The timing of visual speech modulates auditory neural processing

Marc Sato

THE TIMING OF VISUAL SPEECH MODULATES AUDITORY NEURAL PROCESSING

Marc Sato

Laboratoire Parole et Langage, Centre National de la Recherche Scientifique, Aix-Marseille Université, Aix-en-Provence, France

Correspondence can be addressed to Marc Sato, Laboratoire Parole et Langage, UMR 7309 CNRS & Aix-Marseille Université, 5 avenue Pasteur, 13100 Aix-en-Provence, France, or via e-mail: marc.sato@lpl-aix.fr.

CONFLICT OF INTEREST

The author declares no competing financial interests.

Number of pages: 16, Figures: 2

ABSTRACT

In face-to-face communication, visual information from a speaker's face and time-varying kinematics of articulatory movements have been shown to fine-tune auditory neural processing and improve speech recognition. To further determine whether the timing of visual gestures modulates auditory cortical processing, three sets of syllables only differing in the onset and duration of silent prephonatory movements, before the acoustic speech signal, were contrasted using EEG. Despite similar visual recognition rates, an increase in the amplitude of P2 auditory evoked responses was observed from the longest to the shortest movements. Taken together, these results clarify how audiovisual speech perception partly operates through visually-based predictions and related processing time, with acoustic-phonetic neural processing paralleling the timing of visual prephonatory gestures.

KEYWORDS

Audiovisual speech perception, Visual prediction, EEG.

INTRODUCTION

In face-to-face communication, visual information from a speaker's face improves auditory processing and speech recognition. The positions and dynamic patterning of visible vocal tract articulators enhance sensitivity to acoustic speech information by decreasing auditory detection threshold (Grant and Seitz, 2000; Schwartz et al., 2004), and improve auditory phoneme identification and word recognition, notably in case of a degraded acoustic signal (Sumby and Pollack, 1954; Benoît et al., 1994), of a second language (Navarra and Soto-Faraco, 2005), and in hearing-impaired listeners (Grant et al., 1998). These visual-to-auditory benefits stem from a high level of cross-predictability between the acoustic and visual speech signals due to their common underlying motor cause. There is a robust correlation in time between variations of mouth opening and variations of the acoustic envelope, which is maximum in regions of the acoustic spectrum corresponding to the first vocal tract resonances (Grant and Seitz, 2000; Chandrasekaran et al., 2009). In addition, at the beginning of a taking turn or in the case of an isolated syllable/word (i.e., preceded by silence), perceptually relevant prephonatory gestures precede the onset of the acoustic signal with a specific temporal asynchrony (Chandrasekaran et al., 2009; Schwartz and Savariaux, 2014). Taken together, the audiovisual cross-predictability supports the view that audiovisual speech perception partly operates through visually-based predictions based on their joint probability distribution from prior audiovisual speech experience (van Wassenhove, 2013; Rosenblum et al., 2016).

Capitalizing on the temporal precedence of visual articulatory movements on an isolated auditory syllable/word, electro- and magneto-encephalography (EEG/MEG) studies have consistently reported that prephonatory visual movements, before the acoustic speech onset, modulate auditory neural processing early in the supratemporal plane of the auditory cortex. More specifically, the amplitude and latency of P1, N1 and/or P2 auditory evoked potentials (AEPs) were attenuated and speeded up during audiovisual compared to unimodal syllable perception (Lebib et al., 2003; Klucharev et al., 2003; Besle et al., 2004; van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007; Arnal et al., 2009; Winneke and Phillips, 2011; Treille et al. 2014a, 2014b, 2017, 2018; Pinto et al., 2019; Tremblay et al., 2021; Sato, 2022). The observed latency facilitation (time of neural processing) and amplitude suppression (size of neural population and activation synchrony during the component generation) of P1, N1 and P2 AEPs, classically involved in sensory gating, acoustic and phonetic decoding stages of auditory speech processing, are thought to reflect early multisensory integrative mechanisms through visual predictions of the incoming auditory speech events.

