Cogn. Neurosci 21, 1893-1906

Dufour, S., Brunellière, A., Frauenfelder, U.H., 2013. Tracking the time course of word frequency effects in auditory word recognition with event-related potentials. Cogn. Science 34, 489-507

Greenhouse, S.W., Geisser, S., 1959. On methods in the analysis of profile data. Psychometrika 24, 95-111

Jesse, A., Massaro, D.W., 2010. The temporal distribution of information in audiovisual spoken-word identification. Atten. Percept. Psychophys 72, 209-225

Jiang, X., Zhou, X., 2012. Multiple semantic processes at different levels of syntactic hierarchy: Does the higher-level process proceed in face of a lower-level failure? NeuroImage 50, 1918-1928

Kim, J., Davis, C., Krins, P., 2004. Amodal processing of visual speech as revealed by priming. Cognition 93, B39–B47

Klucharev, V., Möttönen, R., Sams, M., 2003. Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. Brain Cogn. Res 18, 65-75

Kutas, M., Hillyard, S.A., 1984. Brain potentials during reading reflect word expectancy and semantic association. Nature 307, 161-163

Ledoux, K., Gordon, P.C., Camblin, C.C., Swaab, T.Y., 2007. Coreference and lexical repetition: mechanisms of discourse integration. Mem. Cognit 35, 801-815

Luce, P.A., Lyons, E.A., 1999. Processing lexically embedded spoken words. J. Exp. Psychol. Hum. Percept. Perform 25, 174-183

Ma, W.J., Zhou, X., Ross, L.A., Foxe, J.J., Parra, L.C., 2009. Lip-reading aids word recognition most in moderate noise: A Bayesian explanation using high-dimensional feature space. PLoS One 4, e4638

Marslen-Wilson, W.D., 1987. Functional parallelism in spoken word-recognition. Cognition 25, 71-102

Marslen-Wilson, W.D., 1990. Activation, Competition, and Frequency in Lexical Access. In G. T. M. Altmann (Ed.), Cognitive Models of Speech Processing: Psycholinguistics and Computational Perspectives. Cambridge, MA: The MIT Press, pp.148-172.

Marslen-Wilson, W.D., Zwitserlood, P., 1989. Accessing spoken words: The importance of word onsets. J. Exp. Psychol. Hum. Percept. Perform 15, 576-585

Mattys, S. L., Bernstein, L. E., Auer, E. T., 2002. Stimulus-based lexical distinctiveness as a general word recognition mechanism. Percept Psychophys 64, 667-679

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.

Mengin, O., Flitton, A., Jones, C.R.G., de Haan, M., Baldeweg, T., Charman, T., 2012. Audiovisual speech integration in autism spectrum disorders: ERP evidence for atypicalities in lexical-semantic processing. Autism Res 5, 39-48

Miller, J., Patterson, T., Ulrich, R. 1998. Jackknife-based method for measuring LRP onset latency differences. Psychophysiology 35, 99-115

Miller, J., Ulrich, R., Schwarz, W. 2009. Why jackknifing yields good latency estimates. Psychophysiology 46, 300-312

Navarra, J., Soto-Faraco, S., 2007. Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. Psychol. Res 71, 4-12

Newman, R.L., Connolly, J.F., 2009. Electrophysiological markers of pre-lexical speech processing: Evidence for bottom–up and top–down effects on spoken word processing. Biol. Psychol, 80, 114-121

Oldfield, R.C.,1971. The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia 9, 97-113

Perrin, F., Pernier, J., Bertrand, O., Giard, M-H., Echallier, J.F., 1987. Mapping of scalp potentials by surface spline interpolation. Electroencephalogr. Clin. Neurophysiol 66, 75-81

Pickering, M.J., Garrod, S., 2007. Do people use language production to make predictions during comprehension? Trends Cogn. Sci 11, 105-110

Ross, L.A., Saint-Amour, D., Leavitt, V.M., Javitt, D.C., Foxe, J.J., 2007. Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cereb. Cortex 17, 1147-1153

Sánchez-García, C., Alsius, A., Enns, J. T., Soto-Faraco, S., 2011. Cross-modal prediction in speech perception. PLoS One 6, e25198

Sánchez-García, C., Enns, J.T., Soto-Faraco, S., 2013. Cross-modal prediction in speech depends on prior linguistic experience. Exp. Brain. Res 225, 499-511

Shahin, A.J., Kerlin, J.R., Bhat, J., Miller, L.M., 2012. Neural restoration of degraded audiovisual speech. NeuroImage 60, 530-538

