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
Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France

ARTICLE INFO

Keywords:
Audiovisual integration
Supra-additivity
Speech processing
Articulatory gestures
Orthography

ABSTRACT

Both visual articulatory gestures and orthography provide information on the phonological content of speech. This EEG study investigated the integration between speech and these two visual inputs. A comparison of skilled readers’ brain responses elicited by a spoken word presented alone versus synchronously with a static image of a viseme or a grapheme of the spoken word’s onset showed that while neither visual input induced audiovisual integration on 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 in P350 time-window and generalized to all electrodes. The finding suggests a strong impact of spelling knowledge on phonetic processing and lexical access. It also indirectly indicates that the dynamic and predictive value present in natural lip movements but not in static visemes is particularly critical to the contribution of visual articulatory gestures to speech processing.

1. Introduction

Natural environment is filled with information from multiple sensory sources. Inputs from different modalities generally provide complementary or redundant information that can be useful in various situations. For instance, in noisy environments where the quality of speech signal is degraded, being able to see speakers’ lip movements improves speech comprehension (Sumby & Pollack, 1954). Likewise, visual deficit or deprivation makes individuals more dependent on, or become more efficient in processing auditory information (Landry et al., 2013). The ability to detect the relationship between different sensory inputs and to integrate them into a coherent percept is fundamental both to basic operations, such as orienting in space, perceiving motions, localizing objects in the environment, as well as to more complex activities such as learning to read or playing musical instruments (Amedi et al., 2005; Calvert et al., 1998).

The present study focuses on the contribution of auditory and visual information in language processing. 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; 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).

* Corresponding author at: Laboratoire Parole et Langage, Centre National de la Recherche Scientifique (UMR 7309), 5, Av. Pasteur, 13100 Aix-en-Provence, France.
E-mail address: chotiga.pattamadilok@lpl-aix.fr (C. Pattamadilok).

https://doi.org/10.1016/j.bandl.2021.105058
Received 30 July 2021; Received in revised form 31 October 2021; Accepted 8 December 2021
0093-934X/© 2021 Elsevier Inc. All rights reserved.
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 orthography has mainly been studied in laboratory settings under synchronous presentation of the auditory speech signal and a static written script, 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. Notably, when considering the temporal AV relationships in visual speech, there is a robust correlation in time between the variations of mouth opening and 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 texts induced illusory changes in the perception of ambiguous syllables (halfway between /aha/ and /aha/) 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 written syllables. 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). 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 et al. (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). A 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 affected 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).

In most studies that investigated the impact of articulatory gestures on speech processing per se (Arnal et al., 2009; Besle, Fort, Delpuech, et al., 2004; Klucharev et al., 2003; Stekelenburg & Vroomen, 2007; Van Wassenhove et al., 2005; Vroomen & Stekelenburg, 2010) as well as in those that compared the impact of articulatory gestures to that of orthography mentioned above (Pinto et al., 2019; Stekelenburg et al., 2018), the articulatory gestures were systematically presented in their natural context, that is, with dynamic lip movements. Nevertheless, due to the differences in the characteristics of the two types of visual cues, directly contrasting their contributions to speech processing remains difficult. Given this constraint, the present study proposed a more strictly controlled experimental protocol 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 dimensions that are specific to articulatory movements were removed by using a static image of lip shape representing the viseme of each spoken word’s initial phoneme instead of dynamic articulatory gestures producing the entire word. Likewise, the orthographic cue corresponded to the first grapheme of each spoken word. To our knowledge, a few former behavioral studies have attempted to dissociate the impact of kinematic facial information and that of static facial features of articulatory gestures on speech processing. For instance, Rosenblum and Saldana (1996) compared the impact of natural visual speech that contained both kinematic and featural dimensions to visual inputs that contained either the kinematic or featural dimension alone. The authors reported that while both kinematic and featural inputs presented alone could affect speech perception, their impacts were much smaller than that of natural visual speech that contained both dimensions. Interestingly, the authors also reported that only the impact of the kinematic input was perceptual and reflected a true integration with auditory speech. On the contrary, the impact of the static features was more related to post-perceptual stages. A dissociation between the two dimensions of visual speech was also reported in a brain-lesion patient who had deficits in matching static images of speech articulations to speech sounds but could nevertheless lipread normally when presented with dynamic articulations (Campbell, 1992). Interestingly, these findings mirror the dissociation between the impact of dynamic articulatory gestures and that of static orthogonal input reported by Stekelenburg et al. (2018). They also further suggest that the contributions of static facial features (visemes) and of static orthographic input to speech processing might share some common properties.

