How are visemes and graphemes integrated with speech sounds during spoken word recognition? ERP evidence for supra-additive responses during audiovisual compared to auditory speech processing

Chotiga Pattamadilok, Marc Sato

To cite this version:

Chotiga Pattamadilok, Marc Sato. How are visemes and graphemes integrated with speech sounds during spoken word recognition? ERP evidence for supra-additive responses during audiovisual compared to auditory speech processing. Brain and Language, 2022, 225, 10.1016/j.bandl.2021.105058 . hal-03472191v2

HAL Id: hal-03472191
https://hal.science/hal-03472191v2

Submitted on 27 Dec 2021

Short title: Audiovisual integration during spoken word recognition

How are visemes and graphemes integrated with speech sounds during spoken word recognition? ERP evidence for supra-additive responses during audiovisual compared to auditory speech processing

Chotiga Pattamadilok[a] & Marc Sato[a]

[a] Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France

Corresponding author

Chotiga Pattamadilok

Laboratoire Parole et Langage

Centre National de la Recherche Scientifique (UMR 7309)

5, Av. Pasteur

13100 Aix-en-Provence

France

Email: chotiga.pattamadilok@lpl-aix.fr

Tel : +33 601323435

ABSTRACT

Articulatory gestures and orthography are connected with speech through a natural and an artificial association, respectively. This EEG study investigated whether the integrations between speech and these two visual inputs rely on the same mechanism, despite their different characteristics. A comparison of skilled readers' brain responses elicited by spoken words presented alone versus synchronously with visemes or graphemes showed that while neither visual input induced audiovisual integration on the N1 acoustic component, both led to a supra-additive integration on P2, with a stronger integration between speech and graphemes on left-anterior electrodes. This pattern persisted on the P350 component and generalized to all electrodes. The finding suggests a strong impact of reading acquisition on phonetic processing and lexical access. It also indirectly indicates that the dynamic and predictive cues present in natural lip movements but not in static visemes are critical to the contribution of visual articulatory gestures to speech processing.

Keywords: Audiovisual integration, supra-additivity, speech processing, articulatory gestures, orthography.

1. Introduction

Natural environment is filled with information from various sensory sources. The ability to detect their causal relationship and to integrate them into a coherent percept is fundamental to form a sense of space, improve perception, and is at the basis of learning. However, while certain associations that we encounter are natural and meaningful, others are artificial, and their acquisitions are far more effortful.

Language processing is an excellent illustration of both natural and artificial multisensory associations. On the one hand, speech sounds are tightly connected with visual information from time-varying kinematic of articulatory movements. This kind of audiovisual (AV) association is natural and develop spontaneously thanks to the biological link between action and perception (Meltzoff & Moore, 1977). Early in their development, infants become aware of the congruency between lip movements and speech sounds, both in terms of temporal synchrony and correspondence between sounds and shape of articulators (Bristow et al., 2009; Barbara Dodd, 1979). Although the full maturation of AV speech integration takes years (Lewkowicz & Flom, 2014; Sekiyama & Burnham, 2008), at around 4 months-old, illusory audiovisual fusion already emerges (i.e., McGurk effect; McGurk & MacDonald, 1976), which indicates infants' ability to integrate speech sounds and articulatory gestures into a unique percept (Bristow et al., 2009; Burnham & Dodd, 2004).

At a later developmental stage, children learn to associate speech sounds with new visual information, that is, orthography. Unlike the previous form of AV association, learning to associate speech sounds with abstract symbols is unnatural and requires extensive practice. Nevertheless, this new association becomes automatic in most adults. Several studies have shown that once reading is acquired, the speech processing system becomes sensitive to, if not dependent on, how speech sounds are orthographically represented (Dijkstra et al., 1995; Lafontaine et al., 2012; Muneaux &

Ziegler, 2004; Pattamadilok et al., 2009, 2014; Seidenberg & Tanenhaus, 1979; Taft, 2006; Ventura et al., 2004).

Even though both articulatory gestures and orthography are tightly connected with speech sounds, considering the nature of their associations, the question remains whether the integration between speech signal and these two kinds of visual inputs is supported by a similar binding process. While the integration between speech sounds and orthographic information has mainly been studied in laboratory settings under synchronous presentation of auditory and visual inputs, the integration between speech sounds and articulatory gestures (visual speech) is known to strongly rely on both spatial and temporal relationships between the auditory and visual signals, and on a high level of cross-predictability related to their common underlying motor cause. When considering the temporal AV relationships in visual speech, there is a robust correlation in time between variations of mouth opening and variations of the acoustic envelope (Chandrasekaran et al., 2009).

To our knowledge, very few studies attempted to compare these two forms of AV speech integration, and the available findings led to diverging conclusions. Using a syllable categorization task, Stekelenburg et al. (2018) examined whether articulatory gestures and written text induced illusory changes in the perception of ambiguous syllables (halfway between /aba/ and /ada/) to the same extent. The authors observed that while both visual cues induced a perceptual bias by shifting the interpretation of the ambiguous speech sound, the bias induced by lip movements was far more robust than that induced by a written syllable. Interestingly, the examination of the neural process underpinning the perceptual bias, using the McGurk-mismatch negativity (MMN) protocol, showed that the illusions induced by lip movements and written syllables were not supported by the same neural mechanism. While the integration between speech sounds and lip movements was associated with a negative deflection corresponding to the MMN typically reported in previous studies (Colin et al., 2002; Saint-Amour et al., 2007; Stekelenburg & Vroomen, 2012), it was not the case for written syllables. The integration between speech sounds and written syllables rather induced a late positive deflection with a frontal distribution that is indicative of a P3a, known to be involved in stimulus selection and decision-making processes. This observation suggests a delayed integration between written text and speech sounds but appears in contradiction with some previous findings that reported their early integration (Froyen et al., 2008; Mittag et al., 2011). Yet, as argued by the authors, the discrepancy could result from the experimental protocols, task demands, or speech materials that varied across studies.

Another recent study that jointly examined the impact of visual articulatory gestures and orthographic cues in AV integration was from Pinto and colleagues (2019). Regarding the impact of the articulatory gestures, which provided information on the timing, the phonetic content, and the articulatory features of speech input, the authors observed a classic reduction of both amplitude and latency of the N1/P2 components in comparison with the sum of the auditory and visual EEG signals (i.e., using and additive model: AV ≠ A+V; for a review, see Baart, 2016). The written syllable, displayed 600 ms prior to the acoustic onset, also significantly reduced the amplitude of the N1/P2 components compared to the condition where the spoken input was presented alone. For the authors, this amplitude reduction of early auditory evoked potentials suggests that the availability of the phonetic content affects an early sensory stage of auditory processing. However, since the written syllable always preceded the spoken input with a constant SOA, in addition to the information on phonetic content, the written syllable also provided a reliable temporal prediction which typically leads to a reduction of N1 amplitude (Stekelenburg & Vroomen, 2007; Vroomen & Stekelenburg, 2010).

Due to the differences in the characteristics of the two types of visual cues described above, comparing their contributions to speech processing remains difficult and needs a strictly controlled experimental protocol. Given this constraint, the present study narrowed down its investigation by focusing on one specific feature that is common to both types of visual cues, that is, the fact that both articulatory gestures and orthography provide information on the phonological content of speech sounds. More specifically, we examined whether, in skilled readers, the phonemic information extracted from visual articulatory cues affects speech processing in the same way as that extracted from orthographic cues. Here, the dynamic and predictive value that is specific to articulatory movements was removed by using static images of lip shapes representing the visemes of spoken words' initial phonemes instead of dynamic articulatory gestures producing the entire words. Likewise, the orthographic cues corresponded to the first graphemes of spoken words. The main aim of the present investigation was to examine whether these two kinds of static visual inputs are similarly integrated with the speech signal when presented synchronously and, if so, whether such integration extended beyond the initial stage of acoustic-phonetic analysis typically reported in the literature.

To address this issue, we conducted an EEG study in which brain responses to spoken words and the two kinds of visual cues were recorded when participants were presented with auditory, visual, and audiovisual stimuli. In the bimodal conditions, the onsets of the auditory and visual inputs were perfectly synchronized. AV integration was examined using the additive model, assuming that integration occurs whenever the activity measured in the audiovisual minus visual-only conditions is different from that observed in the auditory-only condition (i.e., AV - V ≠ A ; for a review, see Baart, 2016). An early integration was expected to induce a modulation of the amplitude and/or latency of the N1 and P2 auditory evoked potentials which reflect the initial acoustic-phonetic stages of auditory processing. In addition to these early components that are classically examined in the AV integration literature, we extended our investigation to a later speech processing stage that remains unexplored, i.e., the P350 component, which reflects the initial activation of a cohort of words in the mental lexicon that overlap with the information presented in the audio (and/or visual) sensory inputs (Friedrich et al., 2013; Schild et al., 2011). Although both viseme and grapheme can activate the initial phoneme of a spoken word, to our best knowledge, their relative contributions to the activation of a cohort of spoken words has never been compared. This could be explained by the fact that most studies that examined AV integration used meaningless syllables, which did not allow an examination of lexical access. As a methodological note, applying the additive model to examine the AV integration on late cognitive processes might also raise some issues. According to Besle et al. (2004, see also Molholm et al., 2002), the additive model is valid only when the auditory, visual, and audiovisual brain responses do not include common activity that would be summed up, such as neural responses related to late semantic processes, target processing, response selection or motor process. In our protocol, the spoken inputs were real words while the visual cues corresponded to a single letter or a static image of viseme without a semantic content. It is thus unlikely that they recruit the same cognitive process. Additionally, as described in the Method section, our tasks only focused on speech inputs, and motor responses were given only on a few auditory and audiovisual catch trials that were removed from EEG analyses.