Skipper, J.I., van Wassenhove, V., Nusbaum, H.C., Small, S.L., 2005. Hearing lips and seeing voices: How cortical areas supporting speech production mediate audiovisual speech perception. Cereb. Cortex 17, 2387-2399

Soto-Faraco, S., Navarra, J., Weikum, W.M., Vouloumanos, A., Sebastián-Gallés N., Werker, J.F., 2007. Discriminating languages by speech-reading. Percept. Psychophys 69, 218-231

Spinelli, E., Segui, J., Radeau, M., 2001. Phonological priming in spoken word recognition with bisyllabictargets. Lang. Cogn. Proc, 16, 367-392

Stekelenburg, J.J., Vroomen, J., 2007. Neural correlates of multisensory integration of ecologically valid audiovisual events. J. Cogn. Neurosci 19,1964-1973

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am 26, 212-215

van Berkum, J.J, Brown, C.M., Zwitserlood, P., Kooijman, V., Hagoort, P., 2005. Anticipating Upcoming Words in Discourse: Evidence From ERPs and Reading Times. J. Exp. Psychol. Learn. Mem. Cogn 31, 443-467

van den Brink, D., Brown, C.M., Hagoort, P., 2001. Electrophysiological evidence for early contextual influences during spoken-word recognition: N200 versus N400 effects. J. Cogn. Neurosci 13, 967-985

van den Brink, D., Hagoort, P., 2004. The influence of semantic and syntactic context constraints on lexical selection and integration in spoken-word comprehension as revealed by ERPs. J. Cogn. Neurosci 16, 1068-1084

Van Petten, C., Coulson, S., Rubin, S., Plante, E, Parks, M., 1999. Time course of word identification and semantic integration in spoken language. J. Exp. Psychol. Learn. Mem. Cogn 25, 394-417

van Wassenhove, V., Grant, K.W., Poeppel, D., 2005. Visual speech speeds up the neural processing of auditory speech. Proc. Natl. Acad. Sci USA 102, 1181-1186

Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr, C.K., Jones, C.J., 1977. Effects of training on the visual recognition of consonants. J. Speech. Hear. Res 20, 130-145

Figure captions

Figure 1. Illustrations of audiovisual stimuli showing snap shots extracted from two sentence examples with various levels of visual articulatory constraints, in which the initial phoneme of the target words was either high visually salient, /p/, or low visually salient, /k/.

Figure 2. Distribution of the 34 channels across the scalp. In red dotted line, the groupings of electrodes included in the statistical analysis.

Figure 3. Grand-average waveforms for the high visual saliency targets according to the semantic constraints (low vs. high semantic constraints) in Experiment 1. The time windows of ERP components were: 110-160 ms (N100; blue); 180-230 ms (N200; grey); 250-500 ms (N400; pink) and 520-600 ms (late N400; green) from the onset of the auditory target word.

Figure 4. Grand-average waveforms for the low visual saliency targets according to the semantic constraints (low vs. high semantic constraints) in Experiment 1. The time windows of ERP components were: 110-160 ms (N100; blue); 180-230 ms (N200; grey); 250-500 ms (N400; pink) and 520-600 ms (late N400; green) from the onset of the auditory target word.

Figure 5. Mean amplitude (error bars) of the ERP difference between low and high semantic constraints over the late N400 component showing each electrode site and each level of visual articulatory constraints in Experiment 1.

Figure 6. Subtraction maps illustrating the ERP difference between low and high semantic constraints at each level of visual articulatory constraints in Experiment 1 and the visual ERP benefit between high and low visual saliency in Experiment 2.

Figure 7. Grand-average waveforms over CPz for target words in audiovisual and auditory-only modalities within each level of visual articulatory constraints (high and low visual saliency) in Experiment 2.

Figure 8. Grand-average waveforms for the ERP difference between audiovisual and auditory-only modalities in high and low visual saliency in Experiment 2. The time windows of ERP components were: 110-160 ms (N100; blue); 180-230 ms (N200; grey); 250-500 ms (N400; pink) and 520-600 ms (late N400; green) from the onset of auditory target word.