The main aim of the present investigation was to examine whether these two kinds of static visual inputs are similarly integrated with speech signal when presented synchronously and, if so, whether such integration extended beyond the initial stage of acoustic–phonetic analysis typically reported in the AV integration literature. To address this issue, we conducted an EEG study in which brain responses to spoken words and the two kinds of static 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 the 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 perceptual components that are classically examined in the AV integration literature, a few studies also extended their investigations to linguistic processing, for instance, by examining whether the presence of articulatory gestures would affect the different phases of the lexico-semantic processing (Basirat et al., 2018; Fort et al., 2010, 2013; Ostrand et al., 2011, 2016). So far, this question remains under debate. Although some researchers argued for a separation between perceptual and linguistic processing, and considered that the contribution of visual speech is restricted to perceptual processing (Baart & Samuel, 2015; Samuel & Lieblich, 2014), others reported evidence suggesting that visual speech might as well play a role during
linguistic encoding of speech (Bassirat et al., 2018; Fort et al., 2012, 2013; see also Doris et al., 2021 for further discussion on the topic). Unlike articulatory gestures, the influence of orthography on lexical access of spoken words has been more consistently reported, both when the manipulation of the relationship between orthography and speech sounds relied on participants’ abstract knowledge of spoken words’ spellings (Chen et al., 2016; Pattamadilok et al., 2009; Perre et al., 2009) and when both written and spoken inputs were actually presented (Grainger et al., 2003; Kiyonaga et al., 2007; Veivo & Järvinen, 2013 but see López Zunini et al., 2020). In the present study, we compared the impact of the articulatory cue (viseme) to that of the orthographic cue (grapheme) on the lexical access of spoken words by examining the modulation of the P350 component. This component reflects the initial activation of a cohort of words in the mental lexicon that overlap with the information present in the audio (and/or visual) sensory inputs (Friedrich et al., 2013; Schild et al., 2011). It is therefore particularly well-suited for the current protocol where the presence of the viseme or the grapheme of spoken word’s onset could activate the initial phoneme of the word and, thus, influence the contact with the lexicon. To our best knowledge, the relative contributions of these two types of visual cues to the activation of a cohort of spoken words has never been compared. As a methodological note, applying the additive model to examine the AV integration during a post-perceptual process might 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. 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 and 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.

1 Given that the visual cues used in the present study involved only the initial phoneme of the spoken inputs, it is unlikely that these cues would have an impact on the N400 component that reflects whole word and semantic processing.

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/, /a/, /pa/, /ta/ and /3a/ syllables were individually recorded by a male native speaker of French, 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 × 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 the 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 Fig. 1.

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 /3/ 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 (Fisher, 1968; Summerfield, 1987). 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 neighborhood, orthographic neighborhood, phonological uniqueness point and orthographic uniqueness point (all ps > 0.15; 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 /3/ 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 the same speaker as in

2.2.3. Analysis


the viseme recognition task 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: 77 dB ± 2).

In addition to the auditory stimuli, two kinds of visual stimuli were used. The first kind (hereafter, V\text{GRAPHEME}) was the static lip shape images corresponding to /\text{s}/, /\text{g}/, /\text{p}/, /\text{n}/, /\text{a}/, /\text{i}/ and /\text{o}/ visemes (see the stimulus description in the viseme recognition task). The second kind was the following graphemes: <\text{f}>, <\text{g}>, <\text{p}>, <\text{t}>, <\text{a}>, <\text{i}>, and <\text{o}>(hereafter, V\text{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 (AV\text{VISEME/CONGRUENT}), auditory word associated with the grapheme of the word’s initial phoneme (AV\text{GRAPHEME/CONGRUENT}), auditory word associated with a viseme that does not correspond to the word’s initial phoneme (AV\text{VISEME/INCONGRUENT}), auditory word associated with a grapheme that does not correspond to the word’s initial phoneme (AV\text{GRAPHEME/INCONGRUENT}), viseme without auditory input (V\text{VISEME}), and grapheme without auditory input (V\text{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. Within each trial, a viseme was simultaneously presented on the screen with its corresponding grapheme and phoneme for 1.5 s. Participants were explicitly required to focus their attention on the speech sound that is associated with the viseme. They were instructed not to memorize the association but to mimic the sound based on the image of the viseme. The association between the three elements was presented ten times for each viseme used in the main tasks: /\text{a}/, /\text{f}/, /\text{o}/, /\text{g}/, /\text{p}/, and /\text{i}/. The presentation order of these 70 trials was randomized.