The above-mentioned studies argue for a key role for visual prediction in facilitating auditory neural processing. Importantly, the effect of visual prediction appears to depend on the degree of visual saliency, with the higher visual recognition, the stronger N1 and P2 latency facilitation (van Wassenhove et al., 2005; Arnal et al., 2009; but see Treille et al., 2004b). The goal of the present EEG study was to further examine the effect of visual predictions by determining the extent to which the timing of prephonatory gestures

modulates auditory neural processing. To this aim, participants were presented with auditory, visual, and audiovisual syllables. One third of the audiovisual syllables consisted of a visually neutral mouth position, followed by 240 ms of visual prephonatory movements before the acoustic speech onset. This prephonatory duration was in line with those previously observed for isolated syllables (Chandrasekaran et al., 2009; Schwartz and Savariaux, 2014). Two additional sets of audiovisual syllables were artificially built by either replacing the first 120 ms or 200 ms of visual prephonatory movements by a still image of the neutral mouth position (i.e., keeping the last 120 ms or 40 ms visual prephonatory movements before the acoustic speech onset). Given the distinct predictive power in the three sets of audiovisual stimuli, with an almost linear decrease of prephonatory duration (i.e., 240 ms vs. 120 ms vs. 40 ms), related differences in P1, N1 and/or P2 amplitudes and latencies would demonstrate that auditory neural processing relies on visual predictions not only as a function of visual saliency but also according to the duration of visual prephonatory movements. In addition to the EEG experiment, since audiovisual speech interaction has been shown to depend on the degree of visual recognition (van Wassenhove et al., 2005; Arnal et al., 2009), a behavioral lip-reading experiment was performed to disentangle the respective contribution of visual timing and visual recognition on auditory evoked responses.

METHODS

Participants

Twenty healthy adults (14 females and 6 males), with a mean age of 27 ±4 years (mean ± SD; range: 20-37 years), participated in the study after giving informed consent. All participants were native French speakers, with an average of 12 ±2 years of education (range: 7-17 years). They were all right-handed according to the standard handedness inventory (Oldfield, 1971) with a mean score of 78 ±14 % (range: 50-100 %), 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 and participants were compensated for the time spent in the study.

Stimuli

Multiple utterances of /pa/, /ta/, and /ka/ syllables, starting with a visually neutral open mouth, were individually recorded by four native French speakers (three females and one male) in a soundproof room. These three syllables all included an initial unvoiced stop consonant, allowing precise detection of the acoustic syllable onset for EEG analyses. In addition, they ensured a gradient of visuo-labial saliency (with the bilabial /p/ consonant known to be more visually salient than the alveolar /t/ and velar /k/ consonants; see Fisher, 1968; Summerfield, 1987). Video digitizing (centered on the speaker's mouth; see Figure 1) was done at 25 frames per second with a resolution of 720 × 576 pixels. Audio digitizing was done at 44.1 kHz with 16-bit quantization recording.