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

Table 2. Psycholinguistic properties of spoken target words

	Log Frequency	Number of syllables	Number of phonemes	Stress pattern	Number of phonological neighbors
High visual salient words	1.01	2.58	5.92	1.69	5.5
Low visual salient words	0.99	2.49	5.74	1.63	6.53

Table 3. Statistical results of AV context and target duration across experimental conditions

Audiovisual context duration		
HV-HS vs. HV-LS	t(190)=1.50	p>0.2
HV-HS vs. LV-HS	t(190)=-0.07	p>0.2
HV-HS vs. LV-LS	t(190)=1.52	p=0.19
HV-LS vs. LV-HS	t(190)=1.3	p>0.2
HV-LS vs. LV-LS	t(190)=-0.01	p>0.2
LV-HS vs. LV-LS	t(190)=1.28	p>0.2
Target duration		
HV-HS vs. HV-LS	t(190)=-0.44	p>0.2
HV-HS vs. LV-HS	t(190)=-0.35	p>0.2
HV-HS vs. LV-LS	t(190)=-0.48	p>0.2
HV-LS vs. LV-HS	t(190)=-0.21	p>0.2
HV-LS vs. LV-LS	t(190)=-0.05	p>0.2
LV-HS vs. LV-LS	t(190)=-0.25	p>0.2

Figure 1

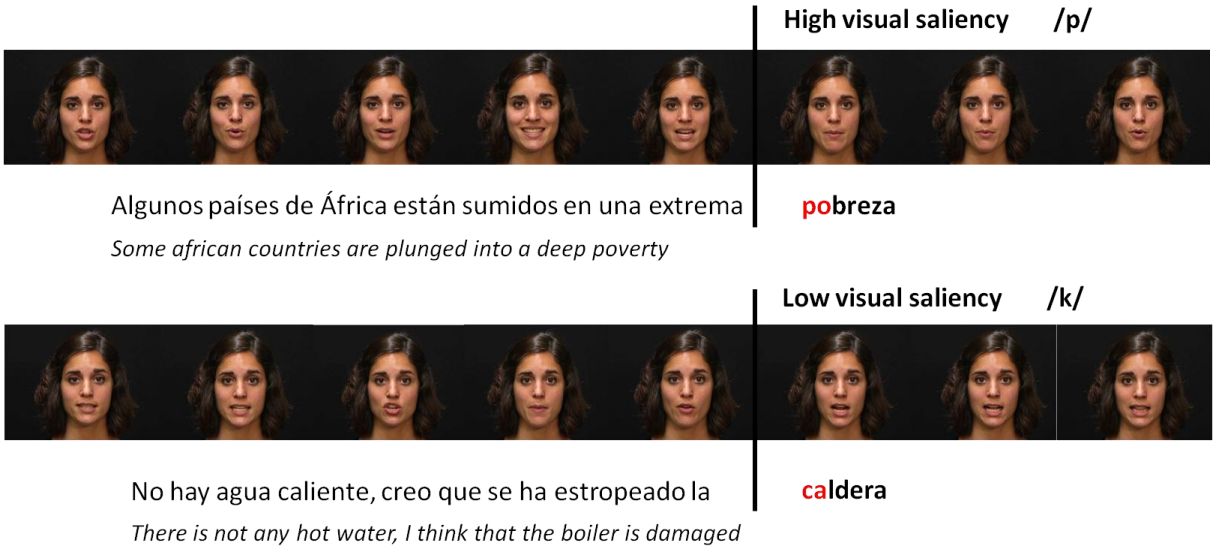

High visual saliency /p/
Algunos países de África están sumidos en una extrema pobreza
Some african countries are plunged into a deep poverty
Low visual saliency /k/
No hay agua caliente, creo que se ha estropeado la caldera
There is not any hot water, I think that the boiler is damaged