Following the exposure phase, the participants completed a viseme recognition task with corrective feedback. Within each trial, a viseme from Fig. 1 was presented on the screen with seven letters: a, i, o, f, g, p, t. The participants had to decide what sound was being produced. To perform the task, they were again encouraged to mimic the sound that corresponded to the image of the viseme and then respond by clicking on the letter that represented the sound within 4 s. Note that, in the present protocol, there was a one-to-one correspondence between the speech sounds and the letters. 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 associated speech sound and letter were simultaneously 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, AV\text{VISEME/CONGRUENT}, AV\text{GRAPHEME/CONGRUENT}, AV\text{VISEME/INCONGRUENT}, AV\text{GRAPHEME/INCONGRUENT}, V\text{VISEME}, 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\text{VISEME} and V\text{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\text{VISEME} and V\text{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 /\text{s}/ consonant. In 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 all the time. The two tasks were of equal duration, each lasting about 20 min. Their order was counterbalanced across participants, with a short break offered between them. At the beginning of each task, a short training was performed.

Fig. 1. The visemes and their corresponding phonemes.
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 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% of the trials in the phonemic decision task (AUD: 16%, AV_VISEME/CONGRUENT: 15%, AV_VISEME/INCONGRUENT: 18%, AV_GRAPHHEME/INCONGRUENT: 17%, VISEME: 16%, GRAPHHEME: 16%), and 16% of the trials in the lexical decision tasks (AUD: 14%, AV_VISEME/CONGRUENT: 16%, AV_GRAPHHEME/CONGRUENT: 18%, AV_VISEME/INCONGRUENT: 16%, AV_GRAPHHEME/INCONGRUENT: 14%, VISEME: 15%, GRAPHHEME: 17%). No significant difference was observed between the tasks [F(1, 19) = 0.55, p = 0.47, p(Δ) = 0.03], the experimental conditions [F(6, 114) = 1.35, p = 0.24, p(Δ) = 0.07], without an interaction between the two factors [F(6, 114) = 1.47, p = 0.20, p(Δ) = 0.07].

For each participant and each task, we used an additive model to test the AV integration (Baart, 2016), 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 VISEME and GRAPHHEME conditions were subtracted from those obtained in the corresponding AV conditions in the following manners: AV_GRAPHHEME/CONGRUENT − VISEME, AV_GRAPHHEME/INCONGRUENT − VISEME, AV_VISEME/CONGRUENT − VISEME, AV_VISEME/INCONGRUENT − VISEME. Each of the resulting ERP difference waves (hereafter, difERP) was compared against the signal obtained in the AUD condition, based on the assumption that AV integration occurs whenever the difERP signals were different from the signal obtained in the AUD condition (AV−V ≠ 0) 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_GRAPHHEME/CONGRUENT − VISEME difERP signal to the AV_VISEME/CONGRUENT − VISEME difERP signal. Finally, the impact of AV congruency for each type of visual input was obtained by comparing the AV_GRAPHHEME/CONGRUENT − VISEME difERP signal, and AV_VISEME/CONGRUENT − VISEME to the 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, FC2, FCz, C1, Cz, C2), left anterior (Fp1, AF3, AF7, F7, F5, F3, FC3, FC5, FT7), left posterior (CP3, CP5, CP7, P3, P5, P7, P03, P07, 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 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 negative and positive peaks corresponded respectively to the lowest and highest individual peaks detected within the time-windows of interest. 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. Fig. 3 (and Appendix 2) showed the waveforms of the ERPs 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

2 The 1Hz high pass filter used here might carry the risk of distorting the data (Tanner et al., 2015), we decided to apply a two-way least-square FIR filtering of 1–30 Hz for two main reasons. First, most of previous EEG studies that examined audiovisual speech integration on N1/P2 AEPs did use a similar filtering (e.g., Besle, Fort, Delpuech, et al., 2004; 1–30Hz; Ganesh et al., 2014: 2–20Hz; Klucharev et al., 2003: 1–25Hz; Pilling, 2009: 1–30Hz; Pinto et al., 2019: 3–30Hz; Stekelenburg & Vroomen, 2007: 0.5–30Hz; Treille et al., 2014: 1–20Hz; Tremblay et al., 2021: 3–30Hz; Van Wassenhove et al., 2005: 1–55Hz; Vroomen & Stekelenburg, 2010: 0.5–30Hz; Winneke & Phillips, 2011: 1–30Hz). Second, in both phonemic and lexical decision tasks, the go/no-go paradigm might have induced task-related neural activity common to all modalities but unrelated to cross-modal integration, such as slow anticipatory potentials that precede perceptual decisions and discriminative responses, characterized by a slow positive deflection on fronto-central sites (Teder-Saljérvi et al., 2002). Since 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, a 1Hz high-pass filter was applied on the EEG data to minimize the contribution of slow anticipatory potentials.

