Title: The contribution of audiovisual speech to lexical-semantic processing in natural spoken sentences

Authors: Angèle Brunellière[1], Laurence Delrue[2], Cyril Auran[2]

Affiliation:

[1] Univ. Lille, CNRS, CHU Lille, UMR 9193 - SCALab - Sciences Cognitives et Sciences Affectives, F-59000 Lille, France

[2] Univ. Lille, CNRS, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France

Please address correspondence to:
Angèle Brunellière
SCALab, CNRS UMR 9193, Université de Lille, Domaine universitaire du Pont de Bois, BP 60149, 59653 Villeneuve d'Acsq, France
Tel: (+33) 3 20 41 72 04
angele.brunelliere@univ-lille.fr

Abstract:

In everyday communication, natural spoken sentences are expressed in a multisensory way through auditory signals and speakers' visible articulatory gestures. An important issue is to know whether audiovisual speech plays a main role in the linguistic encoding of an utterance until access to meaning. To this end, we conducted an event-related potential experiment during which participants listened passively to spoken sentences and a lexical recognition task. The results revealed that N200 and N400 waves had a greater amplitude after semantically incongruous words than after expected words. This effect of semantic congruency was increased over N200 in the audiovisual trials. Words presented audiovisually also elicited a reduced amplitude of the N400 wave and a facilitated recovery in memory. Our findings shed light on the influence of audiovisual speech on the understanding of natural spoken sentences by acting on the early stages of word recognition in order to access a lexical-semantic network.

Count: 150 words

Keywords: audiovisual speech, sentence context, spoken-word recognition, lexical recognition memory, event-related potentials

Introduction

In everyday life, auditory signals are usually accompanied by speakers' visible articulatory gestures. Since human speech is mostly multisensory, a major challenge is to better understand how continuous speech is perceived and analyzed to build the meaning of a sentence from different sensory modalities. Remarkably, the processing of natural spoken sentences depends on interconnection between the sub-lexical, lexical, and sentence-level processes. An intriguing question is to know whether audiovisual speech can exert an influence on the interplay between the processes involved at the different levels (sub-lexical, lexical, sentence level) in natural spoken sentences. However, since the role of audiovisual speech has been mostly studied by considering processing levels of phonemes, syllables and words accompanied by explicit tasks or presented in noisy contexts, the contribution of audiovisual speech to the comprehension of meaning of natural spoken sentences is still unresolved.

It is well known that auditory-visual interactions help disambiguate the perception of speech sounds in noisy contexts (e.g., Ma, Zhou, Ross, Foxe, & Parra, 2009; MacLeod & Summerfield, 1987; Massaro, 1998; Ross, Saint-Amour, Leavitt, Javitt, & Foxe 2007; Sumby & Pollack, 1954; Summerfield & McGrath, 1984) and in non-native languages (e.g., Navarra & Soto-Faraco, 2007). Auditory-visual interactions also can improve speech perception, with the result that reaction times in tasks of syllable and vowel identification were faster when they were presented in an audiovisual modality than in an auditory-only modality (e.g., Besle, Fort, Delpuech, & Giard, 2004; Klucharev, Möttönen, & Sams, 2003). Moreover, electrophysiological recordings demonstrated that auditory-visual interactions even sped up the latency of auditory event-related potentials (Alsius, Möttönen, Sams, Soto-Faraco, & Tiippana, 2014; Arnal, Morillon, Kell, & Giraud, 2009; Stekelenburg & Vroomen, 2007; van

Wassenhove, Grant, & Poeppel, 2005). Although a speeded-up processing of the unfolding speech sounds can be accounted for theoretically by the complementary effect of auditory and visual information, some authors showed that viewing visual movements predicting in advance speech sounds is necessary to observe the speeded-up processing of speech sounds (Stekelenburg, & Vroomen, 2007). In natural face-to-face communication, visual information can sometimes be provided in advance of speech sounds. For instance, visible articulatory movements can precede by tenths or even hundredths of milliseconds the occurrence of corresponding speech sounds in preparatory sequences or the start of a speech utterance (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). However, Schwartz and Savariaux (2014) showed that the temporal relationship between auditory cues and visible articulatory movements is more complex and very variable depending on the different phonemes and on their position within an utterance, including a range of audiovisual asynchronies varying from small auditory lead (50 ms) to large visual lead (200 ms).

Speakers' orofacial movements particularly indicate their lip movements (Benoit, Guiard-Marigny, Le Goff, & Adjoudani, 1996) and somewhat the movements of their teeth and tongue (Badin, Tarabalka, Elisei, & Bailly, 2010; McGrath, 1985). Since speakers' orofacial movements provide some reliable markers of phonological information, the notion of visual phonemes or visemes was proposed (Fisher, 1968), such that visible speech gestures can be produced in a visual distinctive opposition from other phonemes (for example, /p/ and /b/ versus /f/ and /v/). Therefore, the audiovisual presentation of incoming speech input may trigger the activation of sub-lexical phonological representations sharing a visual similarity with the speech gestures and acoustic properties with the auditory input. This impact at the sublexical level can thereafter constrain the activation of lexical candidates. A series of behavioral studies has examined the contribution of visual speech to spoken-word recognition (Buchwald, Winters, & Pisoni, 2009; Fort, Kandel, Chipot, Savariaux, Granjon, & Spinelli,

2013; Dodd, Oerlemans, & Robinson, 1989; Kim, Davis, & Krins, 2004; Mattys, Bernstein, & Auer, 2002; Tye-Murray, Sommers, & Spehar, 2007). Several priming studies have demonstrated that visual speech primes auditory lexical targets in a variety of tasks including semantic categorization and word identification (Buchwald et al., 2009; Fort et al., 2013; Dodd et al., 1989; Kim et al., 2004). Additionally, Mattys, Bernstein, & Auer (2002) showed that visual spoken words were recognized more accurately when they had few visual articulatory competitors (i.e. words that looked like few other words in terms of visemes) than when they had many visual articulatory competitors. Visual articulatory neighbors (i.e. words that look like other words in terms of visemes) thus act on the lexical competition between candidates, suggesting that the visual neighborhood density (i.e., number of words that can be created from a target word by adding, deleting, or substituting a single viseme) can have an impact on speechreading. Tye-Murray, Sommers, & Spehar (2007) replicated the findings of Mattys, Bernstein, & Auer (2002) in visual speech and they also showed that words were recognized more correctly in the audiovisual modality when they had few items at the overlap between the auditory and visual phonological neighborhood densities compared to many items that were both auditory and visual neighbors. Taken together, these findings indicate that both auditory and visual information seem to constrain the activation of lexical candidates, thus influencing lexical competition.

However, how continuous speech is perceived and analyzed to build the meaning of a sentence from different sensory modalities (auditory and visual) is less understood. In this context, the present study explored the impact of audiovisual speech on the interplay between spoken-word recognition and processing at the sentence level in natural spoken sentences in order to access a lexical-semantic network. In light of findings showing that both auditory and visual information can constrain the activation of lexical candidates (Buchwald et al., 2009; Fort et al., 2013; Dodd et al., 1989; Kim et al., 2004; Mattys et al., 2002; Tye-Murray et al.,

2007), we could hypothesize that audiovisual speech affects the interplay between spoken-word recognition and processing at the sentence level by increasing the efficiency in combining the analysis of an incoming word with the contextually based constraints from the meaning of the utterance. In line with this hypothesis, visual speech can guide phonological and timing predictions about the occurrence of upcoming speech sounds, so that audiovisual speech offers substantial benefits of processing in the encoding of the speech by constraining the number of possible candidates in a spoken utterance (for a review, Peelle & Sommers, 2015).

Nonetheless, this hypothesis was drawn from studies that did not focus on sentence processing or use explicit tasks. It should be noted that the integration of auditory and visual information in different speech materials is underpinned by various mechanisms (e.g., Grant & Seitz, 1998; Van Engen, Xie, and Chandrasekaran, 2017). It was found that an individual's susceptibility to the McGurk effect (the perceptual fusion between incongruent auditory, i.e. /ba/ and visual, i.e. [ga]) in isolated syllables did not show any relationship with their ability to use visual cues to understand spoken sentences in noise (Grant & Seitz, 1998; Van Engen et al., 2017). Moreover, our hypothesis of whether audiovisual speech affects the interplay between spoken-word recognition and processing at the sentence level can be challenged by the recent proposal that the role of visual speech might be preferentially limited to the recognition of phonemes (Baart & Samuel, 2015). Studies with behavioral and electrophysiological recordings (e.g., Samuel & Lieblich, 2014; Baart & Samuel, 2015) provided experimental evidence for effects of lip-reading and lexical contexts operating differently on speech sound processing. According to Baart & Samuel (2015), information obtained from visual speech might contribute more to the recognition of phonemes than lexical access and understanding of the meaning of sentences, because visual speech operates particularly along the dorsal pathway of speech perception (Hickock and Poeppel, 2007). In

the dual stream model of speech perception (Hickock and Poeppel, 2007), early stages of speech processing are thought to occur bilaterally in the auditory regions of the dorsal superior temporal gyrus (STG) and superior temporal sulcus (STS). Thereafter, two streams are thought to diverge. The dorsal pathway connects the temporal cortex to the premotor cortex including the parietal-temporal junction for auditory-to-motor mapping. The posterior portion of the STS plays a role in multisensory integration during speech perception (Beauchamp, Argall, Bodurka, Duyn, & Martin, 2004; Calvert, Campbell, & Brammer, 2000; Sekiyama, Kanno, Miura, & Sugita, 2003; Wright, Pelphrey, Allison, McKeown, & McCarthy, 2003) and some neuroimaging studies have shown the involvement of the motor regions in audiovisual speech processing (Fridriksson, Moss, Davis, Baylis, Bonilha, & Rorden, 2008; Hall, Fussell, & Summerfield, 2005; Skipper, Nusbaum, & Small, 2005). In contrast, the second pathway, which is called the ventral pathway, and is located in the temporal lobe supports auditory-to-meaning mapping (i.e., lexical and meaning access). The anatomical and functional distinction between the two streams of speech perception suggests that audiovisual speech helps in the recognition of phonemes rather than in the encoding of the spoken utterance at higher levels (i.e., lexical and meaning comprehension).

In the present study, event-related potentials (ERPs) were used to study cognitive processes in real time with high temporal resolution after semantic anomalies. We thus tracked ERPs in response to the congruency of a final word in the context of the sentence presented either in the auditory-only modality or in the audiovisual modality. While early auditory event-related potentials (ERP), namely N100 and P200, are associated with perceptual and sub-lexical processing, two negative waves (N200, N400) are known to be sensitive to sentence-level processes when a word is semantically incongruous with the previous context in auditory sentences.

In audiovisual speech, the N100 and P200 are attenuated in amplitude when the visual articulatory information is in accordance with auditory information compared to an auditory-only stimulation of syllables (Besle et al., 2004; Klucharev et al., 2003; Pilling, 2009; van Wassenhove et al., 2005). Additionally, the peak of the auditory event-related potentials occurred earlier for audiovisually presented syllables than for the auditory-only presentation of syllables (van Wassenhove et al., 2005). This speeding-up of auditory event-related potentials was also higher for /p/ more visually salient versus /k/ visually ambiguous according to the principle of visual saliency. Stekelenburg and Vroomen (2007) found a similar speeding-up and suppression of auditory N100 amplitude in audiovisual speech perception was found with nonspeech events when the visual information contained anticipatory motion. In contrast, the audiovisual impact on auditory N100 was not observed when the visual information did not predict in advance the speech sound (Stekelenburg, & Vroomen, 2007).