Figure 2

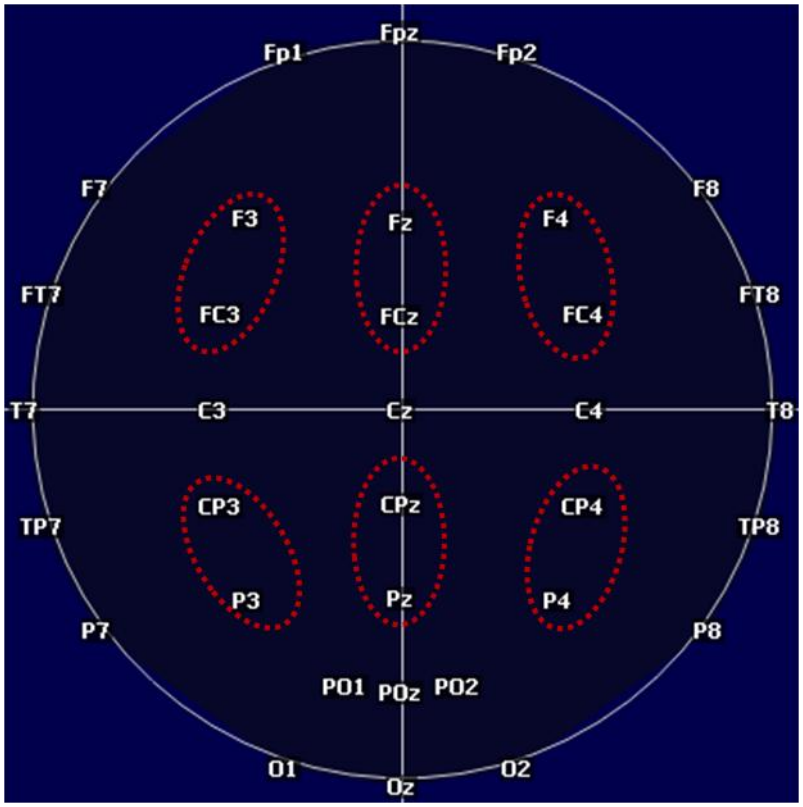

Fpz
Fp1
Fp2
F7
F8
F3
Fz
F4
FT7
FT8
FC3
FCz
FC4
T7
C3
Cz
C4
T8
CP3
CPz
CP4
TP7
TP8
P3
Pz
P4
P7
P8
PO1
POz
PO2
O1
Oz
O2

Figure 3

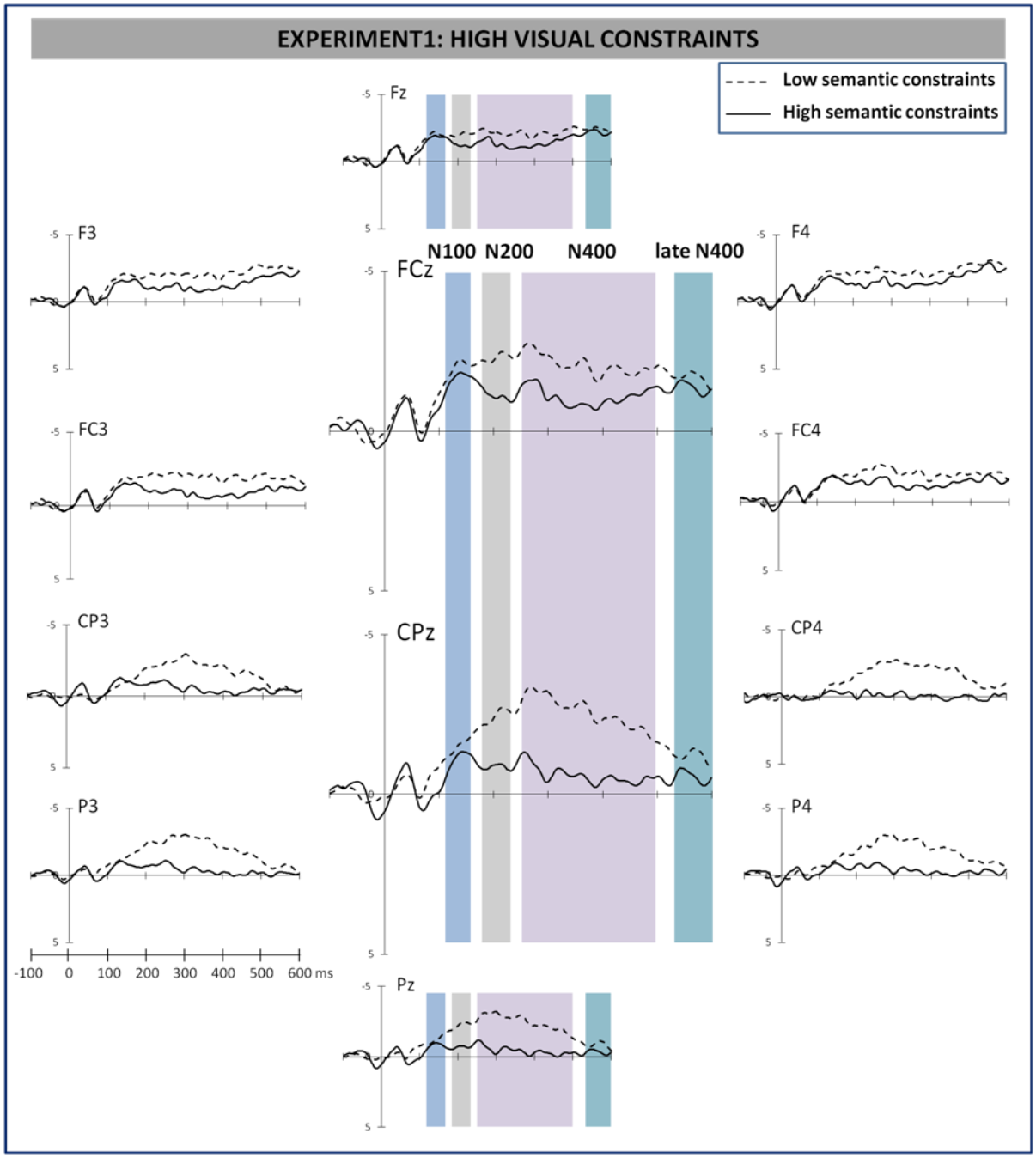

EXPERIMENT1: HIGH VISUAL CONSTRAINTS
Low semantic constraints
High semantic constraints
N100
N200
N400
late N400
Fz
F3
F4
FCz
FC3
FC4
CP3
CPz
CP4
P3
P4
Pz
-100 0 100 200 300 400 500 600 ms