In addition to characterizing the two kinds of AV integration at different processing stages, we also examined the role of two additional factors in the integration process, i.e., task demands and congruency between the auditory and visual inputs. The manipulation of task demands would allow us to examine the automaticity of the integration process. As mentioned above, most studies that examined AV integration typically focused on the processing of meaningless syllables in low-level perceptual tasks, which did not allow examining the role of top-down factors on the integration process. By comparing the impact of AV integration in both phonemic decision and lexical decision tasks, we argued that if AV integration operates independently of the top-down influence of task-demands and participants' attention to the initial phoneme vs. the entire stimulus (lexical status), the same pattern of integration would be observed in both tasks. Finally, the manipulation of the congruency between the initial phonemes contained in speech inputs and in the visual cues would inform us to what extent, the visual information is processed: If AV integration is dependent on whether the visual input matches the phonemic content of the speech signal, an impact of the congruency between the auditory and visual inputs on AV integration would be observed. In contrast, if the mechanism leading to AV integration depends on the mere presence of multisensory inputs regardless of their information content, no impact of audiovisual congruency would be observed.

2. Methods

2.1. Participants

Twenty healthy adults (17 females and 3 males), with a mean age of 22 years (±2 SD, range: 18-28 years) participated in the study. All participants were native French speakers, with a mean of 14 years (±2 SD, range 12-17 years) of education. They were all right-handed according to the standard handedness inventory (Oldfield, 1971) with a mean score of 86% (±16 SD), had normal or corrected-to-normal vision, and reported no history of hearing, speaking, language, neurological and/or neuropsychological disorders. The protocol was carried out in accordance with the ethical standards of the 1964 Declaration of Helsinki. An informed consent was obtained from each participant. All participants were compensated for the time spent in the study.

2.2. Stimuli

2.2.1. Viseme recognition

Multiple utterances of /a/, /i/, /o/, /fa/, /pa/, /ta/ and /ʒa/ syllables were individually recorded by a male native French speaker, using a high-quality digital video camera. The speaker produced each stimulus, maintaining an even intonation, tempo and vocal intensity. Video digitizing (centered on the speaker’s mouth presented against a blue background) was done at 30 frames per second with a resolution of 1920 x 1080 pixels. Audio digitizing was done at 48 kHz with 16-bit analog-to-digital conversion using an AKG C1000S microphone connected to the camera. One clearly articulated token was selected per syllable. Seven images of static lip shape corresponding to the visemes of the seven consonants and vowels were extracted from video recordings using Adobe Premiere (Adobe systems, Inc., San Jose, USA). To do so, the frame immediately preceding the acoustic onset was selected as the most representative and contrastive viseme. The visemes and their corresponding phonemes are shown in Figure 1.

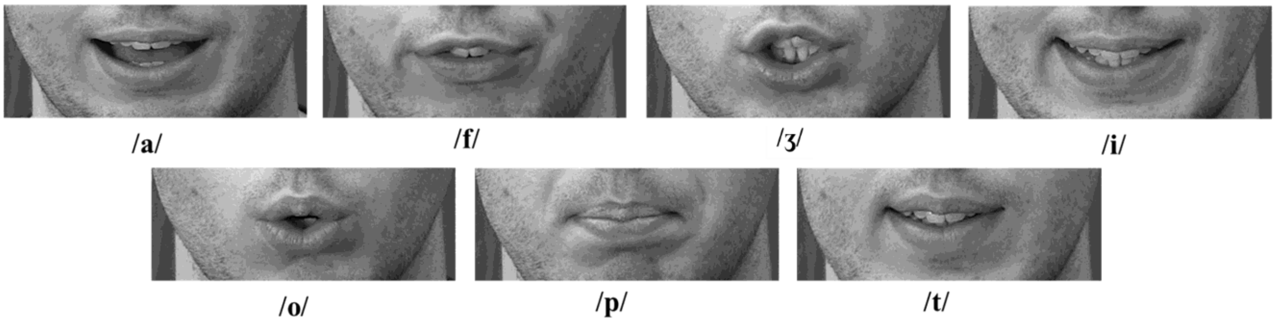

Figure 1: The visemes and their corresponding phonemes

2.2.2. Main tasks

A go/no-go paradigm was used in both phonemic and lexical decision tasks. In the phonemic decision task, participants had to detect spoken words that started with the /ʒ/ consonant. In the lexical decision task, participants had to detect spoken pseudowords.

As critical stimuli presented in the no-go trials, 108 French disyllabic nouns and adjectives beginning with the following phonemes were selected: /f/, /p/, /t/, /a/, /i/ and /o/ (18 words per initial phoneme). These initial phonemes were selected to cover a wide range of visual articulatory gestures and to be visually contrastive from one another. To avoid sound-spelling inconsistencies, the initial phonemes of the selected words were always spelled with the graphemes <f>, <p>, <t>, <a>, <i> and <o>, respectively. Based on the *Lexique* database (New et al., 2004), the stimuli from each phoneme category were further divided into two lists matched on the mean number of phonemes and letters, spoken word frequency, written word frequency, phonological neighbourhood, orthographic neighbourhood, phonological uniqueness point and orthographic uniqueness point (all *ps* > .15; see Appendix A for the characteristics of the stimuli). In half of the participants, the first list was presented in the phonemic decision task and the second list in the lexical decision task. The relationship between list and task was reversed in the other half of the participants.

In the phonemic decision task, 12 additional disyllabic words with the /ʒ/ consonant (spelled with the grapheme <g>) at the initial position were selected for go trials. The psycholinguistic variables associated with these stimuli were within the same range as the no-go trials (see Appendix A). In the lexical decision task, 12 additional disyllabic pseudowords were generated for go trials. All pseudowords began with the six phoneme categories used in the no-go trials.

All stimuli described above were recorded by a male French native speaker in a soundproof room at a sampling rate of 48 kHz with 16-bit analog-to-digital conversion using an AKG C1000S microphone. Each stimulus was recorded twice. The best token of each stimulus was selected based on auditory inspection. Using the Praat software (Boersma & Weenink, 2013), each stimulus was manually cut, at zero crossing points, based on waveform and spectrogram information. The stimuli were matched for intensity (mean value ± SD: 77dB ± 2).

In addition to the auditory stimuli, two kinds of visual stimuli were used. The first kind (hereafter, $\mathrm{V}_{\mathrm{VISEME}}$) was the static lip shape images corresponding to /f/, /ʒ/, /p/, /t/, /a/, /i/ and /o/ visemes (see the stimulus description in the viseme recognition task). The second kind was the following graphemes: <f>, <g>, <p>, <t>, <a>, <i>, and <o> (hereafter, $\mathrm{V}_{\mathrm{GRAPHEME}}$).

The stimuli described above were used to generate the final go and no-go materials. The no-go material consisted of seven experimental conditions: Auditory word (AUD), auditory word associated with the viseme of the word's initial phoneme ($\mathrm{AV}_{\mathrm{VISEME/CONGRUENT}}$), auditory word associated with the grapheme of the word's initial phoneme ($\mathrm{AV}_{\mathrm{GRAPHEME/CONGRUENT}}$), auditory word associated with a viseme that does not correspond to the word's initial phoneme ($\mathrm{AV}_{\mathrm{VISEME/INCONGRUENT}}$), auditory word associated with a grapheme that does not correspond to the word's initial phoneme ($\mathrm{AV}_{\mathrm{GRAPHEME/INCONGRUENT}}$), viseme without auditory input ($\mathrm{V}_{\mathrm{VISEME}}$), and grapheme without auditory input ($\mathrm{V}_{\mathrm{GRAPHEME}}$). In the go trials of both tasks, only the first five conditions were included.

2.3. Procedure

All tasks were conducted in a sound-attenuated room. Participants sat in front of a computer monitor at approximately 50 cm. The acoustic stimuli were presented through insert earphones at the same comfortable sound level for all participants. The E-prime 3.0 software was used for controlling stimulus presentation and collecting participants' responses (Psychology Software Tools, Pittsburgh, PA).

2.3.1. Viseme recognition

To ascertain that all participants were familiar with the visemes used in the study, before performing the phonemic and lexical decision tasks, the participants were exposed to the associations between these visemes and their corresponding speech sounds and graphemes. The associations between the three elements were presented ten times for each viseme used in the main tasks: /a/, /i/, /o/, /f/, /ʒ/, /p/, and /t/. The presentation order of these 70 trials was randomized.

Following the exposure phase, the participants completed an active training with corrective feedback through a viseme-grapheme matching task. Within each trial, a viseme was presented on the screen with seven graphemes: <f>, <g>, <p>, <t>, <a>, <i>, <o>. The participants were instructed to click on the grapheme that represented the sound produced by the viseme within 4 s. Once a response had been registered, or when 4 s had elapsed, a feedback message ("correct", "incorrect" or "please respond faster", in French), the viseme, the correct grapheme and the associated speech sound were presented for 1.5 s. The task was divided in two blocks of 35 trials. Each viseme was presented five times in each block. The presentation order of the visemes was randomized. In total, the familiarization phase lasted about 10 min.