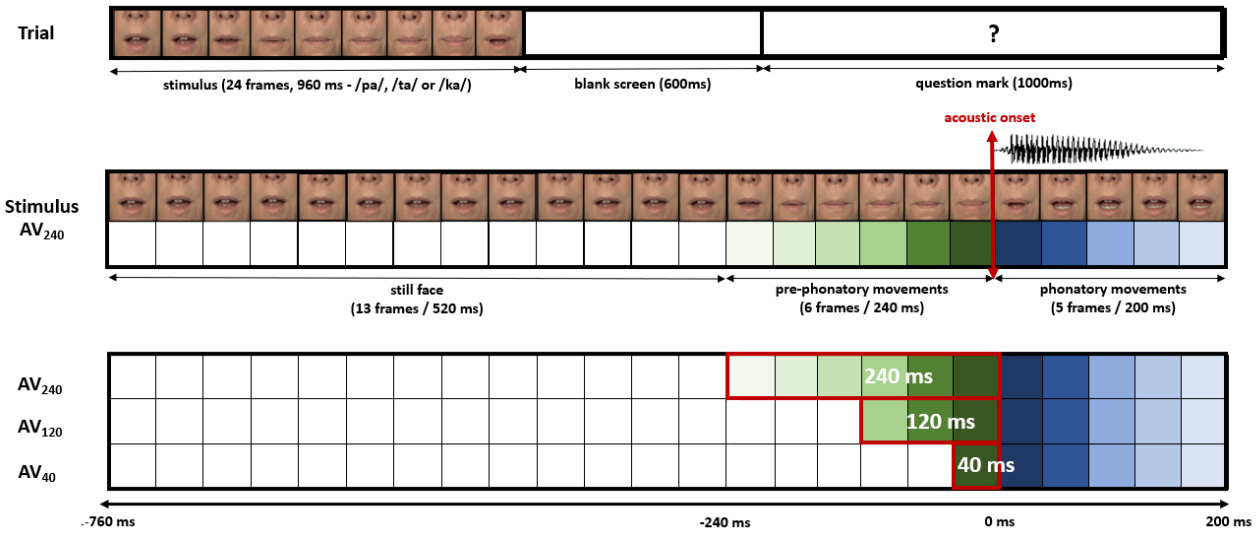

Figure 1. Experimental design. Each trial consisted of a stimulus (960 ms) followed by a blank screen (600 ms) and then by a question mark (1000ms), which served as cue for participants' responses to the syllable identification task. AV_{240} audiovisual stimulus consisted of a visually neutral open mouth (13 frame, 520 ms), followed by visual prephonatory–(6 frames, 240 ms) and phonatory movements (5 frames, 200 ms) before and after the acoustic consonantal burst. AV_{120} and AV_{40} audiovisual stimuli were built by replacing the first three or five frames of the prephonatory movements by a visually neutral open mouth.

Using Adobe Premiere (Adobe systems, San Jose, USA) and Praat (Boersma and Weenink, 2013), one set of clearly articulated /pa/, /ta/, and /ka/ tokens were selected and edited per speaker based on acoustic and visual properties. The editing procedure ensured that all twelve selected audiovisual stimuli (4 speakers x 3 syllables) started with a visually neutral open mouth (1 frame, 40 ms), followed by visual prephonatory (6 frames, 240 ms) and phonatory movements (5 frames, 200 ms) before and after the acoustic consonantal burst. For all stimuli, the first frame corresponding to the visually neutral mid-open mouth was replicated for a total of 13 still images (520 ms) before the prephonatory movements. With this editing procedure, all stimuli were 24 frames long (960 ms; see Figure 1). On average, the onset of the acoustic consonantal burst was of 755 ±9 ms (range: 740-775 ms), the onset of the peak intensity was of 821 ±43 ms (range: 773-896 ms) and the syllable duration was of 182 ±28 ms (range: 127-215 ms). The acoustic intensity was normalized using a common maximal amplitude criterion. For each of the twelve audiovisual stimulus (AV_{240}), auditory-only (A) and visual-only (V_{240}) stimuli were built by either replacing all visual frames by the first still image or by removing the acoustic signal. In addition, two additional audiovisual stimuli were built by replacing the first three or five frames of the prephonatory movements by the first still image (AV_{120}, AV_{40}).

In AV_{240} stimuli, the 240 ms prephonatory duration appeared in the range previously observed by Chandrasekaran et al. (2009; visual lead of 100 ms to 300 ms) and Schwartz and Savariaux (2014; visual lead of 200 ms to 400 ms). Although the extent and timing of lip closure and lip area naturally varied across /pa/, /ta/ and /ka/ syllables (see Schwartz and Savariaux, 2014), prephonatory movements roughly involved two successive temporal events: a closure phase characterized by the initiation of the closing gesture and a progressive decrease of lip area, and a stable phase with the lip area kept minimal. The initial closing phase in AV_{240} stimuli was absent in both AV_{120} and AV_{40} stimuli. For all stimuli, a release phase characterized by an opening onset coincided with the acoustic consonantal burst. In summary, AV_{240} stimuli differed from AV_{120} and AV_{40} stimuli by the presence of the closing phase while AV_{120} stimuli differed from AV_{40} stimuli by the duration of the stable phase (stimuli samples are provided in supplementary materials).

Experimental procedure

The experiment was carried out in a dimly lit sound-attenuated room. Participants sat in front of a computer monitor at approximately 50 cm. The acoustic signal was presented through two loudspeakers, located on each side of the computer monitor, at the same comfortable sound level for all participants. Stimuli were presented using Presentation software (Neurobehavioral Systems, Albany, USA), which was also used to record participants' behavioral responses and to synchronized EEG recordings.