In auditory sentences, a negative wave, peaking around 200 ms, is typically increased in amplitude in semantically anomalous conditions when the word onset is different from that of the word expected from the sentence (e.g., Connolly and Phillips, 1994; Connolly, Stewart, & Phillips, 1990 ; van den Brink and Hagoort, 2004, van den Brink, Brown and Hagoort, 2001). This first negative shift elicited after semantic anomalies is traditionally called the N200 or PMN (phonological mismatch negativity) and is specifically found in the spoken modality (van den Brink and Hagoort, 2004). The N200 indexes phonological processing after the recognition of initial phonemes in contact with the lexical- and sentence-level processes. Therefore, when the initial phonemes of the perceived word do not match the initial phonemes of the expected word from the sentence constraints, a negative shift is elicited (Connolly, Phillips, Stewart, & Brake, 1992).

In addition to the N200, a large negativity occurs later around 400 ms after semantic anomalies with a posterior distribution across the scalp. This second shift is associated with lexical access in both written and spoken modalities (Kutas & Federmeier, 2000) and is known as the N400. Unlike the N200, the N400 component reflects the processing of words at the lexical level in contact with sentence-level processes (Hagoort and Brown, 2000). Hence, both the N200 and the N400 are associated with the interplay between spoken-word recognition and processing at sentence level to access a lexical-semantic network when comparing the processing of incoming words expected from a sentence context to that of semantically incongruous words.

In our previous study (Brunellière, Sánchez-García, Ikumi & Soto-Faraco, 2013) in which we explored the electrophysiological correlates underlying sentence processing with audiovisual speech, we found amplitude modulations of N100, N200 and N400 waves. In the first experiment, we manipulated the semantic constraints of sentence (strong predictable words versus weakly predictable words) within audiovisual sentences as well as visual salience of word onset (/p/ versus /k/). We observed that the high visual salience elicited an increase in amplitude relative to the low one across all ERP components found (i.e., N100, N200, N400). Moreover, the long-lasting N400 after weakly predictable words compared to strongly predictable words was stronger in amplitude and more spatially distributed across the scalp at a late time window when the visual salience was high. In the second experiment, we probed more precisely the effect of visual salience by presenting the sentences in audiovisual or auditory-only modalities without manipulating the semantic constraints. In that case, we replicated the N100 findings of van Wassenhove et al. (2005). While we found the typical amplitude-reduction N100 response independently of visual salience, there was only a temporal facilitation of the auditory-evoked N100 response triggered by audiovisual presentation when the visual salience was high. Thereafter, the audiovisual modality

modulated late processing stages over the late N400 part, producing an increase in amplitude for the high visual salient targets. We interpreted that the highly salient visual cue (/p/) might tend to keep activated lexical candidates sharing the same visual onset for a longer time, thereby producing more lexical competition and thus a cost for lexical selection. In line with this hypothesis, the late interaction between visual salience and semantic constraints in the first experiment could reflect the resolution of this cost thanks to the semantic constraints. Therefore, although the findings of two experiments suggest an impact of audiovisual speech depending on the visual salience at the sublexical and lexical levels, they do not provide direct evidence for a contribution of audiovisual speech to the interplay between spoken-word recognition and processing at the sentence level.

To investigate this issue, we tracked ERPs responses after expected and incongruous words in the context of semantically constraining sentences that were presented in audiovisual and auditory-only modalities. If audiovisual speech were to have an impact on lexical-semantic processing, one would expect that the biphasic negative shift with increased amplitude over the N200 and N400 components for semantically incongruous words in comparison with expected words would be stronger when the sentences are presented audiovisually than when they are presented in an auditory-only modality. On the contrary, if audiovisual speech primarily contributes to the recognition of phonemes rather than the linguistic encoding of the utterance, audiovisual speech would have an impact on the amplitude of the N100, N200 and the N400 waves but without interacting with processing at the sentence level (i.e., lower amplitude when sentences are presented in the audiovisual modality than when presented in an auditory-only modality).

To gain further insight into the impact of audiovisual speech in lexical-semantic processing, we also probed the interactive effects between the semantic congruency of words

regarding the sentence context and the modality of presentation on memory representations in the lexicon. Neville, Kutas, Chesney, & Schmidt (1986) found that old words that previously occurred in sentences during a listening task were better recognized when they fitted with the sentence context than those that did not. We explored whether this effect was strengthened by audiovisual speech. Up to now, richer episodic memory representations have been only found in the audiovisual modality than in the auditory-only modality in free-recall tasks (Goolkasian & Foos, 2005; Thompson & Paivio, 1994; Pichora-Fuller, 1996).

Methods and Materials

Participants. Thirty-two healthy, native monolingual French speakers with normal or corrected-to-normal vision participated in this experiment. They were recruited at the University of Lille and declared no hearing or language impairments as per self-report. All were right-handed as assessed by the Edinburgh handedness inventory (Oldfield, 1971). The participants included 20 women and 12 men, with a mean age of 21.4 (range: 18-34 years, SD=3). Before the beginning of the experiment, participants gave their written informed consent, which was conducted in accordance with the Declaration of Helsinki. The experiment was approved by the Research Ethics Committee of the University of Lille. Participants received monetary compensation for participation (10€) or course credits.

Stimuli. The experimental stimuli were composed of a set of 312 sentences where the context was highly constraining semantically and each context sentence ended either in the final expected noun or in a semantically incongruent noun. The selection of sentence context was based on the classical cloze procedure and the quality of the stimuli during their recordings (normal rate speech, clear productions of words). In addition to the manipulation of the semantic congruency, the whole set of sentences was presented either audiovisually or in the auditory-only modality. By manipulating two factors (semantic congruency and modality of presented), we obtained four experimental conditions (see, Table 1): Auditory-only and expected word (AO-Expected), auditory-only and incongruous word (AO-Incongruous), audiovisual and expected word (AV-Expected) and audiovisual and incongruous word (AV-Incongruous). During the classical cloze procedure, ninety participants not included in the experiment were asked to complete sentence fragments with the first noun that came to mind. The final noun, accompanied by its article, was missing within each sentence fragment. The mean cloze probability was 0.82 (range: 0.53-1) for the 312 selected sentence contexts and the mean length in terms of number of words was 15.3 (range: 8-34). The final words (expected

and incongruous nouns) shared the same gender and the same number, so that there was no violation in terms of grammatical information. Expected and incongruous words differed from the first phoneme, in terms of semantic information. They were matched for lexical frequency, number of phonemes, number of syllables, number of phonological neighbors, and uniqueness point (see, Table 2). The onset of final words was always a plosive (/p/, /t/, /k/), making it easy to align the event-related potentials time-locked to the auditory onset of the final word. Expected and incongruent words were equally likely to begin with /p/, /t/ or /k/ (110 /p/, 114 /k/, 88 /t/). The initial phonemes of the expected and incongruent words thus were visually equally salient. Moreover, the three types of initial phonemes differed in terms of viseme categories (Benoit, Lallouache, Mohamadi, & Abry, 1993; Fisher, 1968; Istria, Nicolas-Jeantoux, & Tamboise, 1982) such that /p/ was replaced with /t/ or /k/ in the incongruous version of sentences. Moreover, 85% of expected words completely differed from incongruent words which replaced them in the incongruous sentences in terms of viseme categories during the unfolding of words.

< Insert Tables 1 and 2 here >

To avoid exposing participants to repeated presentations of the same sentence context, we constructed four experimental lists composed of 78 trials per experimental condition (AO-Expected, AO-Incongruous, AV-Expected, AV-Incongruous). Each sentence context was presented once in one experimental condition per participant and was presented in all conditions to all participants. Eight participants were exposed to one experimental list (i.e., a total number of 32 participants). In addition to the experimental stimuli, we constructed 156 semantically congruent sentences (e.g., « Elle raffole des livres de cet auteur », *She is mad about the books by this author*). These sentences were used as fillers to prevent strategies

related to semantic violations. For all participants, 33.3% of all presented stimuli were semantically incongruent sentences.

All stimuli were produced several times by a French-speaking woman. The speaker was asked to pronounce the sentences with natural prosody at normal speaking rate. To make sure that intonation and speaking rate were kept constant between the two versions of sentence context, ended either in the expected word or the incongruous word, the stimuli were with pairs recorded one after the other. Within each sentence context, the speaker first pronounced the expected version of sentence context twice and then the incongruous version twice, or she proceeded in reverse order. The order of expected and incongruous versions was counterbalanced across the whole set of sentence contexts. The full face of the speaker in frontal view was recorded simultaneously with the auditory stream during sentence production. The audiovisual recordings were clipped out using AVID and a 375 ms linear fade-in ramp and a 250 ms linear fade-out ramp were added. The selection of audiovisual sentences was based on natural intonation and speaking rate. Using Praat (Boersma and Weenink, 2011), we measured the total duration of each sentence, of sentence contexts preceding the critical word and of the final word. *T*-tests comparisons between the expected and the incongruous versions of sentences revealed no significant difference in duration (total duration of sentences, $t(622)=0.27$, $p=.87$; total duration of sentence contexts, $t(622)=-0.70$, $p=.785$; duration of final auditory words $t(622)=0.16$, $p=.482$). In addition, *t*-tests comparisons between the expected and the incongruous versions of sentences showed no significant difference in mean intensity (sentence context, $t(622)=-0.86$, $p=.388$; final words, $t(622)=-0.88$, $p=.378$). Moreover, measures of the mean fundamental frequency of sentence contexts and of final auditory words did not differ significantly as a function of semantic congruency (sentence context, $t(622)=-0.91$, $p=.36$; final words, $t(622)=-0.99$, $p=.21$)". The visual onset of the final word was estimated by visual inspection of the video clips targeting

the onset of relevant lip articulation and tongue position. The visual information preceded the auditory information with a mean lead of 35.3 ms. This earliness of visual information compared to auditory information did not differ as a function of the semantic congruency (incongruous vs. expected versions of sentences, $t(622)=0.45$, $p=.65$).

Experimental procedure. Each trial began with a red fixation cross presented in the center of the monitor for 500 ms, followed by the presentation of a sentence. At the end of a sentence, a black screen appeared for 1000 ms, which was then replaced by a white fixation cross in the center of the monitor for 1000 ms. The auditory stream was presented binaurally at a comfortable sound level via earphones, and the video stream was played on a computer monitor placed 100 cm away from the participant. Sounds were presented through Sennheiser earphones (cx281) at approximately 70 dB sound pressure level at a sample rate of 44.1 kHz. Videos were displayed centered on a 14.1-inch monitor with a refresh rate of 60 Hz. To minimize ocular artefacts, participants were asked to maintain their gaze at the center of the screen and were encouraged to avoid making movements until the white fixation cross was displayed. While their EEG activity was being recorded, participants were asked to listen to the sentences without engaging in any other tasks (for similar approaches, Hagoort and Brown, 2000, van den Brink and Hagoort, 2004, van den Brink, Brown and Hagoort, 2001). They were exposed to 24 practice sentences prior to the set of six blocks of 52 trials. Each block lasted around 12 minutes and was composed of all experimental conditions and fillers presented in random order. While three blocks presented sentences audiovisually, three others presented sentences in an auditory-only modality. The order of blocks was counterbalanced so that the modality of sentence presentation changed after each block. The block order was also randomized.