2.3.2. Main tasks

In both phonemic decision and lexical decision tasks, the participants were presented with a total of 378 no-go trials and 60 go trials in a random order. These corresponded to 54 no-go trials for each of the seven experimental conditions (AUD, $\text{AV}_{\text{VISEME/CONGRUENT}}$, $\text{AV}_{\text{GRAPHEME/CONGRUENT}}$, $\text{AV}_{\text{VISEME/INCONGRUENT}}$, $\text{AV}_{\text{GRAPHEME/INCONGRUENT}}$, V_{VISEME}, $\text{V}_{\text{GRAPHEME}}$) and 12 go trials for each of the five auditory and audiovisual conditions described above. The same spoken words were used in all auditory and audiovisual conditions, which allowed us to control for possible impacts of stimuli's acoustic and psycholinguistic features across conditions. In the V_{VISEME} and $V_{GRAPHEME}$ conditions, each of the six no-go visemes and graphemes was presented nine times. Altogether, the stimuli were presented in five blocks of 88 trials (the first block began with two buffer items to prepare the participants to the task. These items were not analyzed).

In the AUD condition, only an auditory input was presented, and the screen remained blank. In the V_{VISEME} and $V_{GRAPHEME}$ conditions, a viseme or a grapheme was presented at the center of the screen for 500 ms, and no auditory input was provided. In the audiovisual conditions, both auditory and visual inputs were presented. Their onsets were synchronized, and the duration of the visual input was kept constant at 500 ms. In all conditions, the inter-stimulus interval varied between 1.8 and 2.1 s. The screen remained blank during this period. In the phonemic decision task, participants were instructed to press the response button whenever they detected a spoken word that began with the /ʒ/ consonant. During the lexical decision task, they were instructed to respond whenever they detected a pseudoword. By asking the participants to focus on the speech inputs, we kept the participants in a speech processing (rather than visual processing) context. This allowed us to ascertain that if the AV integration occurred, it would not be strategically induced or forced by the tasks. Although only the auditory inputs were relevant to the tasks, the participants were instructed to fixate the center of the screen at all time. The two tasks were of equal duration, each lasting about 20 minutes. Their order was counterbalanced across participants, with a short break offered between them. At the beginning of each task, a short training was performed.

2.4. EEG data recording

During the main tasks, EEG data were continuously recorded from 64 scalp electrodes according to the international 10–20 system and using the Biosemi Active Two AD-box EEG system operating at a sampling rate of 512 Hz. Two additional electrodes served as reference (common mode sense [CMS] active electrode) and ground (driven right leg [DRL] passive electrode). Two other external reference electrodes were placed at the left and right mastoids. The electro-oculograms measuring horizontal (HEOG) and vertical (VEOG) eye movements were recorded using electrodes at the outer canthus of each eye as well as above and below the left eye. Before the experiment, the impedance of all electrodes was adjusted to obtain low offset and stable DC voltages.

2.5. EEG data processing

EEG data from no-go trials were processed using the EEGLAB software (Delorme & Makeig, 2004; version 2020) running on Matlab (Mathworks, Natick, USA; version R2019a). For each participant, each task and each experimental condition, EEG data were first re-referenced to the average of left and right mastoids, and band-pass filtered using a two-way least-square FIR filtering (1–30 Hz). Residual sinusoidal noise from scalp channels was further estimated and removed using the EEGLAB CleanLine plug-in (version 2012). Scalp channels were then automatically inspected, and bad channels interpolated using the EEGLAB Clean_rawdata plug-in (version 0.34). On all channels, speech-related, eye blinks, eye movements and other motion artefacts were detected and removed using the EEGLAB Artifact Subspace Reconstruction plug-in (version 0.13). Based on a sliding-window principal component analysis, this algorithm rejected high-variance bad data periods by determining thresholds based on clean segments of EEG data. EEG data were then segmented into 700-ms epochs including a 100-ms pre-stimulus baseline (from −100 to 0 ms relative to the onset of the acoustic signal in the auditory and audiovisual conditions or to the onset of visual cues in the visual conditions) and lasting until 600 ms post-stimulus onset. Epochs with an amplitude change exceeding ± 100 µV at any channel (including HEOG and VEOG channels) were further rejected. On average, the entire preprocessing pipeline rejected 17% (±4 SD) and 16% (±5 SD) in the phonemic and lexical decision tasks, respectively.

For each participant and each task, we used an additive model to test the AV integration, in which the auditory EEG signal was compared to the difference between audiovisual and visual EEG signals. To this aim, EEG signals obtained in the V_{VISEME} and $V_{GRAPHEME}$ conditions were subtracted from those obtained in the corresponding AV conditions in the following manners: AV$_{\text{GRAPHEME/CONGRUENT}}$ – V$_{\text{GRAPHEME}}$; AV$_{\text{GRAPHEME/INCONGRUENT}}$ – V$_{\text{GRAPHEME}}$; AV$_{\text{VISEME/CONGRUENT}}$ – V$_{\text{VISEME}}$; AV$_{\text{VISEME/INCONGRUENT}}$ – V$_{\text{VISEME}}$. Each of the resulting ERP difference waves (hereafter, *difERP*) was compared against the signal obtained in the AUD condition, based on the assumption of the additive model (Baart, 2016) that AV integration occurs whenever the *difERP* signals were different from the signal obtained in the AUD condition (AV-V ≠ A) in either direction (supra-additive or sub-additive).

Additionally, the impact of the type of visual input (grapheme vs. viseme) on AV integration were examined by comparing the AV$_{\text{GRAPHEME/CONGRUENT}}$ – V$_{\text{GRAPHEME}}$ *difERP* signal to the AV$_{\text{VISEME/CONGRUENT}}$ – V$_{\text{VISEME}}$ *difERP* signal. Finally, the impact of AV congruency for each type of visual input was obtained by comparing the AV$_{\text{GRAPHEME/CONGRUENT}}$ - V$_{\text{GRAPHEME}}$ to the AV$_{\text{GRAPHEME/INCONGRUENT}}$ - V$_{\text{GRAPHEME}}$ *difERP* signal, and AV$_{\text{VISEME/CONGRUENT}}$ – V$_{\text{VISEME}}$ to the AV$_{\text{VISEME/INCONGRUENT}}$ – V$_{\text{VISEME}}$ *difERP* signal, respectively.

Based on the literature and the visual inspection of the grand average ERP signal, three separated time-windows that corresponded to three ERP components of interest were selected: N1 (70–150 ms), P2 (150–250 ms) and P350 (300-400 ms). In each time-window, individual peak amplitude and peak latency of the ERP signals obtained in the AUD condition and the *difERP* signals described above were extracted from six electrode clusters covering the whole brain[1]: Fronto-central (F1, Fz, F2, FC1, FCz, FC2, C1, Cz, C2), left anterior (Fp1, AF3, AF7, F7, F5, F3, FC3, FC5, FT7), left posterior (CP3, CP5, TP7, P3, P5, P7, PO3, PO7, O1), right anterior (Fp2, AF4, AF8, F4, F6, F8, FC4, FC6, FT8), right posterior (CP4, CP6, TP8, P4, P6, P8, PO4, PO8, O2), centro-parietal (C1, Cz, C2, CP1, CPz, CP2, P1, Pz, P2). Because ERP waveforms generally vary from one electrode cluster to another and this common observation is not critical for the present

[1] Since the existing literature does not provide an a priori hypothesis regarding the scalp localization of the AV integration on the P350 component, we chose to conduct the analyses covering the whole brain rather than on a specific electrode cluster.

study, the main effects of cluster that occurred in most analyses are described in detail in Appendix B.

For each component of interest, an ANOVA considering task (phonemic decision; lexical decision), cluster (six levels), and condition (one ERP and four *difERPs*) as within-subject factors was conducted on peak amplitude and peak latency. The Greenhouse–Geisser correction was applied (Greenhouse & Geisser, 1959), and the corrected degrees of freedom and p-values are reported. The significant effects of condition and their interactions with the other factors were further analyzed using planned pairwise comparisons to examine the effects of interest described above. When required, unplanned post-hoc analyses were conducted with Bonferroni corrections. Figure 3 showed the waveforms of the ERP obtained in the auditory condition and the four *difERPs.*

3. Results

3.1. Behavioral data

3.1.1. Viseme recognition

A one-way ANOVA treating viseme as a within-subject factor was conducted on participants' percentage of correct responses. The average performance was above 90%. Coherently with the existing literature, the analysis showed significant differences across visemes [$F(6, 114) = 4.65$, $p = .0003$, $p\eta 2 = .20$] (Fisher, 1968; Summerfield, 1987). This was due to a lower accuracy score obtained on the viseme of the /t/ consonant (81%) compared to the other visemes (/a/ = 96%, /f/ = 95%, /ʒ/ = 94%, /o/ = 96%, /p/ = 96%, all *ps* < 0.01, corrected for multiple comparisons using Bonferroni correction) except /i/ (90%, $p = 0.40$). Although the instructions did not emphasize on response speed, the general tendency of the reaction time (RT) data confirmed that some visemes were more difficult to process than the others [$F(6, 114) = 21.17$, $p < .00001$, $p\eta 2 = .53$: /a/ = 1205 ms, /f/ = 1587 ms, /ʒ/ = 1426 ms, /i/ = 1530 ms, /o/ = 1203 ms, /p/ = 1298 ms, /t/ = 1863 ms]. The mean RT obtained on the viseme of /t/ was significantly longer than those obtained on the other visemes, *ps* ≤ .005; The mean RTs on the visemes of /f/ and /i/ were significantly longer than those obtained on the visemes of /a/, /o/ and /p/, *ps.* < 05, the p values were corrected for multiple comparisons using Bonferroni correction].