Participants were asked to complete a forced-choice syllable identification task. On each trial, they identified one syllable (/pa/, /ta/ or /ka/) by pressing one of three keys on a keyboard with their right hand. No feedback was provided. The response key designation was counterbalanced across participants. To dissociate sensory/perceptual from motor responses on EEG recording, each stimulus (960 ms) was

followed by a blank screen (600 ms) and then by a question mark (1000ms), which served as cue for participants' responses (see Figure 1).

The experiment consisted of 6 blocks of 60 trials (4 speakers x 3 syllables x 5 conditions), each presented in a pseudo-randomized order (i.e., no more than two times the same syllable or the same experimental condition consecutively). In total, there were 72 trials in each of the 5 following experimental conditions: audiovisual (AV_{240}, AV_{120}, AV_{40}), visual-only (V_{240}) and auditory-only (A). The total EEG recording lasted around 20 min and was divided in two sessions of approximately 8 minutes with a short break between sessions.

Following EEG recording, the last ten participants performed a behavioral speeded forced-choice syllable identification task on AV_{240}, AV_{120}, AV_{40} stimuli but without sound (this control experiment was developed through the course of the study). On each trial, they were asked to identify as quickly as possible one syllable (/pa/, /ta/ or /ka/) by pressing one of three keys on a keyboard with their right hand. No feedback was provided. The response key designation was counterbalanced across participants. This lip-reading control experiment consisted of 36 trials (4 speakers x 3 syllables x 3 conditions) presented in a pseudo-randomized order (i.e., no more than two times the same syllable or the same experimental condition consecutively), including 12 trials in each of the 3 experimental conditions (V_{240}, V_{120}, V_{40}).

EEG setup

EEG data were continuously recorded using the Biosemi Active Two AD-box EEG system operating at a 512 Hz sampling rate. Since N1/P2 AEPs have maximal response over fronto-central sites (Scherg and Von Cramon, 1986; Näätänen and Picton, 1987) and in line with previous EEG studies of audiovisual speech perception (Stekelenburg and Vroomen 2007; Treille et al. 2014a, 2014b, 2017, 2018; Pinto et al., 2019; Tremblay et al., 2021; Sato, 2022), EEG were collected from F1, Fz, F2, FC1, FCz, FC2, C1, Cz, C2 fronto-central scalp electrodes (Electro-Cap International, INC), according to the international 10-20 system. Two additional electrodes served as ground electrodes (Common Mode Sense [CMS] active and Driven Right Leg [DRL] passive electrodes). Horizontal (HEOG) and vertical (VEOG) eye movements were recorded using electrodes positioned at the outer canthus of each eye and above the left eye. In addition, two external reference electrodes were attached over the left and the right mastoid bones. Before the experiments, the impedance of all electrodes was adjusted to get low offset voltages and stable DC.

Analyses

In all statistical analyses, the alpha level was set at $p = 0.05$ and Greenhouse–Geisser corrected when appropriate (for violation of the sphericity assumption). To determine the effect size of significant effect and interactions, partial eta squared ($p\eta^2$) were computed. When required, post hoc analyses were conducted with Bonferroni correction.

Lip-reading control experiment

In the lip-reading control experiment, both the percentage of correct responses and median RTs (calculated from the acoustic onset of each syllable) were determined for each participant and each experimental condition. One-way repeated measures ANOVAs were conducted on these measures with the stimulus (V_{240}, V_{120}, V_{40}) as within-participant factors.

EEG experiment

Accuracy

In the EEG experiment, the percentage of correct responses was determined for each participant and each experimental condition. For each experiment, a one-way repeated measures ANOVA was conducted with the stimulus (AV_{240}, AV_{120}, AV_{40}, A, V_{240}) as within-participant factors.

EEG signal

EEG data were processed using the EEGLAB software (Delorme and Makeig, 2004; version 2020.0) running on Matlab (Mathworks, Natick, USA; version R2019a). For each participant, 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 2.00, default parameter settings). Scalp channels were then automatically inspected, and bad channels interpolated using the EEGLAB Clean_rawdata plug-in (version 2.0, default parameter settings). On all channels, eye blinks, eye movements and other motion artefacts were detected and removed using the EEGLAB Artifact Subspace Reconstruction plug-in (version 0.13 merged into the Clean-rawdata plug-in, default parameter settings). 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.