Figure 4

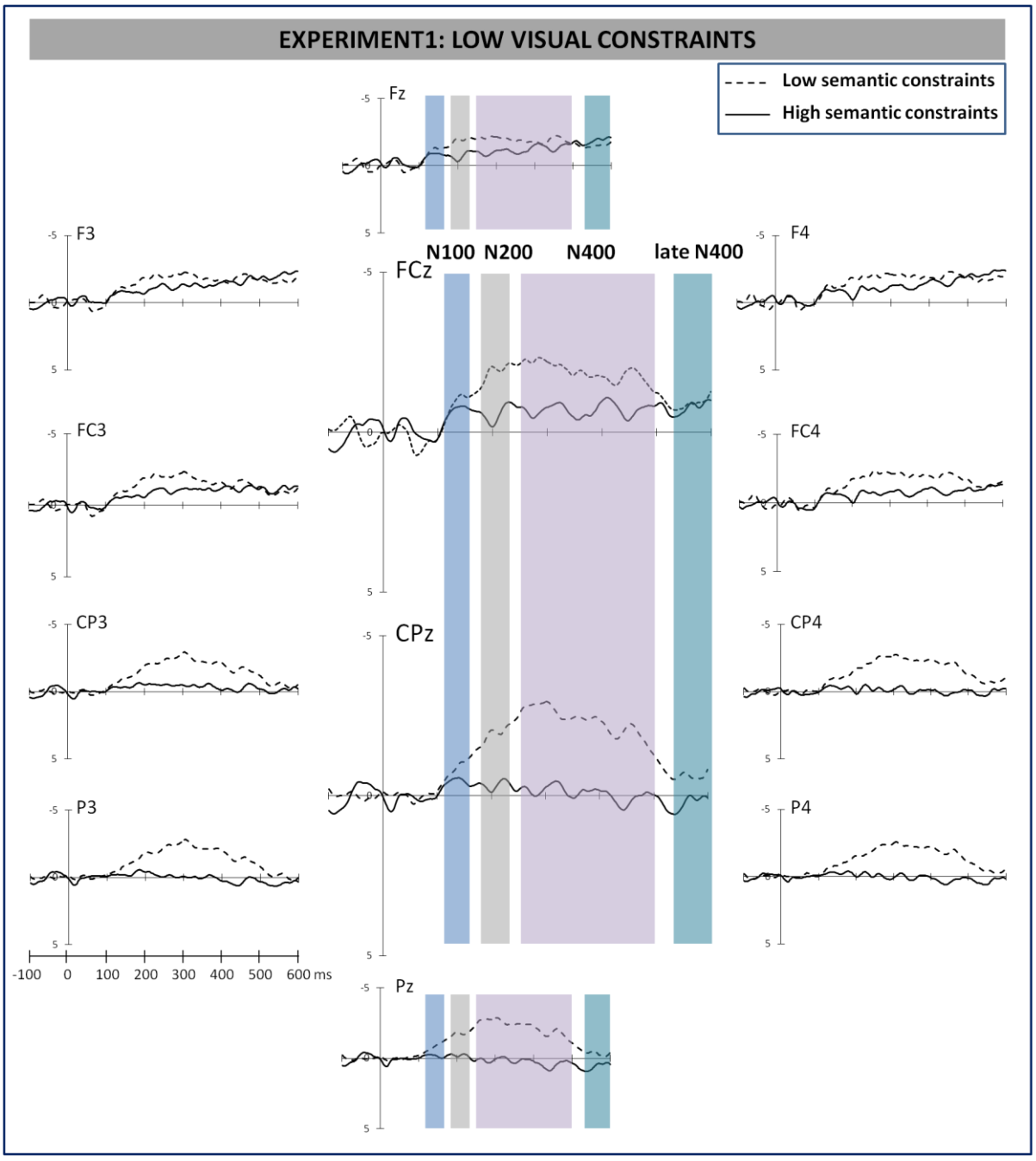

EXPERIMENT1: LOW VISUAL CONSTRAINTS
Low semantic constraints
High semantic constraints
N100
N200
N400
late N400
Fz
F3
F4
FCz
FC3
FC4
CPz
CP3
CP4
P3
P4
Pz
-5
5
-100
0
100
200
300
400
500
600 ms