3.1.2. Main tasks

Statistical analyses were performed on the performance obtained on-go trials. Separated repeated-measure ANOVAs were performed on the percentage of correct responses and the RTs on correct trials. In each task, the RTs smaller or larger than the mean RT of all participants ±2.5 SD were excluded from the analysis. Task (phonemic decision, lexical decision) and condition (AUD, $\text{AV}_{\text{VISEME/CONGRUENT}}$, $\text{AV}_{\text{VISEME/INCONGRUENT}}$, $\text{AV}_{\text{GRAPHEME/CONGRUENT}}$, $\text{AV}_{\text{GRAPHEME/INCONGRUENT}}$) were treated as within-subject factors.

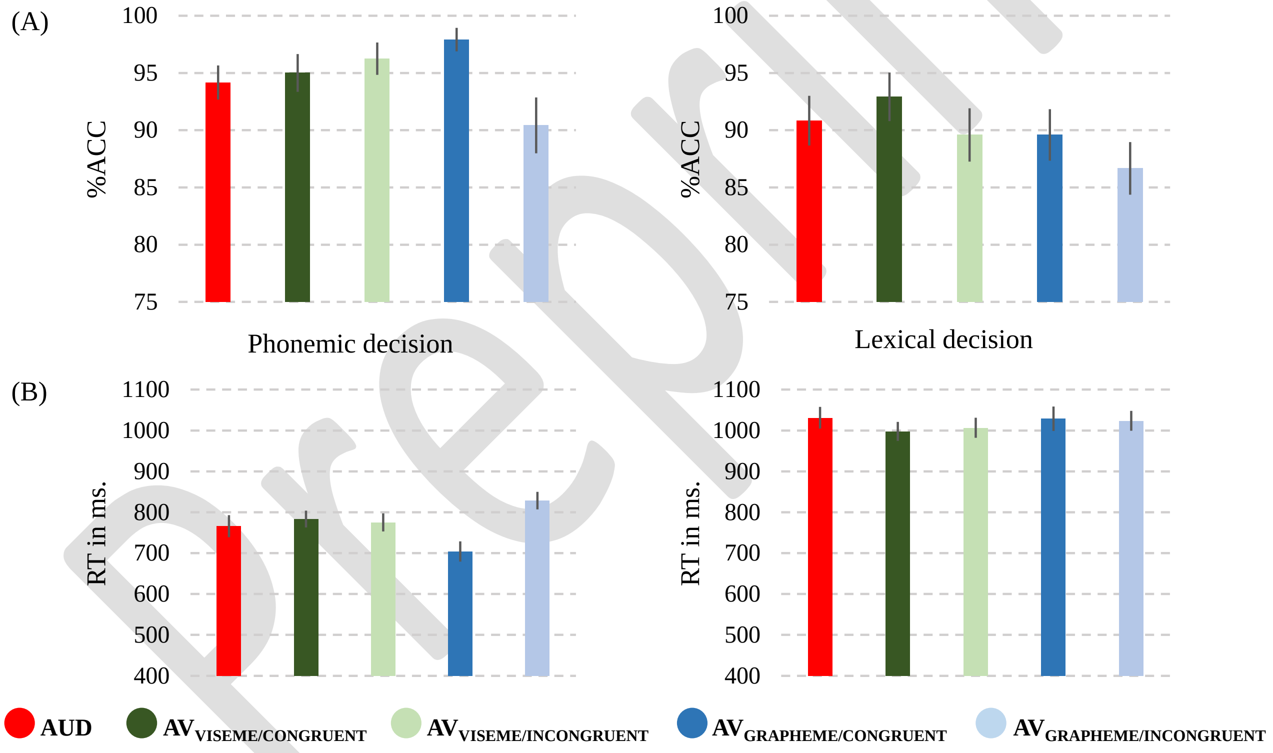

Figure 2: A) Mean percentage of correct responses and B) mean reaction times (in ms) on correct trials obtained in the phonemic decision (left panel) and lexical decision (right panel) tasks. Error bars represent the standard error of the means.

The analysis performed on the percentage of correct responses showed significant main effects of task [*F*(1, 19) = 7.33, *p* = .013, *pη2* = .28] and condition [*F*(4, 76) = 4.33, *p* = .003, *pη2* = .19].

The interaction between the two factors was not significant [*F*(4, 76) = 1.72, *p* = .155, *pη2* = .08]. As illustrated in Figure 2A, the performance obtained in the phonemic decision (94.75%) was higher than that obtained in the lexical decision task (89.82%). The condition effect reflected a lower performance obtained in the $\mathrm{AV}_{\mathrm{GRAPHEME/INCONGRUENT}}$ condition compared to the other conditions (*ps* < .01).

The analysis of the RTs showed significant effects of task [*F*(1, 19) = 119.73, *p* < .0001 *pη2* = .86], condition [*F*(4, 76) = 7.85, *p* < .0001, *pη2* = .29] and their interaction [*F*(4, 76) = 9.06, *p* < .0001, *pη2* = .32]. As illustrated in Figure 2B, participants were faster to identify the initial phoneme than to identify the lexical status of spoken inputs, which was likely because phonemic decisions could be made without waiting until the end of the speech signal. Further analyses of the interaction between task and condition indicated that the condition effect was significant only in the phonemic decision task [*F*(4, 76) = 21.90, *p* < .0001, *pη2* = .53] where the mean RT obtained in the $\mathrm{AV}_{\mathrm{GRAPHEME/INCONGRUENT}}$ condition was longer than those observed in the other conditions and the mean RT obtained in the $\mathrm{AV}_{\mathrm{GRAPHEME/CONGRUENT}}$ condition was shorter that those observed in the other conditions (*ps* < .001).

Altogether, the behavioral measures showed that the performances obtained in both tasks were sensitive to the congruency between speech sounds and orthographic cues, although a more reliable impact was observed in the phonemic decision task. No evidence for the impact of viseme was revealed in the behavioral measures.