For each experimental condition (AV_{240}, AV_{120}, AV_{40}, A, V_{240}), EEG data were segmented into 650 ms epochs, from −350 ms to 300 ms relative to the acoustic onset, corrected from a -350 ms to -250 ms baseline. This baseline ensured that visual movements consisted of a neutral mouth position for all stimuli (i.e., before the prephonatory movements; see Figure 1). Epochs with an amplitude change exceeding ±100 uV at any channels were further removed, and EEG data were averaged over the nine F1, Fz, F2, FC1, FCz, FC2, C1, Cz, C2 fronto-central electrodes. On average, the entire preprocessing pipeline rejected 15 ±10 % of epochs, leaving exactly 61 epochs per each experimental condition.

In order to determine the time windows of analysis for P1, N1 and P2 AEPs in an objective manner, P1, N1 and P2 peak latencies of the grand average waveform relative to all participants and all experimental conditions (i.e., AV_{240}, AV_{120}, AV_{40}, A) were first automatically determined from 0 ms to 100 ms, 50 ms to 150 ms and 150 ms to 250 ms, respectively (P1: 70 ms, N1: 134 ms, P2: 216 ms; see Figure 2). For each participant and each experimental condition, P1, N1 and P2 amplitudes and latencies were then automatically computed based on three fixed temporal windows defined as ±20 ms of P1, N1 and P2 peak

latencies previously calculated from the grand average waveform (Ganesh et al., 2014; Treille et al., 2014b; Sato, 2022). One-way repeated measures ANOVAs on both P1, N1 and P2 amplitudes and latencies was performed with the stimulus (AV_{240}, AV_{120}, AV_{40}, A) as within-participant factors.[1]

[1] In addition, we used an additive model to further test audiovisual integration, in which the bimodal audiovisual EEG signal was compared to the sum of auditory and visual unimodal EEG signals (i.e., $AV_{240} \neq A+V_{240}$; for a recent review, see Baart, 2016). A one-way repeated measures ANOVA on both P1, N1 and P2 amplitudes and latencies was performed with the stimulus type (AV_{240}, $A+V_{240}$) as within-participant factors. Since the results of this analysis appear in line with those reported in previous EEG studies on audiovisual speech integration, they are described in supplementary materials.

RESULTS

Lip-reading control experiment

The lip-reading control experiment performed on V_{240}, V_{120} and V_{40} stimuli showed that the duration of visual articulatory movements did not impact accuracy but RT. The mean proportion of correct responses was 73 %, without any significant difference between the stimuli ($F(2,18) = .1$, $p = .93$; V_{240}: 73 ±15 %, V_{120}: 72 ±12 %, V_{40}: 73 ±12 %. On the contrary, a significant increase of RTs was observed between V_{240}, V_{120} and V_{40} ($F(2,18) = 6.7$, $p = .007$, $p\eta^2 = .43$), with a shorter RT for V_{240} compared to V_{40} (V_{240}: 592 ±158 ms, V_{120}: 620 ±124 ms, V_{40}: 660 ±154 ms.

EEG

Accuracy

The mean proportion of correct responses in the EEG experiment was 90 %, with a near-ceiling effect for all stimuli except visual-only. A significant effect of the stimulus was observed ($F(4,76) = 101.1$, $p < .000001$, $p\eta^2 = .84$), with a lower accuracy for V_{240} compared to all other stimuli (AV_{240}: 95 ±6 %, AV_{120}: 95 ±6 %, AV_{40}: 96 ±6 %, A: 93 ±6 %, V_{240}: 70 ±10 %.