After the listening task, participants performed a lexical recognition task on a set of 144 written words presented one after the other in the center of the monitor. The presentation of written words made it possible to access the memory trace of words in the lexicon and to avoid familiarity effects associated with the acoustic or phonetic properties when the words had previously been heard. During the lexical recognition task, each trial began with a fixation cross for 500 ms followed by the presentation of a written word in the center of the monitor. After a 2,000-ms intertrial interval, another fixation cross was presented. Participants were asked to indicate, as quickly and as accurately as possible, whether they had heard the word or not during the listening task by pressing one of the two buttons on a response box. The response buttons were counterbalanced across all participants. The lexical recognition task lasted around 15 minutes. Half of the words were presented as final nouns (72 old words), while the other half never occurred in the sentences during the listening task (72). Among the words which had not been uttered in the sentences, half of them were new (36) and the other half (36) were expected words from the sentence context, although not presented at the end of the sentence (i.e., in the incongruous version). The words expected but not presented during the listening task were words expected, from either the auditory (18) or the audiovisual sentence context (18). As for the words which had been presented previously as final nouns during the listening task (72 old words), half were expected words from the sentence context (18 words delivered in either an audiovisual modality or an auditory-only modality) and the other half were incongruous words (18 words in each modality of sentence presentation). Therefore, there were four experimental conditions for the old words. The words selected for the lexical recognition task were matched for lexical frequency, number of phonemes and letters, number of syllables, number of phonological and orthographical neighbors, and uniqueness point. This behavioral task allowed us to make sure that participants had paid attention to the words embedded in the spoken sentences by estimating

d-prime using hit responses (i.e., presented words during the sentence listening task and to which participants pressed the button corresponding to 'heard words') and false alarms (i.e., words that had not been presented during the sentence listening task but to which participants pressed the button corresponding to 'heard words'). Interestingly, this task offered the opportunity to probe an effect of richer episodic memory representations for old words according to their modality of presentation and their fitting with the contextually based constraints from the meaning of the utterance during the listening task. Therefore, we performed two-way repeated measures ANOVA on hit rates with the modality of presentation (audiovisual vs. auditory-only modality) and the semantic congruency of the final word (expected vs. incongruous words).

EEG recording and analyses. The EEG signal was recorded on the scalp using a 128-channel Biosemi Active Two AD-box. Two electrodes placed close to the right eye were used to record eye movements. The electrical signal of two additional electrodes placed over the right and left mastoids was also measured. Individual electrodes were adjusted to a stable offset lower than 20 mV during the EEG recording. The electrical signal was digitized at 1024 Hz. The EEG epochs started 100 ms before and lasted 800 ms after the auditory onset of final word. Each epoch was corrected to a 100-ms baseline and was filtered offline with a 0.01–30 Hz band-pass filter and a 50 Hz notch filter. To eliminate eye blinks and other artifacts, the epochs were removed under a rejection criterion of +/-70 µV at any channel within the period of epochs. The ERP waveforms, which were time-locked to the auditory onset of final words, were calculated for every participant, experimental condition, and electrode. The total number of accepted epochs was equal across the experimental conditions (mean accepted epochs, AO-Expected: 71, AO-Incongruous: 70.9, AV-Expected: 71.4, AV-Incongruous: 70.9). Bad channels were interpolated for each participant (Perrin, Pernier, Bertrand, Giard, & Echallier,

1987) and the EEG signal was referenced offline to an average mastoid reference (left and right). The mean number of the interpolated channels was 2.8.

The analyses focused on three ERP components commonly observed during spoken word recognition (N100, N200, and N400). The mean amplitude of each ERP component was extracted across the participants within three time windows in which the maximum peak amplitude was as follows: 100-140 ms (N100); 160-210 ms (N200); 250-600 ms (N400). As in the study by van den Brink, Brown and Hagoort, (2001), the N200 time window was determined from the onset of the first ascending slope until the offset of the descending one in semantically incongruous conditions compared to congruous ones. The latency of the N400 time window was associated with the onset of the ascending slope around 250 ms and the N400 time window corresponded to a large standard time window for measuring the N400 effect (Lau, Phillips, & Poeppel, 2008). We computed three-way repeated measures ANOVA on the mean amplitude of each time window according to semantic congruency of the final noun, modality of presentation[1] and topography (see Table 3). As in previous studies (van den

1

Previous studies examining the N100 explored when multisensory integration takes place (e.g., Besle et al., 2004; Klucharev et al., 2003; Pilling, 2009, van Wassenhove et al., 2005). They used the rationale of the additive model (i.e., differences between the summed unimodal activity and the activity generated by the audiovisual condition). However, this approach can lead to biases (Besle, Fort, & Giard, 2004; Stekelenburg & Vroomen, 2007, and Teder-Salejarvi, McDonald, & Di Russo & Hillyard, 2002). Biases come from the assumption that unimodal auditory stimuli and unimodal visual stimuli are independently processed. In fact, a common activity including attentional modulation, working memory or any higher cognitive processes may be associated with the processing of both types of stimuli (auditory and visual). This issue is very problematic when investigating effects after 200 ms from stimulus onset where the higher cognitive processes are likely to occur. Interestingly, Baart (2016) demonstrated that the suppression of N100/P200 in amplitude and the speeding up of N100/P200 latencies by audiovisual speech were not modulated by whether the visual-only condition was subtracted or not from the audiovisual condition. For all these reasons, it was legitimate to compare audiovisual and auditory-only trials in the present study.

Brink, Brown and Hagoort, 2001; Van den Brink & Hagoort, 2004), the statistical analyses were performed using different topographical sites[2] (left anterior, right anterior, frontocentral, centroparietal, left parietal, right parietal and occipito-parietal) to cover the topography of the components of interests. Each topographical site was composed of seven individual electrodes. To adjust to violations of sphericity (Greenhouse & Geisser, 1959), the Greenhouse-Geisser correction was applied when there was more than one degree of freedom in the numerator. The corrected *p* values are reported. When a significant interaction was found, pairwise Tukey *t*-tests were conducted to interpret the significance of the elicited effects.

Results

Behavioral results. Participants performed the lexical recognition task above chance level (Mean, d'=1.1, $t(31)$=14.2, $p=4.1\times10^{-13}$), suggesting that they paid attention to the words embedded in the spoken sentences during the listening task. The statistical analysis of old words on hit rates revealed main effects of Semantic Congruency ($F(1,31)$=16.55, MSE=227.7, $p=3\times10^{-4}$) and Modality ($F(1,31)$=6.17, MSE=136, p=.019). As seen in Figure 1, participants recognized expected words better than incongruous words. Additionally, words were better recalled when they had been previously presented in audiovisual trials than in an auditory-only modality. No significant interaction between Semantic Congruency and Modality was found ($F(1,31)$=1.11, MSE=95.9, p=.3). Then, to test whether the contextually based constraints from the meaning of the utterance affected the retrieval of the word memory traces, we compared the participants' performance on the 18 new words to that for the words

[2] Left anterior: D10, D11, D12, D13, D14, D18, D19, right anterior: B20, B21, B29, B30, B31, B32, B22, frontocentral: D2, C2, FCz, C24, C22, C11, Fz, centroparietal: Cz, A2, CPz, B1, B2, D15, D16, left parietal: A17, A16, A9, A8, A7, A6, D29, right parietal: A30, A29, B3, B4, B5, B6, B13 and occipito-parietal: A5, A18, A20, Pz, Poz, A31, A32.

that not presented and had been expected, from either the auditory or audiovisual sentence context by conducting one-way repeated measures ANOVA on false alarms rates. This analysis showed a main effect of participants' performance for new words, and those for the two types of expected (but not presented) words ($F(2,62)=14.59$, $MSE=127.04$, $p=6\ 10^{-7}$). Paired Tukey *t*-tests comparisons indicated that participants had more false alarms for expected (but not presented) words than for new words (see Figure 1, for expected words from the auditory-only modality of sentence context presentation versus new words, $p=4.97\ 10^{-4}$; for expected words from the audiovisual modality of sentence context presentation versus new words, $p=1.24\ 10^{-4}$). There was no significant difference between the two types of expected (but not presented) words which differed between them only in the modality of sentence context presentation ($p=.55$).

< Insert Figure 1 >

ERP results. The grand-average ERP waveforms per experimental condition are shown in Figure 2. A first negative response peaks around 100 ms (N100) for all experimental conditions and is then followed by a second negative peak elicited by the semantically incongruent words around 200 ms (N200). The amplitude of this latter component seemed to be stronger for semantically incongruent words than for expected words. A large negativity response appeared between 250 and 600 ms and its amplitude was more pronounced for semantically incongruent words as traditionally observed in the literature for the N400 wave after semantic violations. Based on visual inspection, it appeared that audiovisual speech affected the amplitude of both the N100, N200 and N400 waves.

< Insert Figure 2 >

< Insert Table 3 >

Impact of audiovisual speech on sentence processing at perceptual level

To explore the impact of audiovisual speech in sentence processing at perceptual level, we examined the mean amplitude of the N100 wave (see Figures 2 and 3). Over the N100 time window, the ANOVA revealed main effects of Semantic Congruency ($F(1,31)=8.17$, $MSE=75.53$, $p=.007$, $\eta^2_p=.21$) and of Topography ($F(6,186)= 8.81$, $MSE=7.52$, $p=9.3\times10^{-5}$, $\eta^2_p=.22$). As seen in Figures 2 and 3, the amplitude of the N100 wave was stronger for incongruous words than for expected words. Paired Tukey *t*-tests comparisons regarding the topography factor showed that there were more negative values over the centroparietal sites[3] relative to the left and right anterior sites and the left and right parietal sites ($p<.05$), as shown in Figure 3. Moreover, we found no main effect of Modality ($F(1,31)=1.26$, $MSE=71.84$, $p=.27$, $\eta^2_p=.039$), but a significant interaction between Modality and Topography ($F(6,186)=2.85$, $MSE=4.79$, $p=.049$, $\eta^2_p=.08$). The amplitude of N100 was greater in the audiovisual modality than in the auditory-only modality, only over the frontocentral site ($p=.0013$). As seen in Table 3, no other significant interactions were found. The same statistical analysis was performed over the positivity wave (P50) placed before the N100 time window. This analysis also revealed a main effect of Semantic Congruency ($F(1,31)=7.27$, $MSE=27.95$, $p=.011$, $\eta^2_p=.19$) and a significant interaction between Modality and Topography ($F(6,186)=3.28$, $MSE=3.09$, $p=.039$, $\eta^2_p=.095$) with the same pattern as the N100. There was

[3] ANOVA analysis exclusively based on the mean amplitude of centroparietal sites where the N100 amplitude was found to be the strongest revealed no significant effect of Modality ($F(1,31)=1.86$, $MSE=25.78$, $p=.18$, $\eta^2_p=.005$). The same analysis based on the mean amplitude or the peak amplitude over CPz and Cz again showed no significant effects of Modality (CPz, mean amplitude, $F(1,31)=0.35$, $MSE=2.07$, $p=.56$, $\eta^2_p=.011$, CPz peak amplitude, $F(1,31)=0.003$, $MSE=2.15$, $p=.99$, $\eta^2_p=10^{-5}$, Cz, mean amplitude, $F(1,31)=2$, $MSE=9.1$, $p=.17$, $\eta^2_p=.06$, Cz, peak amplitude, $F(1,31)=1.1$, $MSE=10.98$, $p=.32$, $\eta^2_p=.03$). Furthermore, an analysis based on the peak latency over CPz and Cz again revealed no main effect of Modality (CPz, $F(1,31)=1$, $MSE=407$, $p=.33$, $\eta^2_p=.03$, Cz, $F(1,31)=3.1$, $MSE=327$, $p=.10$, $\eta^2_p=.02$).

an increased N100 amplitude after incongruous words as compared to expected words and the N100 amplitude was greater in the audiovisual modality than in the auditory-only modality over the frontocentral sites (see Figure 3).