3.2. EEG data

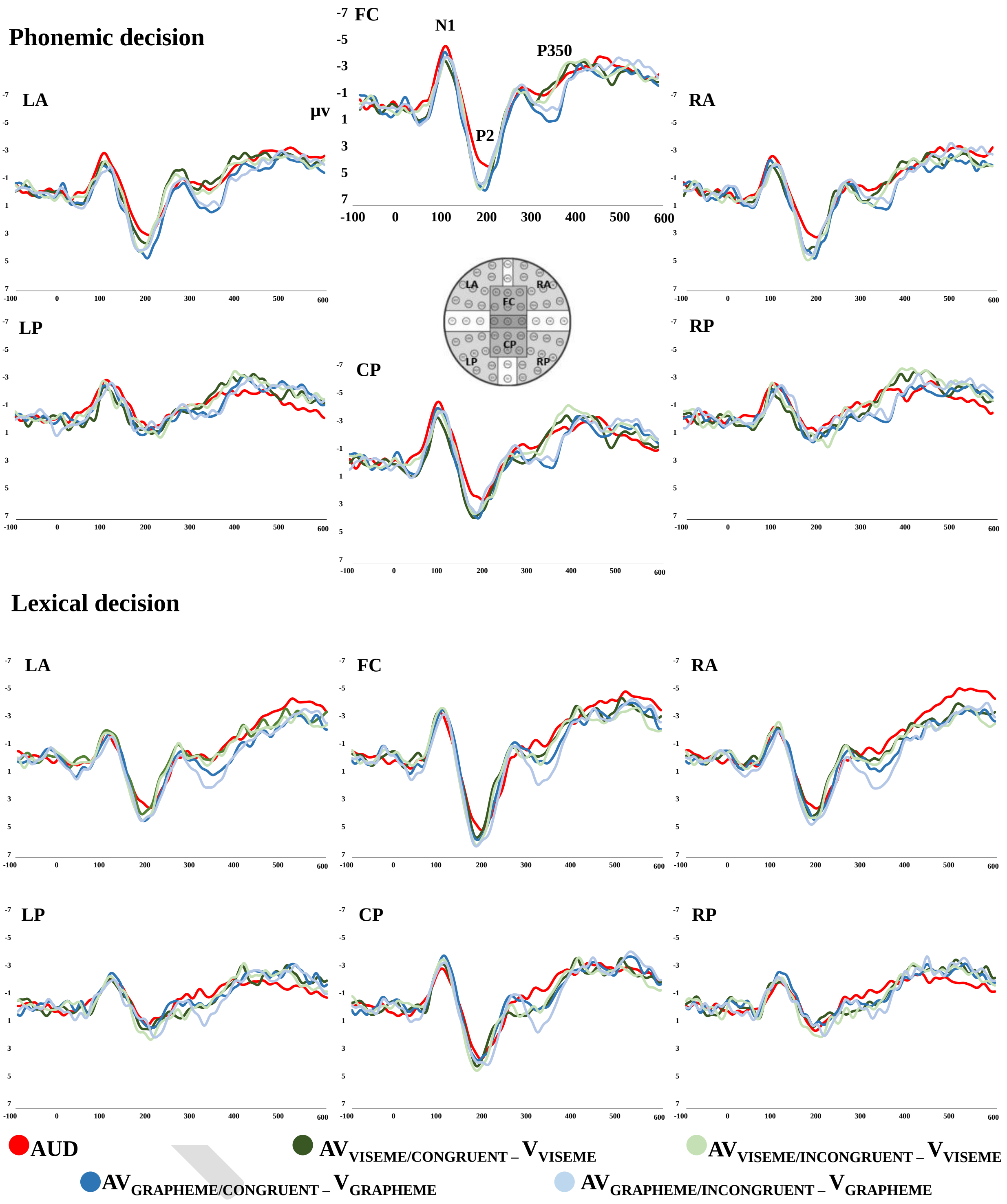

Figure 3: Waveforms of the ERP obtained in the phonemic decision and lexical decision tasks in the auditory condition and the following different ERPs (AV-V): AV$_{VISEME/CONGRUENT}$ – V$_{VISEME}$; AV$_{VISEME/INCONGRUENT}$ – V$_{VISEME}$; AV$_{GRAPHEME/CONGRUENT}$ – V$_{GRAPHEME}$; AV$_{GRAPHEME/INCONGRUENT}$ – V$_{GRAPHEME}$. The six electrode clusters are highlighted in gray in the templates. FC: fronto-central, CP: centro-parietal; LA: left anterior; LP: left posterior; RA: right anterior; RP: right posterior.

3.2.1. N1: 70-150 ms

ANOVA performed on N1 peak amplitude only showed a significant effect of cluster [*F*(2.341, 44.467) = 12.94, *p* < .0001, *pη2* = .41]. A similar finding was obtained in the analysis conducted on peak latency [*F*(2.932, 55.708) = 4.22, *p* = .01, *pη2* = .18]. For both dependent variables, no other main effects or interactions were significant (*ps* ≥ .20), which clearly suggests an absence of AV integration on this early acoustic processing component.

3.2.2. P2: 150-250 ms

ANOVA performed on P2 peak amplitude showed a significant main effect of cluster [*F*(2.426, 46.094) = 48.07, *p* < .0001, *pη2* = .72]. Interestingly, the effect of condition [*F*(2.391, 45.435) = 5.19, *p* = .006, *pη2* = .21] and its interaction with cluster were also significant [*F*(7.327, 139.212) = 2.08, *p* = .047, *pη2* = .10]. Given that these effects did not interact with task (*Fs* < 1), we combined the data obtained in the phonemic and lexical decision tasks together and examined the condition effect within each cluster. The results of these analyses showed a significant effect of condition in all clusters [fronto-central: *F*(2.641, 50.170) = 4.29, *p* = .012, *pη2* = 0.18; centro-parietal: *F*(2.751, 52.269) = 3.25, *p* = .032, *pη2* = 0.15; left anterior: *F*(2.522, 47.914) = 6.30, *p* = .002, *pη2* = 0.25; left posterior: *F*(2.634, 50.043) = 3.66, *p* = .023, *pη2* = 0.16; right anterior: *F*(2.461, 46.751) = 5.05, *p* = .007, *pη2* = 0.21; right posterior: *F*(2.670, 50.732) = 3.90, *p* = .017, *pη2* = 0.17]. As illustrated in Figure 4, the significant effects of condition were due to an enhancement of P2 peak amplitude elicited by both types of AV stimuli (*difERPs*) compared to the ERP elicited by auditory stimuli alone, thus, reflecting a supra-additive AV integration in a wide range of brain areas (ps ≤ .05 for all comparisons except for the $\mathrm{AV}_{\mathrm{VISEME/CONGRUENT}} - \mathrm{V}_{\mathrm{VISEME}}$ vs. AUD contrast in the left anterior cluster and the $\mathrm{AV}_{\mathrm{GRAPHEME/CONGRUENT}} - \mathrm{V}_{\mathrm{GRAPHEME}}$ vs. AUD contrast in the right posterior cluster where the results were marginal, *p* = .09 and *p* = .07, respectively). The analysis comparing the degree of AV integration induced by graphemes and by visemes showed a higher degree of AV integration between speech sounds and graphemes in the left anterior cluster (*p* = .009). No difference between the two kinds of visual cue was found in the other clusters (*ps* > .35). Finally, the degree of integration was not sensitive to the congruency between the auditory and visual inputs for either kind of visual cue (*ps* > .05 in all clusters).

ANOVA performed on P2 peak latency only showed a significant main effect of cluster [*F*(2.686, 51.043) = 13.07, *p* < .0001, *pη2* = .41]. No other main effects or interactions showed significant result (*ps* > 0.05).

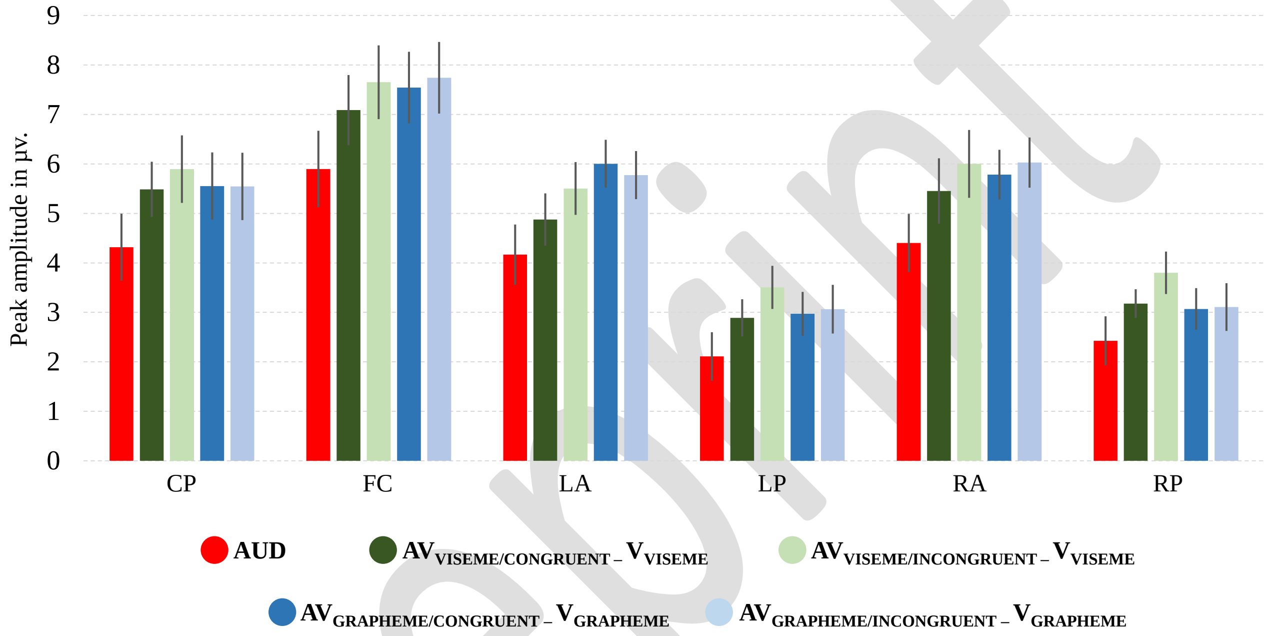

Figure 4: Mean P2 peak amplitudes (in µv) corresponding to the ERP obtained in the auditory condition and the following different ERPs (AV-V): $AV_{VISEME/CONGRUENT} - V_{VISEME}$; $AV_{VISEME/INCONGRUENT} - V_{VISEME}$; $AV_{GRAPHEME/CONGRUENT} - V_{GRAPHEME}$; $AV_{GRAPHEME/INCONGRUENT} - V_{GRAPHEME}$. The findings obtained in the phonemic decision and lexical decision are averaged. Error bars represent the standard error of the means. FC: fronto-central, CP: centro-parietal; LA: left anterior; LP: left posterior; RA: right anterior; RP: right posterior.

3.2.3. P350: 300-400 ms

ANOVA performed on P350 peak amplitude showed a significant main effect of cluster [*F*(2.387, 45.356) = 4.33, *p* = .014, *pη2* = .19]. The effect of condition was significant [*F*(2.563, 48.703) = 6.04, *p* = .002, $p\eta 2$ = .24] but did not interaction with the other main factors. As illustrated in Figure 5, the condition effect reflected an overall enhancement of P350 peak amplitude observed in all the AV conditions compared to the auditory alone condition. Interestingly, the degree of AV integration seems stronger when the visual inputs were graphemes (*p* = .0002 and *p* = 0.00003 for the $AV_{GRAPHEME/CONGRUENT}$ – $V_{GRAPHEME}$ vs. AUD and $AV_{GRAPHEME/INCONGRUENT}$ – $V_{GRAPHEME}$ vs. AUD contrast, respectively) than when they were visemes (*p* = .04 and *p* = 0.015 for the $AV_{VISEME/CONGRUENT}$ – V_{VISEME} vs. AUD and $AV_{VISEME/INCONGRUENT}$ – V_{VISEME} vs. AUD contrast, respectively). The direct comparison of the $AV_{GRAPHEME/CONGRUENT}$ – $V_{GRAPHEME}$ and the $AV_{VISEME/CONGRUENT}$ – V_{VISEME} *difERPs* showed a marginal difference for this tendency (*p* = .08). For either type of visual cues, the congruency between the auditory and visual input did not affect the degree of AV integration (*ps* > .50). No significant difference was found in the ANOVA conducted on P350 peak latency (*ps* > .10).