EEG results

For P1 amplitude, a significant effect of the stimulus was observed ($F(3,57) = 12.3$ $p < .000001$, $p\eta^2 = .39$), with a higher amplitude for AV_{120} compared to all other stimuli (AV_{240}: 3.3 ±2.7 uV, AV_{120}: 6.1 ±2.7 uV, AV_{40}: 3.0 ±2.4 uV, A: 2.8 ±2.9 uV. For P1 latency, a significant increase was observed from AV_{240} and AV_{120} to AV_{40} ($F(3,57) = 7.0$, $p < .001$, $p\eta^2 = .27$), with a shorter latency for AV_{240} and AV_{120} compared to AV_{40} (AV_{240}: 66 ±12 ms, AV_{120}: 70 ±13 ms, AV_{40}: 80 ±12 ms, A: 73 ±10 ms).

For N1 amplitude, a significant effect of the stimulus was observed ($F(3,57) = 9.7$ $p < .0001$, $p\eta^2 = .34$), with a higher negative amplitude for A compared to AV_{40} and AV_{120}, and for AV_{240} compared to AV_{40} (AV_{240}: -2.3 ±3.3 uV, AV_{120}: -1.0 ±3.7 uV, AV_{40}: -0.2 ±2.7 uV, A: -3.5 ±4.0 uV. For N1 latency, the main effect of the stimulus did not reach significance ($F(3,57) = 1.8$, $p = .18$; AV_{240}: 131 ±10 ms, AV_{120}: 135 ±9 ms, AV_{40}: 135 ±12 ms, A: 137 ±10 ms.

Finally, for P2 amplitude, a significant increase was observed from AV_{240} to AV_{120} and AV_{40} and to A (i.e., $AV_{240} < AV_{120} = AV_{40} < A$). A significant effect of the stimulus was observed ($F(3,57) = 15.8$, $p < .000001$, $p\eta^2 = .45$), with a lower amplitude for AV_{240} compared to all other stimuli as well as for AV_{120} and AV_{40} compared to A (AV_{240}: 6.4 ±3.6 uV, AV_{120}: 8.3 ±3.5 uV, AV_{40}: 8.4 ±3.9 uV, A: 11.9 ±3.5 uV. For P2 latency, the main effect of the stimulus also reached significance ($F(3,57) = 5.9$, $p = .006$, $p\eta^2 = .24$), with a shorter latency for AV_{240} and AV_{120} compared to A (AV_{240}: 211 ±12 ms, AV_{120}: 214 ±13 ms, AV_{40}: 218 ±14 ms, A: 222 ±11 ms.