3 Since the existing literature does not provide an a priori hypothesis regarding the scalp location 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.
indicating that the visual saliency of the articulatory gestures varies from one viseme to another (Fisher, 1968; Summerfield, 1987), the analysis showed significant differences across visemes [F(6, 114) = 4.65, p = .0003, ps = 0.20]. This was due to a lower accuracy score obtained on the viseme of the /t/ consonant (81%) compared to the other visemes (/a/ = 96%, /i/ = 95%, /s/ = 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 < .0001, ps = 0.53: /a/ = 1205 ms, /t/ = 1587 ms, /y/ = 1426 ms, /i/ = 1530 ms, /o/ = 1203 ms, /p/ = 1298 ms, /s/ = 1863 ms]. The mean RT obtained on the viseme of /t/ was significantly longer than those obtained on the other visemes, ps ≤ 0.005. The mean RTs on the visemes of /t/ and /i/ were significantly longer than those obtained on the visemes of /a/, /o/ and /p/, ps < 0.05 (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, AVVIKEMBERG, AVVIKEMBERG, AVVIKEMBERG, AVAVVIKEMBERG, AVAVVIKEMBERG, AVAVVIKEMBERG) were treated as within-subject factors.

The analysis performed on the percentage of correct responses showed significant main effects of task [F(1, 19) = 7.33, p = 0.013, ps = 0.28] and condition [F(4, 76) = 4.33, p = 0.003, ps = 0.19]. The interaction between the two factors was not significant [F(4, 76) = 1.72, p = 0.155, ps = 0.08]. As illustrated in Fig. 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 AVVIKEMBERG condition compared to the other conditions (ps < 0.01).

The analysis of the RTs showed significant effects of task [F(1, 19) = 119.73, p < .0001, ps = 0.86], condition [F(4, 76) = 7.85, p < .0001, ps = 0.29] and their interaction [F(4, 76) = 9.06, p < .0001, ps = 0.32]. As illustrated in Fig. 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, ps = 0.53] where the mean RT obtained in the AVVIKEMBERG condition was longer than those observed in the other conditions and the mean RT obtained in the AVAVVIKEMBERG condition was shorter than those observed in the other conditions (ps < 0.001).

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

#### 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, ps = 0.41]. A similar finding was obtained in the analysis conducted on peak latency [F(2.932, 55.708) = 4.22, p = 0.18]. For both dependent variables, no other main effects or interactions were significant (ps > 0.20), which clearly suggests the 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, ps = 0.72]. Interestingly, the effect of condition [F(2.391, 45.435) = 5.19, p = 0.006, ps = 0.21] and its interaction with cluster were also significant [F(7.327, 139.212) = 2.08, p = 0.047, ps = 0.10]. 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, 7.017) = 4.78, p = 0.011, ps = 0.39]. Further analyses showed that the effect of condition was significant for both clusters (ps < 0.01). No other main effects or interactions were significant (ps > 0.20).
3.2.3. P350: 300–400 ms

ANOVA performed on P350 peak amplitude showed a significant main effect of cluster \( \left[ F(2.387, 45.356) = 4.33, p = .014, \eta^2_p = 0.19 \right] \). The effect of condition was significant \( \left[ F(2.563, 48.703) = 6.04, p = .002, \eta^2_p = 0.24 \right] \) but did not interact with the other main factors. As illustrated in Fig. 5, the condition effect reflected an overall enhancement of P350 peak amplitudes observed in all the AV conditions \((\text{difERPs})\) 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 \(\text{AV}_{\text{GRAPHEME/CONGRUENT}} - \text{V}_{\text{GRAPHEME}} \) vs. AUD contrast and \(\text{AV}_{\text{GRAPHEME/INCONGRUENT}} - \text{V}_{\text{GRAPHEME}} \) vs. AUD contrast, respectively) than when they were visemes \((p = .04\) and \(p = 0.015\) for the \(\text{AV}_{\text{VISME/CONGRUENT}} - \text{V}_{\text{VISME}} \) vs. AUD and \(\text{AV}_{\text{VISME/INCONGRUENT}} - \text{V}_{\text{VISME}} \) vs. AUD contrast, respectively). For either type of visual cues, the congruency between the auditory and visual inputs did not affect the degree of AV integration \((p > 0.50)\). The direct comparison of the impacts of viseme and grapheme on the degree of audiovisual integration conducted on congruent trials showed a marginally stronger integration between the auditory input and graphemic cues \((p = .08)\). However, this advantage of the graphemic cues was significant when the analysis was conducted across congruency levels \((p = .034)\). No significant difference was found in the ANOVA conducted on P350 peak latency \((p > 0.10)\).