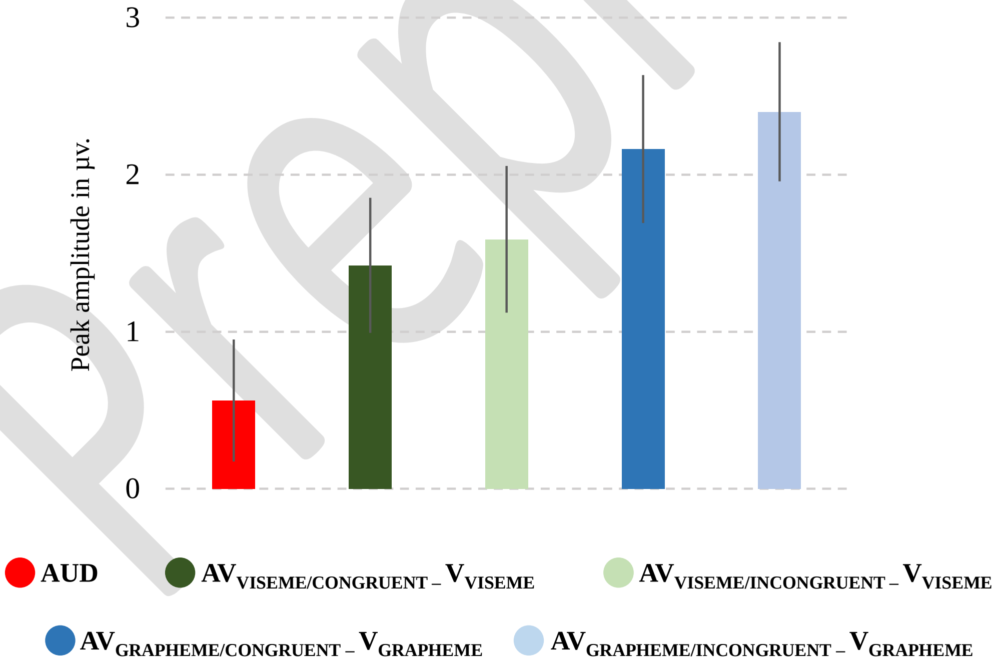

Figure 5: Mean P350 peak amplitudes (in µv) corresponding to the ERP obtained in the auditory condition and the following different ERPs (AV-V): $AV_{VISEME/CONGRUENT}$ – V_{VISEME}; $AV_{VISEME/INCONGRUENT}$ – V_{VISEME}; $AV_{GRAPHEME/CONGRUENT}$ – $V_{GRAPHEME}$; $AV_{GRAPHEME/INCONGRUENT}$ –

V_{GRAPHEME}. The findings obtained in the phonemic decision and lexical decision tasks and in the six clusters are averaged. Error bars represent the standard error of the means.

4. Discussion

The result obtained in the viseme recognition task showed that participants were able to match the visemes with their corresponding phonemes at a good level of performance, although as previously reported in the literature (Fisher, 1968; Summerfield, 1987), the performance was not uniform across all visemes (mean %ACC = 92.8%, range = 81%-96%; mean RT = 1444 ms, range = 1203 -1863 ms). Despite this ability, presenting a viseme in synchrony with speech sounds did not have a significant impact on the performance obtained either in the phonemic decision task or in the lexical decision task. Only the presence of the graphemes revealed behavioral evidence of multisensory processing: In both tasks, the incongruency between grapheme and initial phoneme of spoken words significantly reduced the accuracy scores. In the phonemic decision task, RTs were shortest in the graphemic congruent condition and longest in the graphemic incongruent condition compared to the others. Overall, the behavioral measures showed that skilled-reader participants were sensitive to the congruency between speech sounds and orthographic cue in both sublexical and lexical tasks, although a stronger sensitivity (as shown by the modulation of RTs) was found in the former task. This latter observation is coherent with the fact that the manipulation of the relationship between speech sounds and the visual input was restricted to the sublexical (phonemic) unit.

Interestingly, this behavioral outcome differed from the pattern of AV integration revealed by brain responses in three main aspects. First, ERP evidence of AV integration was observed for both visemic and graphemic cues, although some analyses revealed a stronger degree of integration between speech sounds and graphemes. Second, no significant effect of congruency between the auditory and visual input was observed on any ERP component of interest. Finally, the same pattern of AV integration was observed in both phonemic and lexical decision tasks. Detailed discussions on the characteristics of AV integration observed at the three stages leading to spoken word recognition are presented below.

4.1. Absence of AV integration on N1: Absence of prediction

N1 has been considered as an ERP component that reflects the initial stage of acoustic processing, which is not specific to speech (Näätänen & Picton, 1987). We observed no hint of AV integration on this early auditory evoked potential either on its amplitude or on its latency. To our knowledge, most previous studies that reported the modulation of N1 during AV processing used either articulatory movements with their natural temporal dynamics or other leading visual cues (including written text) that provided a valid prediction of the auditory input. This prediction typically leads to a reduction of N1 amplitude and latency which indicates a reduction of the computational demands within the primary auditory cortex (Besle, Fort, Delpuech, et al., 2004; Klucharev et al., 2003; Pinto et al., 2019; Stekelenburg & Vroomen, 2007; Van Wassenhove et al., 2005; Vroomen & Stekelenburg, 2010). The absence of evidence of AV integration on N1 reported here is coherent with the fact that, in our protocol, the auditory and visual inputs were presented synchroneously, thus, eliminating the predictive value of the visual inputs.

4.2. Impact of AV integration on P2: A combination of a non-specific AV integration and a specific increase of sensitivity to the graphemic cues in the left anterior electrode cluster

In the absence of prediction, the pattern of AV integration observed on P2 is clearly distinct from that observed on N1, which suggests that P2 modulation does not strictly depend on the predictive value of visual input. In this time-window, a visual cue presented in synchrony with speech signal led to a significant increase of neural responses compared to the combination of ERPs elicited by unimodal auditory and visual inputs (i.e., AV > A+V, or AV-V > A as computed in the presented study). This supra-additive integration was found in all electrode clusters.

Interestingly, two distinct patterns of AV integration were found on different electrode sites, likely indicating that different neural generators of multisensory integration operate under different constraints. First, in most clusters (including the fronto-central one where the AV integration on the P2 component is typically reported in the literature), AV integration occurred independently of task demands, congruency between the auditory and visual inputs and type of visual cue. This pattern seems to reflect a general multisensory integration mechanism by which a co-occurrence of multisensory inputs, regardless of their relationship, would be sufficient to enhance (or reduce) neural responses beyond the combination of the activity elicited by unimodal inputs (Meredith, 2002; Stein & Stanford, 2008). Second, a more specific pattern was observed in the left anterior cluster. Here, both types of visual cues also led to a supra-additive AV integration but the degree of integration between speech sounds and graphemes was significantly stronger than that observed between speech sounds and visemes. While it is not possible to directly infer the cortical source(s) of brain responses recorded from surface electrodes, the fact that this specific pattern of AV integration occurred in the left hemisphere strongly suggests that its underlying mechanism might be related to language processing. In line with this assumption, Xu et al. (2019) reported MEG evidence for an integration between speech sounds and written characters that was located on left anterior sensors. At the cortical source level, the authors found that the integration that took place around 205-365 ms. (which corresponds to the P2 and P350 time-windows reported here) was located in left angular and supramarginal gyri, thought as heteromodal areas related to linking orthographic representation encoded in the occipital lobe to phonological representation encoded in the superior temporal gyrus (Price, 2000; Pugh et al., 2000; Schlaggar & McCandliss, 2007).

Two plausible cognitive processes leading to the bias towards speech sounds-graphemes association could be advanced at this stage of research. The first explanation is related to the strength of the link between speech sounds and visual cues: Graphemes generally provide more salient and less ambiguous information on phonemic content than visemes, especially in skilled readers. Thus, the access to phonemic information and the subsequence formation of a coherent AV perceptual unit would be facilitated. Indeed, despite the overall high performance obtained in the viseme recognition task, some visemes (e.g., /t/, /i/) were more difficult to recognize than the others. Such ambiguity is not expected in the association between graphemes and phonological representations, at least for the sound-spelling associations used here. However, the assumption that the facility of access to phonemic content is the only critical factor seems somehow incompatible with the absence of the congruency effect on the ERP data: In the incongruent condition, neither type of visual cue provided a valid phonemic content of speech sounds. If this information played the key role in the current integration process, one would also expect the degree of AV integration to depend on the congruency between the auditory and visual inputs. While the absence of the congruency effect on the ERP data remains intriguing and deserves further investigation, it could, to some extent, be due to the fact that in the present study the overlap between the auditory and visual inputs was only partial, since it involved only the first phoneme of disyllabic words. This partial (mis)match might drastically reduce the impact of audiovisual congruency compared to what has been reported in previous studies where the overlap typically involved the entire stimuli (Stekelenberg et al., 2007; Klucharev et al., 2003; Raij 2000). However, the absence of the congruency effect on ERP responses does not imply that this information was not processed. In fact, the presence of the congruency effect on task performances suggests otherwise, i.e., while this factor might not have a significant influence on the processing stages leading to lexical access, it could play a significant role during the decision-making stage.