< Insert Figure 3 >

Impact of audiovisual speech on sentence processing at sub-lexical, lexical, and sentence-levels

The N200 and the N400 are associated with the interplay between spoken-word recognition and processing at sentence level to access a lexical-semantic network. If audiovisual speech had an impact on lexical-semantic processing, one would expect that the biphasic negative shift with increased amplitude over the N200 and N400 components for semantically incongruous words in comparison with expected words would be stronger when the sentences are presented audiovisually than when they are presented in an auditory-only modality. Over the N200 time window, we found a main effect of Semantic Congruency ($F(1,31)=31.35$, $MSE=143{,}02$, $p=4\times10^{-6}$, $\eta^2_p=.5$) and a significant interaction between Semantic Congruency and Modality ($F(1,31)=5.97$, $MSE=30.96$, $p=.02$, $\eta^2_p=.16$). As seen in Figures 2 and 4, the amplitude of the N200 wave was stronger for incongruous words than for expected words. Paired Tukey *t*-tests comparisons revealed only differences between incongruous words and expected words ($p<.001$). The enhanced negativity for incongruous words (as compared to expected words) was greater in the audiovisual trials than in the auditory-only modality ($t(31)=2.44$, $p=0.021$, $d=.38$). There was also a main effect of Topography ($F(6,186)=8.58$, $MSE=8.41$, $p=9\times10^{-6}$, $\eta^2_p=.22$) and a significant interaction between Semantic Congruency and Topography ($F(6,186)=4.19$, $MSE=2.38$, $p=.016$, $\eta^2_p=.12$). Paired Tukey *t*-tests comparisons showed that there were more negative values over the centroparietal sites relative to the anterior and parietal sites (including the left and right side). The enhanced

negativity for incongruous words than for expected words was stronger over centroparietal, occipito-parietal sites, left and right parietal sites than over left and right anterior sites (p=2.6×10^{-6}). As seen in Figure 4C, the amplitude of N200 seemed to be greater after incongruous words in the audiovisual trials than in the auditory-only modality over centroparietal sites. In contrast, a lower amplitude of N200 for expected in the audiovisual trials than in the auditory-only modality seemed to appear over occipito-parietal sites. Supplementary ANOVAs over centroparietal and occipito-parietal sites where the effect of semantic congruency was the strongest revealed a significant interaction between Semantic Congruency and Modality for each one (centroparietal, F(1,31)=7.07, MSE=6.196, p=.01, η^2_p=.18, occipito-parietal, F(1,31)=7.74, MSE=6.03, p=.009, η^2_p=.2). Moreover, paired t-test comparisons over the centroparietal sites indicated that the amplitude of N200 after incongruous words was stronger in the audiovisual modality than in the auditory-only modality (p=.04) as shown in Figure 4. There was a lower amplitude of N200 after expected words in the audiovisual modality than in the auditory-only modality over the occipito-parietal sites (p=.02).

< Insert Figure 4 >

As with the N200 time window, a main effect of Semantic Congruency (F(1,31)=101.51, MSE=55.7, p=2.6×10^{-11}, η^2_p=.77) was observed over the N400. This increased N400 amplitude for incongruous words relative to expected words was topographically localized, as suggested by the significant interaction between Semantic Congruency and Topography (F(6,186)=14.42, MSE=3.59, p=2.2×10^{-7}, η^2_p=.32). The increased N400 amplitude for incongruous words (as compared to expected words) was stronger over the centroparietal and occipito-parietal sites than the other sites (p<.05). Moreover, while no main effect of Modality was found (F(1,31)=0.3, MSE=55.7, p=.59, η^2_p=.01), there was a significant interaction

between Modality and Topography ($F(6,186)=8.88$, $MSE=3.59$, $p=.0001$, $\eta^2_p=.22$). The amplitude of N400 was reduced in the audiovisual trials as compared to the auditory-only modality, only over the left and right parietal sites and the occipito-parietal sites (respectively, $p=.0015$, $p=.017$, $p=.017$), as shown in Figure 5. There was also a main effect of Topography ($F(6,186)=45.08$, $MSE=8.51$, $p=2.1\times10^{-14}$, $\eta^2_p=.59$). Paired *t*-tests comparisons showed that values were more negative over the anterior, frontocentral and centroparietal sites than over the left and right parietal sites and the occipito-parietal sites ($p<.05$). No other significant interactions were found.

< Insert Figure 5 >

To confirm the apparent topographical differences between N200 and N400, we compared mean amplitudes after normalization to the global field power between these two ERP components and performed a four-way repeated measures ANOVA using component (N200 versus N400), semantic congruency of the final noun, modality of presentation and topography. The ANOVA analysis showed a significant interaction between Component and Semantic Congruency ($F(1,31)=18.77$, $MSE=8.31$, $p=1.4\times10^{-4}$, $\eta^2_p=.38$) revealed a stronger effect of semantic congruency over the N400. Incongruous words elicited a greater amplitude of N400 than that of N200 ($p=.02$) and expected words elicited a lower amplitude of N400 than that of N200 ($p=.006$). Significant interactions between Component and Topography ($F(6,186)=24.69$, $MSE=1.31$, $p=2.3\times10^{-9}$, $\eta^2_p=.44$) and between Component, Semantic Congruency and Modality ($F(1,31)=4.76$, $MSE=14.95$, $p=.04$, $\eta^2_p=.13$) were also shown. The topography of N400 differed from that of N200 in that more negative responses were found over the left anterior and frontocentral sites (see Figure 6). Whereas values of N200 were more negative over the centroparietal sites than over the anterior and parietal sites (including the left and right side) and the occipito-parietal sites ($p<.05$), the topography of N400 showed

more negative values over left anterior, frontocentral and centroparietal sites than over the left and right parietal sites and the occipito-parietal sites (p<.05) as seen in Figure 6. As observed in separate analysis on each ERP time window (i.e., N200 and N400) and in Figures 4 and 5, the significant interaction between Component, Semantic Congruency and Modality indicated that although the effect of semantic congruency was greater in the audiovisual trials than in the auditory-only trials over N200, the effect of semantic congruency was equal whatever the modality of presentation over N400. Taken together, this analysis confirmed that the N200 wave is a separate ERP component from the N400 wave.

< Insert Figure 6 >

Discussion

The present study investigated whether speakers' orofacial movements contribute to the interplay between the stages of spoken-word recognition and the sentence-level processes to build the meaning of sentences. After a large sample of participants had listened passively to natural spoken sentences presented either in the audiovisual or the auditory-only modality, we recorded ERPs elicited by contextually expected and semantically incongruous words. The rationale was that if audiovisual speech affected the interplay between spoken-word recognition and processing at sentence level, the audiovisual modality of presentation would increase the efficiency of combining the analysis of the incoming word with the contextually based constraints from the meaning of the utterance. As a result, an increase would be observed in the brain response associated with the effect of semantic congruency between expected and semantically incongruous words in the audiovisual modality of presentation compared to the auditory-only modality. Moreover, if audiovisual speech affects lexical-semantic processing, a facilitated recovery of episodic memory representations associated with final words would be shown in a lexical recognition task, such that the facilitated recovery due to audiovisual speech could depend on the semantic congruency of final words regarding the sentence context. An alternative hypothesis is that audiovisual speech contributes to the recognition of phonemes rather than the linguistic encoding of the utterance at higher levels. In that case, the brain response associated with the effect of semantic congruency should not differ as a function of the modality of presentation.

The N100 associated with perceptual processing was affected by the modality of presentation, so that a greater amplitude was seen over frontocentral sites in audiovisual trials than in auditory-only trials. Regarding the semantic congruency of word in the sentence context, the amplitude of the N100, N200 and N400 waves was greater after semantically

incongruous than contextually expected words. In line with the goal of the present study, a greater efficiency to combine the analysis of the incoming word with the contextually based constraints from the meaning of the utterance was observed over the N200 component when words were presented in the audiovisual modality than in the auditory-modality alone. While an increased N200 amplitude for incongruous words in the audiovisual modality was found over centroparietal sites, a reduced N200 amplitude for expected words in the audiovisual modality was observed over occipito-parietal sites. In contrast, the semantic congruency effect did not vary as a function of the modality of presentation over the N400 and audiovisually presented words elicited a reduced N400 wave amplitude over parietal sites. In the lexical recognition task, words were better recalled when they had been previously presented in audiovisual trials than in an auditory-only modality. As for the N400, this facilitated recovery of episodic memory representations associated with final words was found independently of the semantic congruency. The implications of these findings are discussed below.

Impact of audiovisual speech on sentence processing at perceptual level

Although the N100 was affected by the modality of presentation, our findings do not replicate previous electrophysiological studies showing the suppression of N100 amplitude when the visual articulatory information was in accordance with the auditory information compared to an auditory-only stimulation (Besle et al., 2004; Brunellière, et al., 2013; Klucharev et al., 2003; Pilling, 2009, van Wassenhove et al., 2005). In the present study, we used grammatical cataphora in French spoken sentences to create strong predictions of the critical word. Grammatical cataphora is the use of a pronoun to refer ahead to another word in a sentence. For example, in the sentence context “J’ai eu du mal à me garer, il est complet”, the pronoun “il” appears earlier than “le parking” to which it refers (“I found it difficult to park my car, the parking lot is full”). The use of this structure gives strong predictions about

when the critical word will be uttered, as a pause occurs after the sentence context (e.g., after the following context “J’ai eu du mal à me garer, il est complet”). Therefore, the mechanisms underlying temporal attention may be highly involved in our study. Interestingly, in electrophysiological studies exploring the effect of temporal attention, an increased amplitude of the auditory N100 was found when the stimulus was temporally expected (Lange, 2012; Lange, & Schnuerch, 2014). We reported the same effect in the audiovisual modality as compared to the auditory-only modality. The increase in N100 amplitude with the audiovisual modality compared to the auditory-only modality might reflect the facilitation of audiovisual speech on temporal attention to speech sounds. This interpretation is in line with recent behavioral findings showing that participants benefit more from the audiovisual modality than from unimodal stimulation when adapting their temporal attention in non-speech events (Ball, Michels, Thiele, & Noesselt, 2018).