Figure 5

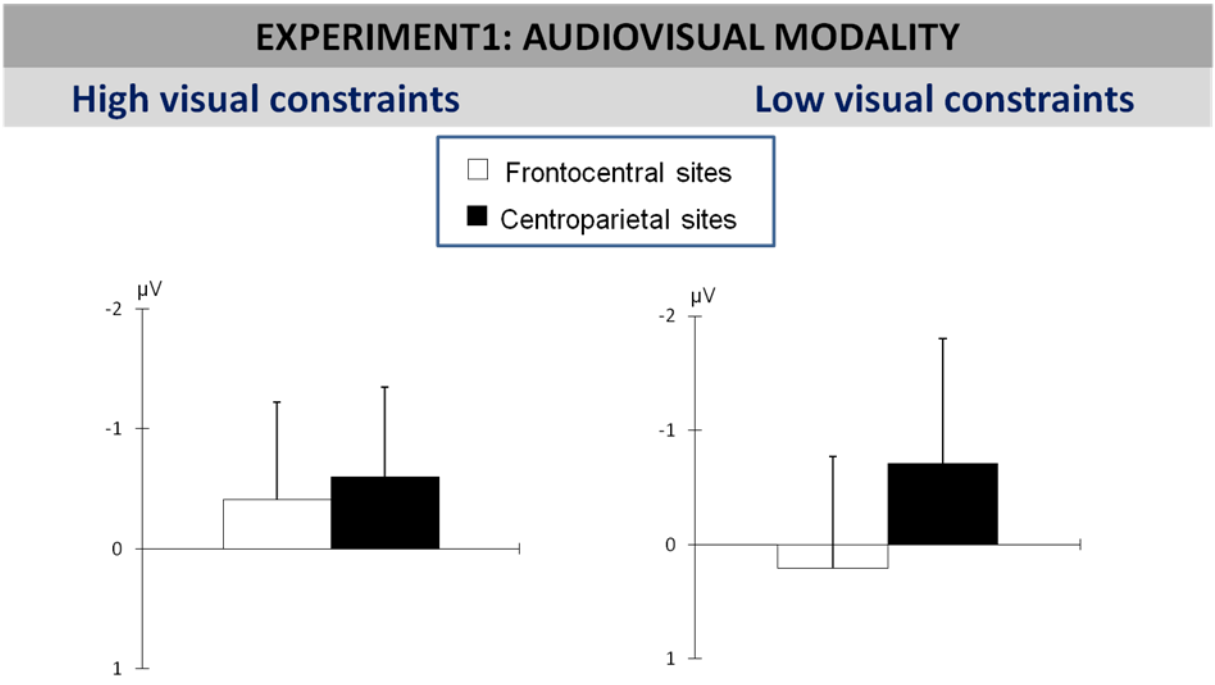

EXPERIMENT1: AUDIOVISUAL MODALITY
High visual constraints
Low visual constraints
Frontocentral sites
Centroparietal sites
μV
-2
-1
0
1
μV
-2
-1
0
1