The second explanation of the stronger integration between speech sounds and graphemes relies on the relative role of the two types of visual cue in the current speech processing contexts. During the experiment, the participants had to perform tasks that required an analysis of either a sublexical phonological unit or the lexical status of isolated spoken inputs. In these non-ecological and academic tasks, our participants, who were skilled readers, might be more sensitive to the graphemic cues than to the visemic cues. It has indeed been demonstrated that the neurocognitive state of the human brain is automatically adapted to the task to be performed (Sakai & Passingham, 2003). This pre-task adaptation may reflect an increase of attention to a specific feature of the stimuli even before the stimuli are actually presented (Corbetta & Shulman, 2002). The stronger sensitivity to the graphemic cues observed on the left anterior electrodes at this early stage of AV integration could be supported by such neural adaptation mechanism.

4.3. Impact of AV integration on P350: A trend towards a generalized sensitivity to the graphemic cues in all clusters.

Interestingly, the enhanced sensitivity to graphemes that was restricted to the left anterior cluster, observed during the phonetic processing stage, was generalized to all clusters in the subsequent stage. Once again, the same pattern of integration was found in both phonemic and lexical decision task which suggests that, at least up to lexical access, AV integration occurred regardless of task-demands, i.e., whether participants' attention was focused on the initial phoneme or on the entire stimulus. The P350 component has been reported to reflect the initial activation of a cohort of words in the mental lexicon that overlap with the information presented in the sensory inputs (Friedrich et al., 2013; Schild et al., 2011). This enhanced sensitivity to graphemes observed on P350 is coherent with findings from a number of ERP studies that reported significant contributions of orthographic knowledge to spoken word recognition. For instance, using a semantic decision task in which orthographic consistency of spoken words' onset was manipulated, Pattamadilok et al. (2009) reported that processing spoken words that began with a syllable that has more than one possible spelling (e.g., in French, the sound /e/ at the onset position could be spelled 'ai', 'é', 'e' or 'hé') induced a stronger ERP response at the early phase of lexical access compared to processing spoken words that began with a syllable that has only one possible spelling (see Chen et al., 2016; Perre et al., 2011; Perre & Ziegler, 2008 for similar observations). Also, several cross-modal priming studies consistently showed that orthographic primes facilitated the recognition of subsequent spoken words (Holcomb et al., 2005; Kiyonaga et al., 2007; Slowiaczek et al., 2003). While some behavioral findings suggest that articulatory gestures may also help to trigger lexical access and constrain lexical competition during speech recognition (Fort et al., 2013; Tye-Murray et al., 2007), to our best knowledge, no evidence on the underlying neural processes has been reported, and the locus of AV integration is still under debate (Ostrand et al., 2016). Our observation of speech sound-viseme integration at the initial phase of lexical access provides ERP evidence that is in line with these behavioral findings. However, it also suggests that, although visemes play a role during lexical access, their contribution appears to be more modest than that of graphemes, at least in skilled readers and in the context of academic speech processing tasks used here (see also Pattamadilok et al., in press, for a similar conclusion on the relatively reduced contribution of articulatory compared to orthographic cues to spoken word acquisition).

5. Conclusion

This study is among the very few that directly compared the two main types of AV integration that are specific to speech processing. Furthermore, it complements the existing findings by extending the analyses beyond the classical N1/P2 components, to the lexical processing stage. By focusing our investigation on the common feature of the graphemic and visemic visual cues, that is, their ability to provide information on the phonemic content of speech sounds, we found that the role of the two visual inputs in AV integration vary across processing stages. In the N1 time-window, which reflects the initial acoustic analysis, no AV integration was found. However, a supra-additive integration emerged in the P2 time-window. At this processing stage, we found a general increase of neural responses to both types of bimodal inputs in most electrode clusters which indicates a general multisensory integration mechanism (Meredith, 2002; Stein & Stanford, 2008). Interestingly, the cluster located in the anterior regions of the language dominant left hemisphere showed a more specific pattern, i.e., a higher sensitivity to association between speech sounds and graphemes than between speech sounds and visemes. This observation suggests that already at the stage where the cognitive system starts to differentiate speech from non-speech acoustic inputs (Baart et al., 2014), AV integration process itself also becomes more sensitive to the link between speech sounds and the abstract orthographic code than to the link between speech sounds and articulatory gestures. The higher sensitivity to orthography persisted and generalized in all brain regions in the subsequent P350 time-window, which emphasizes the role of orthography at the initial phase of lexical activation. This bias for the integration between speech sounds and the abstract orthographic code reported here provides further evidence for the claim that, even though orthography is linked with speech sounds in an arbitrary and abstract manner, once the link had become automatic, it strongly affects the way speech is processed. It also suggests that a significant part of the contribution of articulatory gestures on speech processing previously reported in the literature might be due to the dynamic and predictive cues present in natural visual speech. Once these cues are removed, the benefit of articulatory gestures on speech processing seems to be severely reduced.

Appendix A: Characteristics of the *no-go* words used in the phonemic decision and lexical decision tasks and the *go* words used in the phonemic decision task. The mean values (and standard deviations) were computed from the *Lexique* database (New et al., 2004)

	No-go trials	Go trials
Number of phonemes	4.61 (0.96)	4.83 (0.83)
Number of letters	5.94 (1.22)	6.25 (0.97)
Spoken word frequency	3.11 (4.03)	2.50 (2.55)
Written word frequency	5.69 (5.85)	4.92 (5.56)
Phonological neighbourhood	4.91 (4.77)	2.58 (2.43)
Orthographic neighbourhood	1.59 (1.75)	1.00 (1.35)
Phonological uniqueness point	4.44 (0.95)	4.50 (0.90)
Orthographic uniqueness point	5.37 (1.34)	5.00 (1.41)

Appendix B: Full descriptions of the cluster effects on peak amplitude and peak latency

	Peak amplitude (in µv)			Peak latency (in ms.)		
	N1	P2	P350	N1	P2	P350
Fronto-central	-4.72	7.19	1.77	103	189	332
Centro-parietal	-4.69	5.36	1.54	103	193	332
Left anterior	-3.05	5.27	2.00	106	191	339
Left posterior	-3.74	2.91	1.16	109	201	333
Right anterior	-3.09	5.53	2.10	104	192	335
Right posterior	-3.71	3.12	1.20	111	201	335

The following pairwise comparisons showed significant differences at p <.05 after Bonferroni corrections:

Peak amplitude

N1 Fronto-central > Left anterior, Left posterior, Right anterior, Right posterior

Centro-parietal > Left anterior, Left posterior, Right anterior, Right posterior

P2 Fronto-central > Centro-parietal, Left anterior, Left posterior, Right anterior, Right posterior

Left posterior < Centro-parietal, Left anterior, Right anterior

Right posterior < Centro-parietal, Right anterior

P350 Left anterior > Left posterior

Right anterior > Right posterior, Left posterior

Peak latency

N1 Right posterior > Fronto-central, Centro-parietal

P2	Left posterior > Fronto-central, Centro-parietal, Left anterior, Right anterior
	Right posterior > Fronto-central, Centro-parietal, Left anterior, Right anterior
P 350	/

Acknowledgements

C.P. was supported by the French National Research Agency: ANR-19-CE28-0001-01, ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) and the Excellence Initiative of Aix Marseille University (A*MIDEX). We warmly thank Pierre Martin and Agathe Mestrallet for stimulus preparation and data collection.

References

Baart, M. (2016). Quantifying lip-read-induced suppression and facilitation of the auditory N1 and P2 reveals peak enhancements and delays. *Psychophysiology*, *53*(9), 1295–1306. https://doi.org/10.1111/psyp.12683

Baart, M., Stekelenburg, J. J., & Vroomen, J. (2014). Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia*, *53*(1), 115–121. https://doi.org/10.1016/j.neuropsychologia.2013.11.011

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*(8), 2225–2234. https://doi.org/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*(3). https://doi.org/10.1007/s10339-004-0026-y

Boersma, P., & Weenink, D. (2013). *Praat: doing phonetics by computer* (5.3.42).

Bristow, D., Dehaene-Lambertz, G., Mattout, J., Soares, C., Gliga, T., Baillet, S., & Mangin, J.-F.

(2009). Hearing faces: How the infant brain matches the face it sees with the speech it hears. *Journal of Cognitive Neuroscience*, *21*(5), 905–921. https://doi.org/10.1162/jocn.2009.21076

Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology*, *45*(4), 204–220. https://doi.org/10.1002/dev.20032

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, *5*(7), e1000436. https://doi.org/10.1371/journal.pcbi.1000436

Chen, W.-F., Chao, P.-C., Chang, Y.-N., Hsu, C.-H., & Lee, C.-Y. (2016). Effects of orthographic consistency and homophone density on Chinese spoken word recognition. *Brain and Language*, *157–158*, 51–62.