Moreover, at a perceptual level, it appears that the contribution of audiovisual speech operates independently of sentence-level processing. Like previous ERP studies focusing on the influence of semantic constraints during written word recognition (e.g., Penolazzi, Hauk, & Pulvermüller, 2007; Kim, & Lai, 2012) and auditory word recognition (Brunellière and Soto-Faraco, 2015), we found that the semantic constraints of sentence context affected the recognition of spoken words in the perceptual stages around 100 ms. The earliness of these effects even before 100 ms could be explained by the structure of the sentences used in the present experiment. The early impact of semantic constraints is in accordance with psycholinguistic models of spoken word recognition (e.g., TRACE model, McClelland & Elman, 1986) that posit top-down feedback between the lexical and lower levels (phonemes and acoustic features), and with a predictive coding framework (Friston and Kiebel, 2009) assuming that the brain continuously infers the probabilities of sensory input across the hierarchy of multi-level representations to be able to predict upcoming input. The distinctive

contribution of audiovisual speech with respect to sentence-level processing over the N100 can be accounted for by the nature of audiovisual benefits at the perceptual level. It has indeed been suggested that modulations in N100 due to audiovisual speech stem from the integration of spatial and temporal audiovisual properties, because the speeding-up of the N100 occurred both in speech and non-speech events (Baart, Stekelenburg, & Vroomen, 2016).

Impact of audiovisual speech on sentence processing at sub-lexical, lexical, and sentence-levels

The N100 was followed by the two other components known to be sensitive to sentence-level processes when a word is semantically incongruous regarding the sentence context. The latency of the N200 component was similar to that observed in previous ERP studies (Basirat, Brunellière, & Hartsuiker, 2018; Brunellière, et al., 2013) examining the impact of audiovisual presentation in sentence or word processing. Contrary to the present study, the aforementioned ERP studies investigated the benefits of audiovisual speech by focusing on either the effect of word repetition or the saliency of visual articulatory onset. Above all, our study is the first to show that audiovisual speech can influence the stages of spoken-word recognition by interacting with the sentence-level processes in constructing the meaning of sentences. This was evidenced by the finding that audiovisual speech increased the effect of semantic congruency over the N200 component. This increased effect of semantic congruency by audiovisual speech was due to a reduction in the N200 amplitude when the incoming word was expected by the semantic constraints of sentence context and an increase in N200 amplitude when the incoming word was semantically incongruous over particular sites. Therefore, it appeared clearly that audiovisual speech facilitated the processing of the incoming word by interacting with the sentence-level processes.

Although this finding seems to be in accordance with behavioral findings during the processing of isolated words (Buchwald et al., 2009; Fort et al., 2013; Dodd, Oerlemans, and Robinson, 1989; Kim et al., 2004), it should be more precisely discussed in the light of the prior literature on audiovisual speech and word recognition in sentence contexts. Studies on audiovisual speech with isolated syllables usually reveal a positivity, called P200, occurring at the same moment as the N200 between 100 ms and 200 ms. Several authors have found that the P200 is more associated with the phonological content of speech sounds than the N100, since its amplitude reacted to incongruency between auditory and visual information (Stekenburg & Vroomen, 2007; Klucharev et al., 2003) and its speeding-up occurred only in speech events (Baart, Stekelenburg, & Vroomen, 2016). The function of the P200[4] may be understood as being somewhat similar to that of the N200 in word recognition in sentence contexts, since it is thought to reflect the phonological processing that occurs during the recognition of the initial phonemes in a word (Connolly, Phillips, Stewart, & Brake, 1992). In the context of auditory sentence processing, phonological processing however interacts with the lexical- and sentence-level processes over the N200 component (Connolly, Phillips, Stewart, & Brake, 1992). A possible explanation for the increased effect of semantic congruency in audiovisual speech is that the latter acts on lexical activation thanks to phonological processing at the sub-lexical level. Given the phonological and timing predictions about the occurrence of upcoming speech sounds (for a review, Peelle & Sommers, 2015), a speaker's orofacial movements may pre-activate a set of lexical candidates, that share similar visual articulatory cues (Fort et al., 2013). Moreover, the

[4] The P200 is apparent during the listening of natural speech when auditory evoked spread spectrum (AESPA) analysis is used (Power, Foxe, Forde, Reilly, & Lalor, 2012). The AESPA method is sensitive to electrophysiological brain response related to the amplitude envelope of a natural continuous speech. Contrary to this approach, we tracked to ERPs responses after expected and incongruous words in the context of semantically constraining sentences. In such experimental designs (e.g., Connolly and Phillips, 1994; Connolly, Stewart, & Phillips, 1990 ; van den Brink and Hagoort, 2004, van den Brink, Brown and Hagoort, 2001), a N200 is elicited when the initial phonemes of the perceived word do not match the initial phonemes of the expected word from the sentence constraints.

complementarity of visual and auditory speech may reduce the number of pre-activated lexical candidates (Jesse & Massaro, 2010; Mattys, Bernstein, & Auer, 2002; Tye-Murray, Sommers, & Spehar, 2007). For instance, the target word "*pin*" will be recognized more easily in the audiovisual modality because visual speech can rule out lexical competitors such as "*fin*" or "*kin*". By acting at lexical level, the audiovisual speech may help to detect the semantic incompatibility of incongruous words in the sentence context and the semantic compatibility of expected words. Therefore, integrative processes at sentence level could explain the effects of semantic congruency by integrating the word, which was better recognized in audiovisual speech, into the representation of the sentence.

An alternative explanation is that listeners use the semantic constraints arising from sentence contexts to generate on-line lexical predictions about upcoming words and lexical predictions interact with phonological processing and lexical activation by the incoming word. There is a growing body of experimental evidence obtained from electrophysiological recordings and visual world eye-tracking paradigms for lexical predictions about the sentence context in spoken-sentence comprehension (Altmann and Kamide, 1999; Brunellière & Soto-Faraco, 2013, 2015; Foucart, Ruiz-Tada, Costa, 2015; Kaiser and Trueswell, 2004; Kamide, Altmann, & Haywood, 2003; van Berkum, van den Brink, Tesink, Kos, & Hagoort, 2005; Otten, Nieuwland, & Van Berkum, 2007; Wicha, Bates, Moreno, & Kutas, 2003). The early impact of semantic constraints in the present study is in accordance with on-line lexical predictions about upcoming words from the semantic constraints of sentence contexts. The two explanations that we have described however could arise at the same period of time across neural loops. Although we cannot provide more explanations between the strength of top-down and bottom-up processing involved at the different levels (sub-lexical, lexical, sentence level) in natural spoken sentences, this study shows the clear contribution of audiovisual speech to the interplay between spoken-word

recognition and processing at sentence level to access the lexical-semantic network by increasing semantic congruency effects.

Contrary to the findings on the N200, the amplitude of N400 elicited in the audiovisual modality was lower than in the auditory-only modality, independently of the semantic plausibility. This phenomenon was not observed over the entire scalp but was limited to certain topographical sites, including the left and right parietal sites and the occipito-parietal sites. During the N400 wave, audiovisual speech seemed to act on the recognition of phonemes without interacting with the sentence-level processes. Given the different pattern of findings between N200 and N400, our findings suggest that audiovisual speech mainly functions at the early stages of word recognition to access a lexical-semantic network. The distinctive pattern of N200 and N400 could correspond to the two main stages of spoken word recognition usually described by models of spoken-word recognition (Gaskell & Marslen-Wilson, 1997; Marslen-Wilson, 1984; McClelland & Elman, 1986). Although not all models of spoken-word recognition claim that sentence context plays a role during phonological processing in the initial stages of word recognition, they all assume that once the first lexical candidates have been activated from word onset, the following speech sounds of the incoming final word are delivered until a unique lexical candidate is selected. The time window of the N200 could correspond to the initial stages of word recognition described by these models, whereas the N400 could be associated with the later stages of word recognition when the lexical candidate is selected (for similar interpretations, van den Brink and Hagoort, 2004, van den Brink, Brown and Hagoort, 2001). Unlike the initial stages of word recognition, audiovisual speech did not affect the interplay between spoken-word recognition and processing at the sentence level during the later stages but guided the recognition of the incoming word as reflected by the N400 wave. When sufficient phonological information about the incoming word has been recognized and the word is selected, audiovisual speech

operates independently of the semantic constraints given by the sentence context. Further studies in magnetoencephalography and with electrical intracranial recordings should be conducted to clarify whether modality effects in spoken language comprehension are related to interactions between auditory and visual information or can be due to simply the superposition of the two modalities with respect to an auditory unimodal situation.

Impact of audiovisual speech on lexical memory

The interest of the present study was to explore how listeners process the incoming speech to extract the meaning of an utterance in different modalities (audiovisual modality vs. auditory alone). It was therefore relevant to examine episodic memory representations in the lexicon to better understand the impact of audiovisual speech in lexical-semantic processing. We provide experimental evidence for richer episodic memory representations in the audiovisual modality. Old words were better recognized in the audiovisual modality than those presented in the auditory-only modality. Such an increase in memory capacity has been reported in free-recall tasks with audiovisual material compared to unimodal material (Goolkasian & Foos, 2005; Thompson & Paivio, 1994; Pichora-Fuller, 1996). In those studies, improved performance in memory was found to be due to the combination of multisensory information and not simply due to the redundancy of the sensory information. In the present study, we did not explicitly ask our participants to memorize a set of words nor did we present them auditory noise-degraded stimuli. Instead, we tested the recall of words after listening to natural spoken sentences. The findings from the lexical recognition task suggest that audiovisual speech clearly influences spoken-word recognition in natural spoken sentences and therefore reinforces the memory trace of words in the lexicon. Importantly, improved performance in memory with audiovisual material compared to auditory-only material was found in the present study during which the recovery of old words was evaluated

from their written presentation. Hence, the findings could be not simply associated with familiarity effects based on the acoustic or phonetic properties when the words had previously been heard or a better encoding of phonetic or phonological forms thanks to the audiovisual presentation.

Additionally, richer episodic memory representations may occur with old words that fit with the contextually based constraints of the content of the utterance than with old words that are semantically incongruous. In particular, we replicate the findings of Neville, Kutas, Chesney, & Schmidt (1986) by showing that old words that previously occurred in sentences during a listening task were better recognized when they fitted with the sentence context than those that did not. Moreover, the contextually based constraints arising from the meaning of the utterance led the listeners to consider more as heard the words they expected from the sentence context than the new words although the former had not been presented in the speech input (for similar findings, Foucart, Ruiz-Tada, & Costa, 2015). This could be accounted for by the notion of prediction such that listeners generating on-line lexical predictions for upcoming words from the sentence contexts reinforce the memory trace of predicted words. Interestingly, no additive effects were observed between the presentation modality and the semantic congruency. The findings obtained during the later stages of spoken-word recognition were in accordance with those regarding episodic memory representations. We could hypothesize that the contribution of audiovisual speech during the later stages of spoken-word recognition could have led to the absence of additive effects between the modality of presentation and the semantic congruency on the memory trace of words.