4. Discussion

The result obtained in the viseme recognition task showed that the participants were able to match the visemes with their corresponding phonemes at a good level of performance only after a short
familiarization phase (mean %ACC = 92.8%, range = 81%-96%; mean RT = 1444 ms, range = 1203–1863 ms). We also observed a variation in the performance across visemes, which clearly reflects the typical pattern of viseme recognition performance previously reported in the literature, with some visemes being intrinsically harder to map to speech sounds than others, even when the natural dynamic gestures were used (Fisher, 1968; Summerfield, 1987). The observation of such variation suggested that the participants had followed our instructions to mimic the speech sounds based on the images of visemes rather than to memorize the associations between the speech sounds and the images of visemes in the way they would have done if they had had to learn novel associations between speech sounds and arbitrary symbols.

The analyses of the behavioral data obtained in the main tasks showed that presenting a viseme in synchrony with a spoken word did not have a significant impact on the performance either in the phonemic decision task or in the lexical decision task. Only the presence of graphemes 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 cues 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 our manipulation of the relationship between speech sounds and visual inputs 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 inputs was observed on the ERP components of interest. Finally, the same pattern of AV integration was observed in both phonemic and lexical decision tasks, which suggests that the AV integration reported here occurred at the processing stages that were common to both tasks and it was independent of the top-down influence of task-demands and participants’ attention to the initial phoneme vs. the entire stimulus (but see Lopez-Barroso et al., 2020 for a different interpretation of the locus of AV integration). 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 (Naatanen & 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 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 texts) 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.,...
The absence of evidence for AV integration on N1 reported here is coherent with the fact that in our protocol no visual predictive cues, which have been argued to be critical for the modulation of N1, were available before the acoustic onset.

4.2. Impact of AV integration on P2: A combination of a non-specific AV integration and a specific increase of sensitivity to 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. 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 at electrode clusters. Importantly, the supra-additive AV integration effect observed on P2 appears in the opposite direction of what was usually found in previous EEG studies on AV speech integration (i.e., sub-additive AV integration effect, AV − V < A). As discussed above, this is likely due to the absence of a predictive visual cue before the acoustic onset. In line with this hypothesis, Molholm et al. (2002) similarly observed a supra-additive AV integration on P2 at fronto-central electrodes, when visual (a disk) and auditory (a beep) stimuli were presented alone and simultaneously without any predictive visual cues. Likewise, Vroomen and Stekelenburg (2010) observed a supra-additive effect on P2, although non-significant, when an auditory stimulus (a pure tone) was presented together with a non-predictive visual stimulus (a rectangle).

Interestingly, we observed two distinct patterns of AV integration at 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 and where the P2 amplitude is highest, AV integration occurred for both types of visual cues, independently of task demands or congruency between the auditory and visual inputs. 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 at 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. To our knowledge, previous studies investigating the impact of AV integration on the N1 and P2 components generally focused on central or fronto-central electrodes and did not report whether the integration was also found on other sites. Despite the fact that the advantage of speech sound-grapheme integration was observed in the P200 time-window, the left anterior region is not the typical scalp location of this component, and it is not recommended to directly infer the cortical source(s) of brain responses recorded from surface electrodes. However, at the very least, 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 and the current observation of a stronger AV integration in the grapheme condition, Xu et al. (2019) reported MEG evidence for an integration between speech sounds and visemes, with the highest amplitude of late components (P3) when a graphemic cue was presented. Additionally, the authors noted that the integration that took place around 205–365 ms. (which corresponds to the P2 and P350 time-windows reported here) was located in the left anterior and supra-marginal gyri, thought as heteromodal areas involved in linking orthographic representations encoded in the occipital lobe to phonological representations encoded in the superior temporal gyrus (Price, 2000; Pugh et al., 2000; Schlaggar & McCandliss, 2007). This MEG finding provides an interesting direction for future examination of the underlying mechanism of the AV integration effect reported here.