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

Corbetta, M., & Shulman, G. . (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*, 201–215.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21.

Dijkstra, T., Roelofs, A., & Fieuws, S. (1995). Orthographic effects on phoneme monitoring. *Canadian Journal of Experimental Psychology*, *49*(2), 264–271.

Dodd, Barbara. (1979). Lip reading in infants: Attention to speech presented in- and out-of-synchrony. *Cognitive Psychology*, *11*(4), 478–484. https://doi.org/10.1016/0010-0285(79)90021-5

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech and*

Hearing Research, *11*, 796–804.

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*(8), 1207–1223. https://doi.org/10.1080/01690965.2012.701758

Friedrich, C. K., Felder, V., Lahiri, A., & Eulitz, C. (2013). Activation of words with phonological overlap. *Frontiers in Psychology, 4*(August), 1–11. https://doi.org/10.3389/fpsyg.2013.00556

Froyen, D., Van Atteveldt, N., Bonte, M., & Blomert, L. (2008). Cross-modal enhancement of the MMN to speech-sounds indicates early and automatic integration of letters and speech-sounds. *Neuroscience Letters*, *430*(1), 23–28.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*(2), 95–112.

Holcomb, P. J., Anderson, J., & Grainger, J. (2005). An electrophysiological study of cross-modal repetition priming. *Psychophysiology, 42*(5), 493–507. https://doi.org/doi:10.1111/j.1469-8986.2005.00348.x

Kiyonaga, K., Grainger, J., Midgley, K., & Holcomb, P. J. (2007). Masked cross-Modal repetition priming: An event-related potential investigation. *Language and Cognitive Processes*, *22*(3), 337–376.

Klucharev, V., Möttönen, R., & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Cognitive Brain Research*, *18*(1), 65–75. https://doi.org/10.1016/j.cogbrainres.2003.09.004

Lafontaine, H., Chetail, F., Colin, C., Kolinsky, R., & Pattamadilok, C. (2012). Role and activation time course of phonological and orthographic information during phoneme judgments. *Neuropsychologia*, *50*(12), 2897–2906.

Lewkowicz, D. J., & Flom, R. (2014). The audiovisual temporal binding window narrows in early childhood. *Child Development, 85*(2), 685–694. https://doi.org/10.1111/cdev.12142

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

Meltzoff, A. N., & Moore, M. K. (1977). Imitation of Facial and Manual Gestures by Human Neonates. In *Science* (Vol. 198, Issue 4312, pp. 75–78). https://doi.org/10.1126/science.198.4312.75

Meredith, M. A. (2002). On the neuronal basis for multisensory convergence: A brief overview. *Cognitive Brain Research*, *14*(1), 31–40. https://doi.org/10.1016/S0926-6410(02)00059-9

Mittag, M., Takegata, R., & Kujala, T. (2011). The effects of visual material and temporal synchrony on the processing of letters and speech sounds. *Experimental Brain Research*, *211*, 287–298.

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., & Foxe, J. J. (2002). Multisensory auditory-visual interactions during early sensory processing in humans: A high-density electrical mapping study. *Cognitive Brain Research*, *14*(1), 115–128. https://doi.org/10.1016/S0926-6410(02)00066-6

Muneaux, M., & Ziegler, J. C. (2004). Locus of orthographic effects in spoken word recognition: Novel insights from the neighbour generation task. *Language and Cognitive Processes*, *19*(5), 641–660.

Näätänen, R., & Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: A review and an analysis of the component structure. *Psychophysiology*, *24*(4), 375–425.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods*, *36*(3), 516–524.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9,* 97–1113.

Ostrand, R., Blumstein, S. E., Ferreira, V. S., & Morgan, J. L. (2016). What You See Isn't Always What You Get: Auditory Word Signals Trump Consciously Perceived Words in Lexical

Access. *Cognition*, *151*, 96–107.

Pattamadilok, C, Morais, J., Colin, C., & Kolinsky, R. (2014). Unattentive speech processing is influenced by orthographic knowledge: Evidence from mismatch negativity. *Brain and Language*, *137*, 103–111.

Pattamadilok, C, Perre, L., Dufau, S., & Ziegler, J. C. (2009). On-line orthographic influences on spoken language in a semantic task. *Journal of Cognitive Neuroscience*, *21*(1), 169–179.

Pattamadilok, Chotiga, Welby, P., & Tyler, M. D. (n.d.). The contribution of visual articulatory gestures and orthography to speech processing: Evidence from novel word learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Perre, L., Bertrand, D., & Ziegler, J. C. (2011). Literacy affects spoken language in a non-linguistic task: An ERP study. *Frontiers in Psychology*, *2*, 1–8. https://doi.org/10.3389/fpsyg.2011.00274

Perre, L., & Ziegler, J. C. (2008). On-line activation of orthography in spoken word recognition. *Brain Research*, *1188*, 132–138.

Pinto, S., Tremblay, P., Basirat, A., & Sato, M. (2019). The impact of when, what and how predictions on auditory speech perception. *Experimental Brain Research*, *237*(12), 3143–3153. https://doi.org/10.1007/s00221-019-05661-5

Price, C. J. (2000). The anatomy of language: contributions from functional neuroimaging. *Journal of Anatomy*, *197 Pt 3*, 335–359. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1468137&tool=pmcentrez&rendertype=abstract

Pugh, K. R., Mencl, W. E., Shaywitz, B. A., Shaywitz, S. E., Fulbright, R. K., Constable, R. T., Skudlarski, P., Marchione, K. E., Jenner, A. R., & Fletcher, J. M. (2000). The angular gyrus in developmental dyslexia: task-specific differences in functional connectivity within posterior cortex. *Psychological Science*, *11*(1), 51–56.

Saint-Amour, D., De Sanctis, P., Molholm, S., Ritter, W., & Foxe, J. J. (2007). Seeing voices: High-density electrical mapping and source- analysis of the multisensory mismatch negativity evoked during the McGurk illusion. *Neuropsychologia*, *45*(3), 587–597.

Sakai, K., & Passingham, R. E. (2003). Prefrontal interactions reflect future task operations. *Nature Neuroscience*, *6*(1), 75–81.

Schild, U., Röder, B., & Friedrich, C. K. (2011). Learning to read shapes the activation of neural lexical representations in the speech recognition pathway. *Developmental Cognitive Neuroscience*, *1*(2), 163–174.

Schlaggar, B. L., & McCandliss, B. D. (2007). Development of neural systems for reading. *Annu. Rev. Neurosci.*, *30*, 475–503.

Seidenberg, M. S., & Tanenhaus, M. K. (1979). Orthographic effects on rhyme monitoring. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(6), 546–554.

Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory-visual speech perception. *Developmental Science*, *11*(2), 306–320. https://doi.org/10.1111/j.1467-7687.2008.00677.x

Slowiaczek, L. M., Soltano, E. G., Wieting, S. J., & Bishop, K. L. (2003). An investigation of phonology and orthography in spoken-word recognition. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *56*(2), 233–262. https://doi.org/10.1080/02724980244000323

Stein, B. E., & Stanford, T. R. (2008). Multisensory integration: Current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, *9*(4), 255–266. https://doi.org/10.1038/nrn2331

Stekelenburg, J. J., Keetels, M., & Vroomen, J. (2018). Multisensory integration of speech sounds with letters vs. visual speech: only visual speech induces the mismatch negativity. *European Journal of Neuroscience*, *47*(9), 1135–1145. https://doi.org/10.1111/ejn.13908

Stekelenburg, J. J., & Vroomen, J. (2007). Neural Correlates of Multisensory Integration of Ecologically Valid Audiovisual Events. *Journal of Cognitive Neuroscience*, *19*(12), 1964–1973. https://doi.org/10.1162/jocn.2007.19.12.1964

Stekelenburg, J. J., & Vroomen, J. (2012). Electrophysiological evidence for a multisensory speech-specific mode of perception. *Neuropsychologia*, *50*(7), 1425–1431. https://doi.org/10.1016/j.neuropsychologia.2012.02.027

Summerfield, Q. . (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing By Eye: The Psychology of Lip-reading* (pp. 3–51). Lawrence Erlbaum Associates Ltd.

Taft, M. (2006). Orthographically influenced abstract phonological representation: Evidence from non-rhotic speakers. *Journal of Psycholinguistic Research*, *35*(1), 67–78.

Tye-Murray, N., Sommers, M., & Spehar, B. (2007). Auditory and Visual Lexical Neighborhoods in Audiovisual Speech Perception. *Trends in Amplification*, *11*(4), 233–241. https://doi.org/10.1177/1084713807307409

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(4), 1181.

Ventura, P., Morais, J., Pattamadilok, C., & Kolinsky, R. (2004). The locus of the orthographic consistency effect in auditory word recognition. *Language and Cognitive Processes*, *19*(1), 57–95.

Vroomen, J., & Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *Journal of Cognitive Neuroscience*, *22*(7), 1583–1596. https://doi.org/10.1162/jocn.2009.21308

Xu, W., Kolozsvári, O. B., Oostenveld, R., Leppänen, P. H. T., & Hämäläinen, J. A. (2019). Audiovisual processing of Chinese characters elicits suppression and congruency effects in

MEG. *Frontiers in Human Neuroscience*, *13*(February), 1–12. https://doi.org/10.3389/fnhum.2019.00018