To conclude, audiovisual speech can contribute to the interplay between spoken-word recognition and sentence-level processing. Nonetheless, its contribution to sentence-level processing is limited to the initial stages of spoken-word recognition when a set of lexical

candidates is activated. When sufficient phonological information about the incoming word has been recognized and the word is selected, audiovisual speech operates independently of the semantic constraints of sentence context.

Acknowledgements: This research was supported by visual studies grant (SCV2013-2014) from the French National Research Agency (ANR-11-EQPX-0023), and European funds through the FEDER SCV-IrDIVE program. It was also funded by the University of Lille (AAPEtablissement2014) and the municipal authorities in Lille (AppelLMCU2014). We are very grateful to Adèle Delalleau, Benjamin Lob, Amandine Lepachelet, Laurent Ott and Maeva Veber for their help in the selecting the stimuli and the running of the experiment. We also thank Perrine Janssoone for recording the stimuli. ERP analyses were performed with the Cartool software (supported by the Center for Biomedical Imaging in Geneva and Lausanne). The manuscript was proofread by a native-speaking English copy-editor. We thank the anonymous reviewers for their helpful comments.

References

Alsius, A., Möttönen, R., Sams, M. E., Soto-Faraco, S., & Tiippana, K. (2014). Effect of attentional load on audiovisual speech perception: evidence from ERPs. *Fontiers in Psychology, 5*, e727. doi: 10.3389/fpsyg.2014.00727

Altmann, G.T., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, *73*, 247-264. doi: 10.1016/S0010-0277(99)00059-1

Arnal, L.H., Morillon, B., Kell, C.A., & Giraud, A.-L. (2009). Dual neural routing of visual facilitation in speech processing. *The Journal of Neuroscience, 29,* 13445-13453. doi: 10.1523/JNEUROSCI.3194-09.2009

Baart, M. (2016). Quantifying lip-read-induced suppression and facilitation of the auditory N1 and P2 reveals peak enhancements and delays. *Psychophysiology,* 53, 1295-1306. doi: 10.1111/psyp.12683

Baart, M., & Samuel, A. G. (2015). Turning a blind eye to the lexicon: ERPs show no cross-talk between lip-read and lexical context during speech sound processing. *Journal of Memory and Language, 85,* 42-59. doi: 10.1016/j.jml.2015.06.008

Baart, M., Stekelenburg, J. J., & Vroomen, J. (2016). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia, 53,* 115-121. doi: 10.1016/j.neuropsychologia.2013.11.011

Badin, P., Tarabalka, Y., Elisei, F., & Bailly, G. (2010). Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding. *Speech Communication, 52,* 493-503. doi: 10.1016/j.specom.2010.03.002

Ball, F., Michels, L.E., Thiele, C., & Noesselt, T. (2018). The role of multisensory interplay in enabling temporal expectations. *Cognition, 170,* 130-146. doi: 10.1016/j.cognition.2017.09.015

Basirat, A., Brunellière, A., & Hartsuiker, R. (2018). The role of audiovisual speech in the early stages of lexical processing as revealed by ERP word repetition effect. *Language Learning, 68,* 80-101. doi: 10.1111/lang.12265

Beauchamp, M.S., Argall, B.D., Bodurka, J., Duyn, J.H., & Martin, A. (2004). Unraveling multisensory integration: patchy organization within human STS multisensory cortex. *Nature Neuroscience, 7,* 1190-1192. doi: 10.1038/nn1333

Benoit, C., Guiard-Marigny, T., Le Goff, B., & Adjoudani, A. (1996). Which components of the face do humans and machines best speechread? In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by Humans and Machines*: Berlin: NATO-ASI Series 150 Springer.

Benoit, C., Lallouache, T., Mohamadi, T., & C. Abry. (1994). A set of French visemes for visual speech synthesis", Les cahiers de l'ICP, research report, 3, 113-129.

Besle, J., Fort, A., Delpuech, C., & Giard, M.H. (2004). Bimodal speech: Early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience, 20,* 2225-2234. doi: 10.1111/j.1460-9568.2004.03670.x

Besle, J., Fort, A., & Giard, M.H., (2004). Interest and validity of the additive model in electrophysiological studies of multisensory interactions. *Cognitive Processing, 5,* 189-192. doi: 10.1007/s10339-004-0026-y

Boersma, P. & Weenink, D. (2011). Praat: doing phonetics by computer [Computer program]. Version 3.4, retrieved 2 Jan 2011 from http://www.praat.org/

Brunellière, A. Sánchez-García, C., Ikumi, N., & Soto-Faraco, S. (2013). Visual information constrains early and late stages of spoken-word recognition in sentence context. *International Journal of Psychophysiology, 89*, 136-147. doi: 10.1016/j.ijpsycho.2013.06.016

Brunellière, A., & Soto-Faraco, S. (2013). The speakers' accent shapes the listeners' phonological predictions during speech perception. *Brain and Language*, *125*, 82-93. doi: 10.1016/j.bandl.2013.01.007

Brunellière, A., & Soto-Faraco, S. (2015). The interplay between semantic and phonological constraints during spoken-word comprehension. *Psychophysiology, 52*, 46-58. doi: 10.1111/psyp.12285

Buchwald, A.B., Winters, S.J., & Pisoni, D.B. (2009). Visual speech primes open-set recognition of spoken words. *Language and Cognitive Processes*, *24*, 580-610. doi: 10.1080/01690960802536357

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, 10*, 649-657. doi: 10.1016/S0960-9822(00)00513-3

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A.A. (2009). The natural statistics of audiovisual speech. *PLoS Computing Biology, 5*, e1000436. doi: 10.1371/journal.pcbi.1000436

Connolly, J.F., & Phillips, N.A. (1994). Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. *Journal of Cognitive Neuroscience, 6*, 256-266. doi: 10.1162/jocn.1994.6.3.256

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 and Language, 43*, 1-18. doi: 10.1016/0093-934X(92)90018-A

Connolly, J.F., Stewart, S.H., & Phillips, N.A. (1990). The effects of processing requirements on neurophysiological responses to spoken sentences. *Brain and Language, 39,* 302-318. doi: 10.1016/0093-934X(90)90016-A

Dodd, B., Oerlemens, M., & Robinson, R. (1989). Cross-modal effects in repetition priming: A comparison of lip-read graphic and heard stimuli. *Visible Language, 22,* 59-77.

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research, 11,* 796-804. doi: 10.1044/jshr.1104.796

Foucart. A, Ruiz-Tada, E., Costa, A. (2015). How do you know I was about to say "book"? Anticipation processes affect speech processing and lexical recognition. *Language, Cognition and Neuroscience, 30,* 768-780. doi: 10.1080/23273798.2015.1016047

Fort, M., Kandel, S., Chipot, J., Savariaux, C., Granjon, L., & Spinelli, E. (2013). Seeing the initial articulatory gestures of a word triggers lexical access. *Language and Cognitive Processes, 28,* 1207-1223. doi: 10.1080/01690965.2012.701758

Fridriksson, J., Moss, J., Davis, B., Baylis, G.C., Bonilha, L., & Rorden, C. (2008). Motor speech perception modulates the cortical language areas. *NeuroImage, 41,* 605-613. doi: 10.1016/j.neuroimage.2008.02.046

Friston, K., & Kiebel, S. (2009). Cortical circuits for perceptual inference. *Neural Networks, 22,* 1093-1104. doi: 10.1016/j.neunet.2009.07.023

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12,* 613-656. doi: 10.1080/016909697386646

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

Grant, K.W., & Seitz, P.F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *Journal of the Acoustical Society of America, 104,* 2438-2450. doi: 10.1121/1.423751

Greenhouse, S.W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24,* 95-111. doi: 10.1007/BF02289823

Hagoort, P., & Brown, C.M. (2000). ERP effects of listening to speech: semantic ERP effects. *Neuropsychologia, 38,* 1518-1530. doi: 10.1016/S0028-3932(00)00052-X

Hall, D.A., Fussell, C., & Summerfield, A.Q. (2005). Reading fluent speech from talking faces: Typical brain networks and individual differences. *Journal of Cognitive Neuroscience, 17,* 939-953. doi: 10.1162/0898929054021175

Hickock, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8,* 393-402. doi: 10.1016/j.jcomdis.2012.06.004

Istria, M., Nicolas-Jeantoux, C., & Tamboise, J. (1982). *Manuel de lecture labiale. Exercices d'entraînement*. Paris: Masson.

Jesse, A., & Massaro, D.W. (2010). The temporal distribution of information in audiovisual spoken-word identification. *Attention, Perception, Psychophysics, 72,* 209-225. doi: 10.3758/APP.72.1.209

Kaiser, E., & Trueswell, J.C. (2004). The role of discourse context in the processing of a flexible word-order language. *Cognition, 94,* 113-147. doi: 10.1016/j.cognition.2004.01.002

Kamide, Y., Altmann, G.T., & Haywood, S.L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49,* 133-156. doi: 10.1016/S0749-596X(03)00023-8

Kim, J., Davis, C., & Krins, P. (2004). A modal processing of visual speech as revealed by priming. *Cognition, 93,* B39-B47. doi: 10.1016/j.cognition.2003.11.003

Kim, A., & Lai, V. (2012). Rapid interactions between lexical semantic and word form analysis during word recognition in context: evidence from ERPs. *Journal of Cognitive Neuroscience, 24*, 1104-1112. doi: 10.1162/jocn_a_00148

Klucharev, V., Möttönen, R., & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Brain Cognitive Research, 18*, 65-75. doi: 10.1016/j.cogbrainres.2003.09.004

Kutas, M., & Federmeier, K.D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *12*, 463-470. doi: 10.1016/S1364-6613(00)01560-6

Lange, K. (2012). The N1 effect of temporal attention is independent of sound location and intensity: Implications for possible mechanisms of temporal attention. *Psychophysiology*, 49, 1468-1480. doi: 10.1111/j.1469-8986.2012.01460.x

Lange, K. & Schnuerch, R. (2014). Challenging perceptual tasks require more attention: The influence of task difficulty on the N1 effect of temporal orienting. *Brain and Cognition*, *84*, 153-163. doi: 10.1016/j.bandc.2013.12.001

Lau, E.F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience*, *9*, 920-933. doi: 10.1038/nrn2532

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. doi: 10.1371/journal.pone.0004638

MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology, 21*, 131-141. doi: 10.3109/03005368709077786

Marslen-Wilson, W. D. (1984). Function and process in spoken word recognition: A tutorial review. In H. Bouma & D. G. Bouwhis (Eds.), *Attention and performance X: Control of language processes* (pp. 125–150). Hillsdale, NJ: Erlbaum.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

Mattys, S.L., Bernstein, L.E., & Auer, E.T. (2002). Stimulus-based lexical distinctiveness as a general word recognition mechanism. *Perception and Psychophysics*, *64*, 667-679. doi: 10.3758/BF03194734

McClelland, J. L., & Elman, J. L. (1986). The Trace model of speech perception. *Cognitive Psychology, 18,* 1-86. doi: 10.1016/0010-0285(86)90015-0

McGrath. (1985). An examination of cues for visual and audiovisual speech perception using natural and computer generated faces. PhD Dissertation. University of Nottingham, United Kingdom.