Figure 6

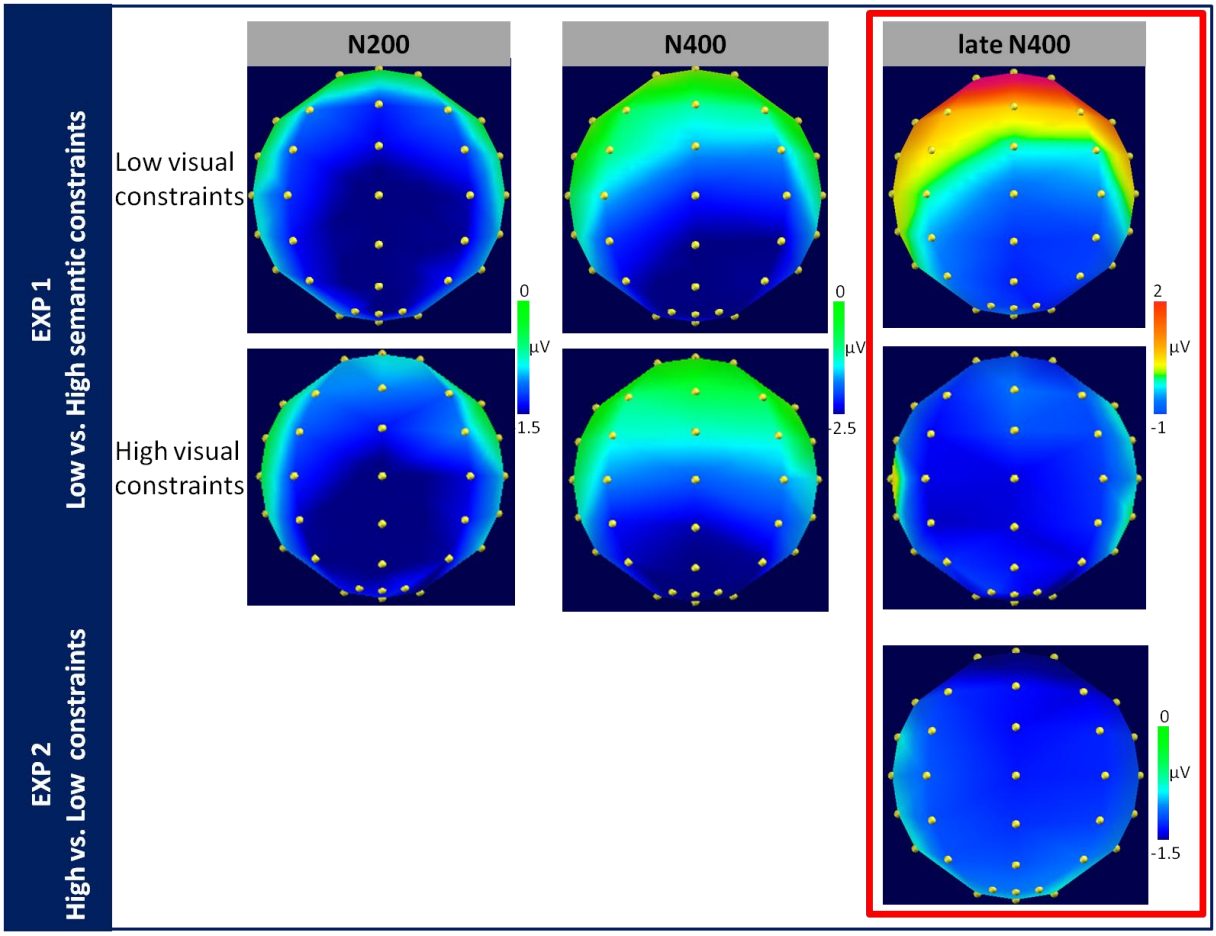

EXP 1
Low vs. High semantic constraints
EXP 2
High vs. Low constraints
N200
N400
late N400
Low visual constraints
High visual constraints
0
µV
1.5
0
µV
-2.5
2
µV
-1
0
µV
-1.5