Two plausible complementary cognitive processes leading to the advantage of speech sound-grapheme association could be advanced at this stage of research. The first explanation is related to the nature of the mapping between speech sounds and graphemes, which is different from the nature of the mapping between speech sounds and visemes. Graphemes, which are abstract symbols, provide a direct and unambiguous access to abstract phonemic representations, especially in skilled readers. In contrast, given that visemes indicate how speech sounds are produced, the retrieval of abstract phonemic representations from visemes is likely operated through the motor system, even when the visemes are static. Therefore, the access to these abstract representations is less direct and can be ambiguous in some cases (e.g., /l/ and /i/). As mentioned earlier, the variation of viseme recognition performance is not specific to the static images used here but it is also observed with natural dynamic articulatory gestures (Fisher, 1968; Summerfield, 1987). Even though the absence of lip movements certainly increased the difficulty in mapping visemes to speech sounds, it is worth noting that the degree of visual saliency itself (i.e., the facility to identify speech sounds based on the visual information from articulatory gestures) does not seem to affect the amplitude of the P2 component. Indeed, a previous finding by Van Wassenhove et al. (2005), who examined ERP responses to AV integration of single syllables (ipa, ta, ka), showed that the visual saliency of dynamic articulatory gestures only affected the latency, but not the amplitude, of the N1 and P2 component measured in the audiovisual condition. Although the by-item analysis could not be conducted in the present study, we could reasonably expect the same outcome for static visemes, i.e., no modulation of P2 amplitude according to the visual saliency of the viseme.

However, the assumption that the nature of the mapping between speech sounds and visual cues, which determines the automaticity of the access to abstract phonemic representations, is the only critical factor may seem incompatible with the absence of the congruency effect on the ERP data. In the incongruent condition, neither type of visual cues provided a valid phonemic content of speech sounds. If this information was the only critical factor for the AV integration process, one would also expect the degree of AV integration to depend on the congruency between the auditory and visual inputs. The absence of the congruency effect is inconsistent with several existing findings (e.g., Caffarra et al., 2021; Karipidis et al., 2018; Raji et al., 2000; Xu et al., 2019). Nevertheless, a close look at the literature revealed that most studies that reported the congruency effect used experimental designs that were optimal to reveal the impact of AV (in)congruency. For instance, the auditory and visual inputs were either fully (or almost) matched or mismatched, both auditory and visual inputs were task-relevant, or a delay between the auditory and visual inputs was added to increase the possibility to observe an early audiovisual congruency effect (cf. Froyen et al., 2009). These conditions were not met in our protocol where 1) the overlap (or the mismatch) between the auditory and visual inputs involved only the first phoneme of disyllabic words, 2) even though the visual cues were presented, the tasks relied on the auditory inputs alone, and 3) there was no delay between the onset of the auditory and the visual inputs. The combination of these factors might drastically reduce the impact of audiovisual congruency on the integration process compared to what has been reported in some previous studies. Moreover, 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 cues 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 (as well as at the subsequent processing stage discussed below) could be supported by such neural adaptation mechanism.

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

Interestingly, the enhanced sensitivity to graphemes that was restricted to the left anterior cluster 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 the top-down influence of task-demands and 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 present 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’, ‘è’, ‘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), 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., 2021, for a similar conclusion on a weaker 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 verbal cues, i.e., 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. This is likely due to the absence of the predictive value of the visual cues used here. 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 (Raart et al., 2014), AV integration process itself also becomes more sensitive to the link between speech sounds and the abstract orthographic code than 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. The stronger degree of 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 allows a fast and direct access to abstract phonological representations. The retrieval of these representations from visemes, which is operated through the motor system, seems to be less straightforward. Finally, by using static visemes rather than dynamic articulatory gestures, our finding also suggests that a significant part of the advantage of visual speech compared to auditory-alone speech reported in the literature could be attributed to the dynamic and predictive cues present in natural lip movements. Once these cues are removed, the impact of articulatory gestures especially at the early perceptual stages of speech processing seems to be severely modulated.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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.

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)
Appendix B

Full descriptions of the cluster effects on peak amplitude and peak latency.

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

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

P350: /

Appendix C

Waveforms of the ERPs obtained in the phonemic decision and lexical decision tasks in the auditory condition and the following different ERPs (AV-V): AVVISEME/CONGRUENT – VVISEME; AVVISEME/INCONGRUENT – VVISEME; AVGRAPHEME/CONGRUENT – VGRAPHEME; AVGRAPHEME/INCONGRUENT – VGRAPHEME. 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.
References