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. *Psychological Research, 71*, 4-12. doi: 10.1007/s00426-005-0031-5

Neville, H. J., Kutas, M., Chesney, G., & Schmidt, A. L. (1986). Event-related brain potentials during initial encoding and recognition memory of congruous and incongruous words. *Journal of Memory and Language, 25*, 75-92. doi : 10.1016/0749-596X(86)90022-7

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE. *L'Année Psychologique*, *101*, 447-462. doi: 10.3406/psy.2001.1341

Oldfield, R.C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9,* 97-113. doi: 10.1016/0028-3932(71)90067-4

Otten, M., Nieuwland, M.S., & van Berkum, J.A. (2007). Great expectations: specific lexical anticipation influences the processing of spoken language. *BMC Neuroscience*, *8*, 89. doi: 10.1186/1471-2202-8-89

Peelle, J.E, & Sommers, M.S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169-181. doi: 10.1016/j.cortex.2015.03.006

Penolazzi, B., Hauk, O., & Pulvermüller, F. (2007). Early semantic context integration and lexical access as revealed by event-related brain potentials. *Biological Psychology*, *74*, 374-388. doi: 10.1016/j.biopsycho.2006.09.008

Perrin, F., Pernier, J., Bertrand, O., Giard, M.-H., & Echallier, J.F. (1987). Mapping of scalp potentials by surface spline interpolation. *Electroencephalography and Clinical Neurophysiology*, *66*, 75-81. doi: 10.1016/0013-4694(87)90141-6

Pichora-Fuller, M.K. (1996). Working memory and speechreading. In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by humans and machines: Models, systems, and applications* (pp.257-274). Berlin: Springer.

Pilling, M. (2009). Auditory event-related potentials (ERPs) in audiovisual speech perception. *Journal of Speech, Language and Hearing Research*, *52*, 1073-1081. doi: 10.1044/1092-4388(2009/07-0276

Power, A. J., Foxe, J. J., Forde, E.-J., Reilly, R. B., & Lalor, E. C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. *European Journal of Neuroscience*, *35*, 1497-1503. doi: 10.1111/j.1460-9568.2012.08060.x

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. *Cerebral Cortex*, *17*, 1147-1153. doi: 10.1093/cercor/bhl024

Samuel, A.G., & Lieblich, J. (2014). Visual speech acts differently than lexical context in supporting speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 40,* 1479-1490. doi: 10.1037/a0036656

Schwartz, J.-L., Savariaux, C. (2014). No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. *PLoS Computational Biology, 10,* e1003743. doi: 10.1371/journal.pcbi.1003743

Sekiyama, K., Kanno, I., Miura, S., & Sugita, Y. (2003). Auditory-visual speech perception examined by fMRI and PET. *Neuroscience Research, 47,* 277-287. doi: 10.1016/S0168-0102(03)00214-1

Skipper, J.I., Nusbaum, H., & Small, S.L. (2005). Listening to talking faces: Motor cortical activation during speech perception. *NeuroImage, 25,* 76-89. doi: 10.1016/j.neuroimage.2004.11.006

Stekelenburg, J.J., & Vroomen, J., (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience, 19,* 1964-1973. doi: 10.1162/jocn.2007.19.12.1964

Sumby, W.H., & Pollack, I. (1954) Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26,* 212-215. doi: 10.1121/1.1907309

Summerfield, Q., & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology, 36,* 51-74. doi: 10.1080/14640748408401503

Teder-Salejarvi, W.A., McDonald, J.J., Di Russo, F. & Hillyard, S.A. (2002). An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings. *Cognitive Brain Research, 14,* 106-114. doi :10.1016/S0926-6410(02)00065-4

Thompson, V., & Paivio, A. (1994). Memory for pictures and sounds: independence of auditory and visual codes. *Canadian Journal of Experimental Psychology, 48*, 380-398. doi: 10.1037/1196-1961.48.3.380

Tye-Murray, N., Sommers, M.S., & Spehar, B. (2007). Auditory and visual lexical neighborhoods in audiovisual speech perception. *Trends in Amplification, 11*, 233-241. doi: 10.1177/1084713807307409

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. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 443-467. doi; 10.1037/0278-7393.31.3.443

van den Brink, D., Brown, C.M., & Hagoort, P. (2001). Electrophysiological evidence for early contextual influences during spoken-word recognition: N200 versus N400 effects. *Journal of Cognitive Neuroscience, 13*, 967-985. doi: 10.1162/089892901753165872

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. *Journal of Cognitive Neuroscience, 16*, 1068-1084. doi: 10.1162/0898929041502670

Van Engen, K.J., Xie, Z., & Chandrasekaran, B. (2017). Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect. *Attention, Perception & Psychophysics*, *79*, 396-403. doi: 10.3758/s13414-016-1238-9

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, 102*, 1181-1186. doi: 10.1073/pnas.0408949102

Wicha, N.Y.Y., Bates, E.A, Moreno, E.M., & Kutas, M. (2003). Potato not Pope: human brain potentials to gender expectation and agreement in Spanish spoken sentences. *Neuroscience Letters*, *346*, 165-168. doi: 10.1016/S0304-3940(03)00599-8

Wright, T.M., Pelphrey, K.A., Allison, T., McKeown, M.J., & McCarthy, G. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cerebral Cortex*. *13*, 1034-1043. doi: 10.1093/cercor/13.10.1034

Figure captions

Figure 1. Mean Hit rates in % for old words per experimental condition (AO-Expected, AO-Incongruous, AV-Expected, AV-Incongruous) and false alarms in % for unheard words (Expected-Auditory Context, Expected-Audiovisual Context, New) and SEM bars. *p<.05; **p<.01; ***p<.001; AO: Auditory-only modality, AV: Audiovisual modality.

Figure 2. Grand-average ERP waveforms time-locked to the auditory onset of final word across four experimental conditions. AO: Auditory-only modality, AV: Audiovisual modality.

Figure 3. A. ERP waveforms over the frontocentral and centroparietal sites across four experimental conditions (AO-Expected, AO-Incongruous, AV-Expected, AV-Incongruous). AO: Auditory-only modality, AV: Audiovisual modality. B. Topographical maps of modality and semantic congruency effects over the N100 time window.

Figure 4. A. Mean and SEM bars of ERP amplitude in the N200 time windows across four experimental conditions (AO-Expected, AO-Incongruous, AV-Expected, AV-Incongruous). AO: Auditory-only modality, AV: Audiovisual modality. *p<.05; **p<.01; ***p<.001. B. Topographical maps of semantic congruency effect on the N200 according to the modality of presentation. C. Topographical maps of modality effect on the N200 according to the semantic congruency.

Figure 5. A. ERP waveforms over the critical sites (left and right parietal and occipito-parietal sites) of the modality effect on the N400. B. Topographical map of modality effect on the N400

Figure 6. Topographical maps averaged across all experimental conditions after gfp normalization over the N200 and the N400 time window.

Tables

Table 1. Examples of experimental conditions during listening task

Conditions	Modality presentation	Semantic congruency	Examples
AO-Expected	Auditory-only	Expected	J'ai eu du mal à me garer, il est complet le **parking.** *I found it difficult to park my car, the **parking lot** is full.*
AO-Incongruous	Auditory-only	Incongruous	*J'ai eu du mal à me garer, il est complet le **coussin.** **I found it difficult to park my car, the **cushion** is full.*
AV-Expected	Audiovisual	Expected	J'ai eu du mal à me garer, il est complet le **parking.** *I found it difficult to park my car, the **parking lot** is full.*
AV- Incongruous	Audiovisual	Incongruous	*J'ai eu du mal à me garer, il est complet le **coussin.** **I found it difficult to park my car, the **cushion** is full.*

Asterisk indicates when final noun (here, in bold) was semantically incongruent with respect to semantic constraints of sentence context. AO: Auditory-only modality, AV: Audiovisual modality

Table 2. Mean values of psycholinguistics properties associated with final words

Conditions	Lexical Frequency	Number of phonemes	Number of syllables	Number of phonological neighbors	Uniqueness point
Expected	7.1	5.7	2.3	3.9	5
Incongruous	6.6	5.8	2.3	3.7	5.1

The psycholinguistic properties of final words were extracted from the Lexique database (New, Pallier, Ferrand &, Matos, 2001, http://www.lexique.org), lexical frequency in number of occurrences per million words across corpora of film dialogue. A phonological neighbor is any word that can be created by changing one phoneme without changing the other.

Table 3. Statistical results of ERP data across various time windows (N100, N200, N400)

	N100 time window	**N200 time window**	**N400 time window**
Semantic Congruency	$F(1,31)=8.17$, $p=.007$	$F(1,31)=31.35$, $p=4\times10\text{-}6$	$F(1,31)=101.51$, $p=2.6\times10\text{-}11$
Modality	$F(1,31)=1.26$, $p=.27$	$F(1,31)= 0.007$, $p=.93$	$F(1,31)=0.3$, $p=.59$
Topography	$F(6,186)= 8.81$, $p=9.3\times10\text{-}5$	$F(6,186)=8.58$, $p=9\times10\text{-}6$	$F(6,186)=45.08$, $p=2.1\times10\text{-}14$
Semantic Congruency x Modality	$F(1,31)=1.83$, $p=.19$	$F(1,31)=5.97$, $p=.02$	$F(1,31)=0.06$, $p=.8$
Semantic Congruency x Topography	$F(6,186)=1.5$, $p=.22$	$F(6,186)=4.19$, $p=.016$	$F(6,186)=14.42$, $p=2.2\times10\text{-}7$
Modality x Topography	$F(6,186)=2.85$, $p=.049$	$F(6,186)=2.36$, $p=.09$	$F(6,186)=8.88$, $p=.0001$
Semantic Congruency x Modality x Topography	$F(6,186)=0.67$, $p=.52$	$F(6,186)=0.7$, $p=.48$	$F(6,186)=1.47$, $p=.23$