Figure 7

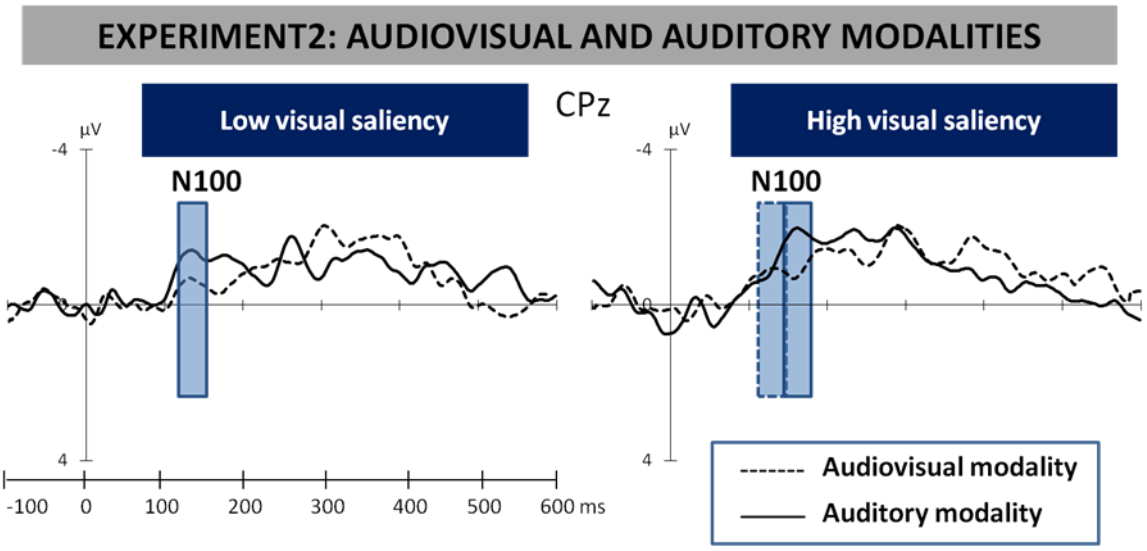
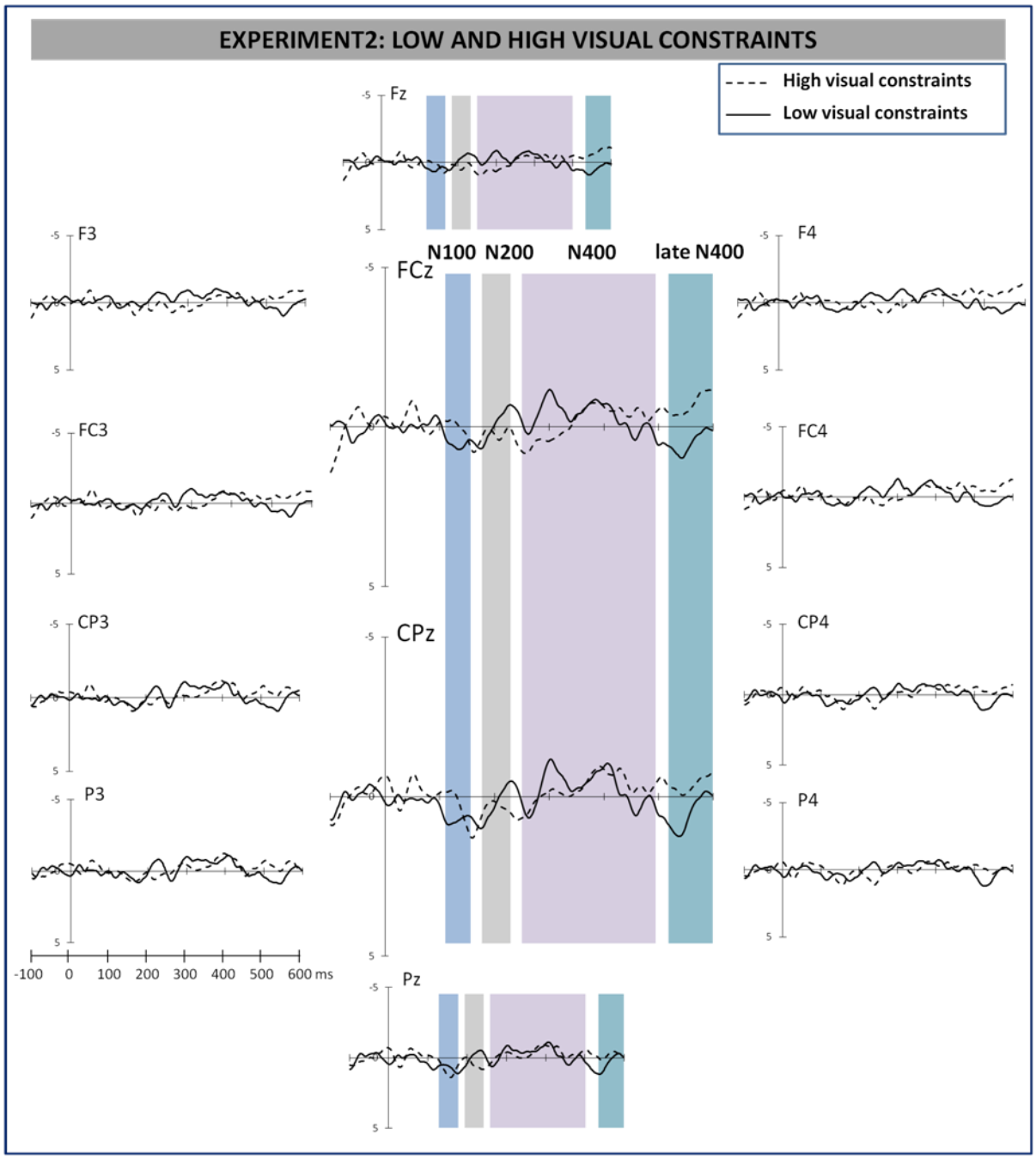

EXPERIMENT2: AUDIOVISUAL AND AUDITORY MODALITIES
Low visual saliency
CPz
High visual saliency
µV
-4
N100
4
-100
0
100
200
300
400
500
600 ms
Audiovisual modality
Auditory modality

Figure 8


EXPERIMENT2: LOW AND HIGH VISUAL CONSTRAINTS
High visual constraints
Low visual constraints
N100
N200
N400
late N400
Fz
F3
F4
FCz
FC3
FC4
CP3
CPz
CP4
P3
P4
Pz
-5
5
-100
0
100
200
300
400
500
600 ms