Figure 1

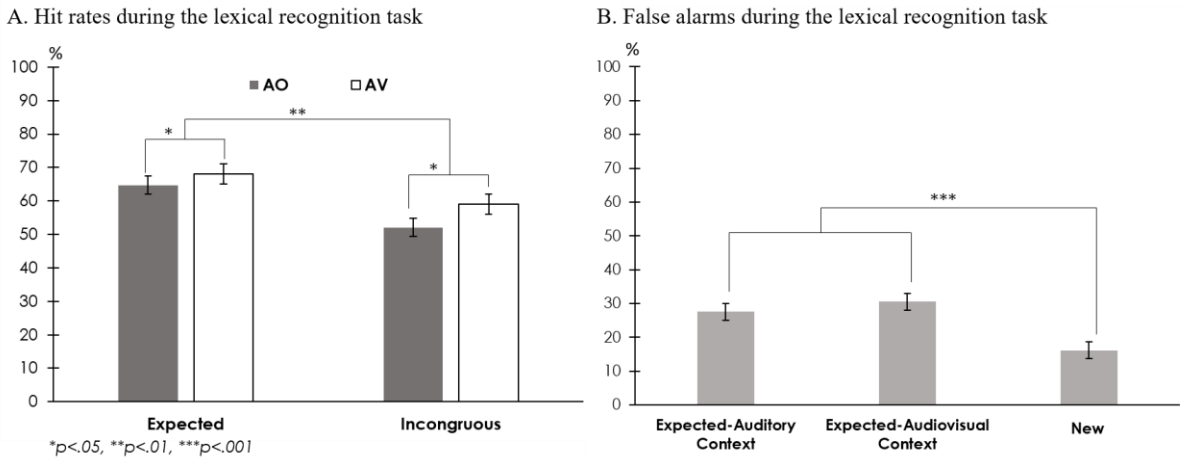

Figure 2

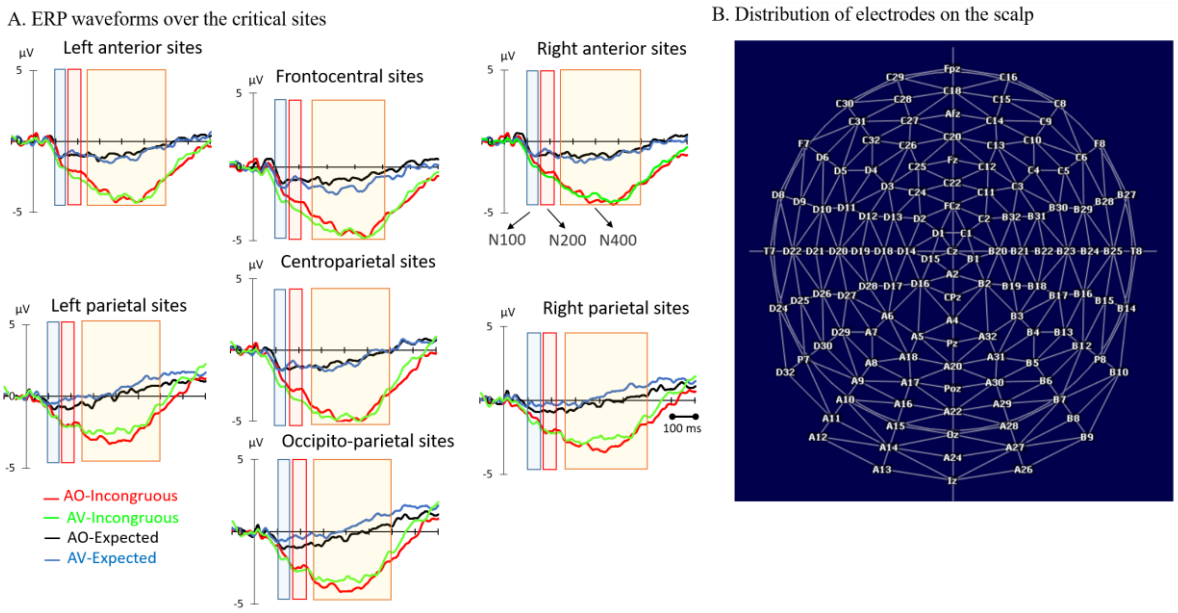

Figure 3

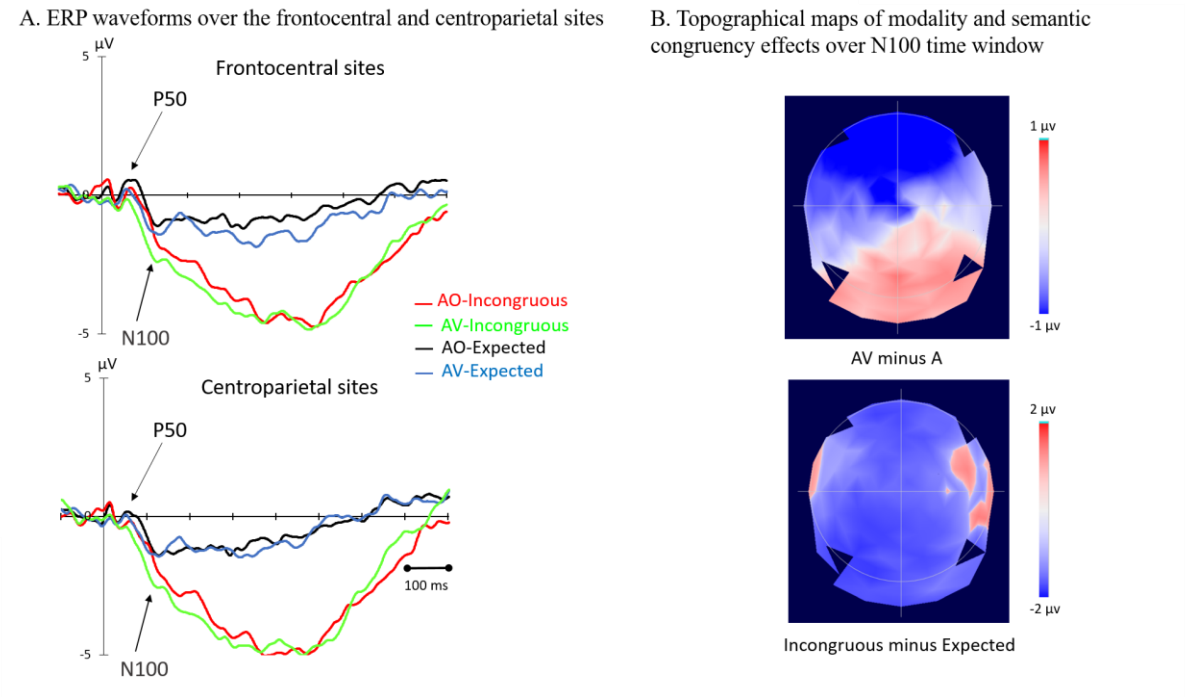

Figure 4

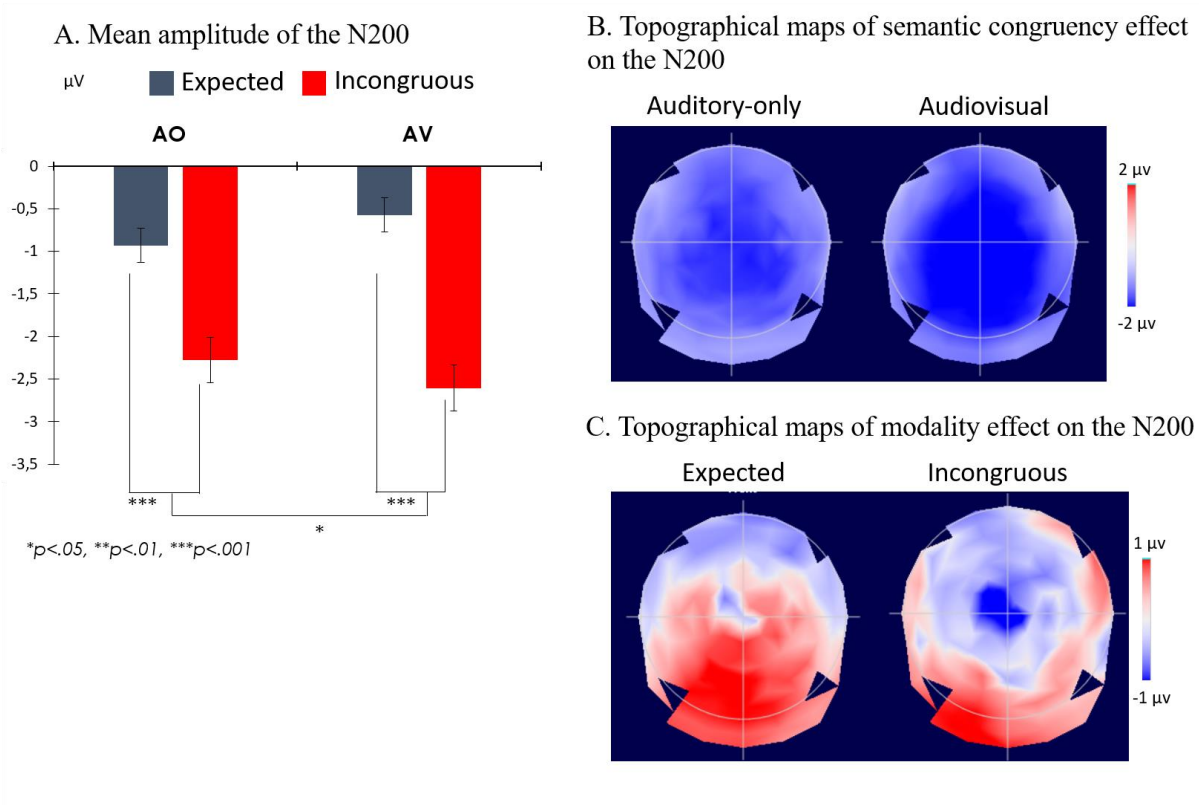

Figure 5

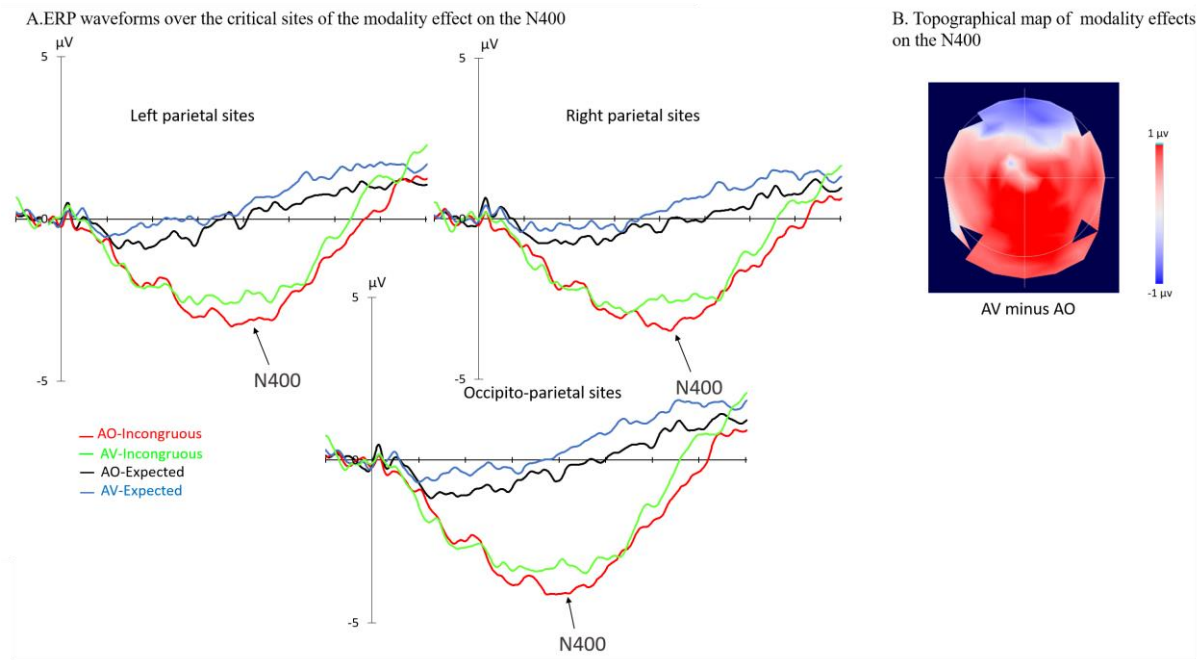

Figure 6