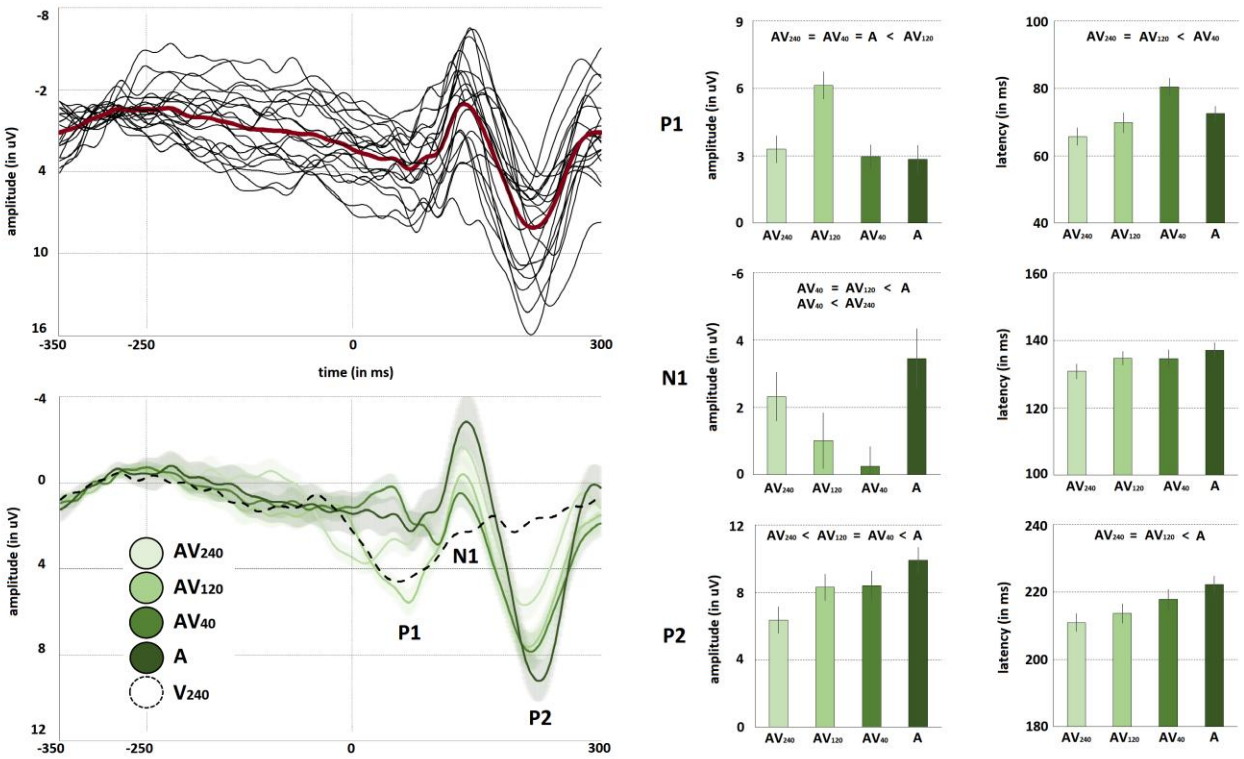

Figure 2. Left-Top: Individual EEG waveforms averaged over all experimental conditions (AV_{240}, AV_{120}, AV_{40}, A) and grand average EEG waveform averaged over all participants and experimental conditions (in red) on fronto-central electrodes. Left-Bottom: Average EEG waveform for each experimental condition (and V_{240} condition). Right: Mean P1, N1 and P2 AEP amplitudes and latencies in each experimental condition (the error bars represent the standard error of the mean).

DISCUSSION

Previous EEG/MEG studies on audiovisual speech perception have consistently reported that visual prediction modulates auditory neural processing, particularly as a function of the degree of visual saliency (van Wassenhove et al., 2005; Arnal et al., 2009). To further examine the effect of visual prediction, three sets of identical syllables only differing in the onset and duration of silent prephonatory movements were here compared. The main findings of the study are as follows. In the lip-reading control experiment, despite similar visual recognition rates, the timing of visual speech gestures linearly impacted visual recognition times, with the shortest duration of visual prephonatory movements, the longest RT. In the EEG experiment, a significant increase of P2 amplitude was observed during audiovisual speech perception from the longest to the shortest visual prephonatory movements. Taken together, these results demonstrate that audiovisual speech perception partly operates through visually-based predictions and related processing time, with acoustic-phonetic neural processing paralleling the timing of prephonatory articulatory gestures.

Limitations of the present study

Before to discuss these findings, it is important to consider two limitations of the present study. First, due to the limited number of trials per syllable for reliable EEG analyses, the observed results in each experimental conditions were based on the average of /pa/, /ta/ and /ka/ syllables (for similar averaging method on speech stimuli, see Lebib et al., 2003; Klucharev et al., 2003; Besle et al., 2004; Stekelenburg and Vroomen, 2007; Winneke and Phillips, 2011; Treille et al. 2014a, 2017, 2018; Pinto et al., 2019; Tremblay et al., 2021; Sato, 2022). As previously noted, these syllables are well-known to differ in terms of time-varying kinematic of articulatory movements, visuo-labial saliency, and visual recognition (Fisher, 1968; Summerfield, 1987; Schwartz and Savariaux, 2014). Second, it is also obvious that our result cannot be generalized to all speech contexts. Notably, one highly relevant assumption for visual-induced predictions is that the visual speech signal precedes the auditory one. As a matter of fact, the material choice in most EEG studies on audio-visual speech perception, as in the present study, consisted of isolated syllables/words in which the visual speech signal preceded the acoustic speech signal by tens and even hundreds of milliseconds (Chandrasekaran et al., 2009). It should however be noted that in more ecological and naturalistic situations, that is with continuous speech, the temporal relationship between auditory and visual speech onsets appears more variable and spans a range of 30–50 ms auditory lead to 170–200 ms visual lead (Schwartz and Savariaux, 2014). Given these limitations, future studies are needed to clarify whether the timing of visual prephonatory movements modulates auditory neural processing in a more phonemic/syllabic specific way and in more ecological situations.

The timing of visual speech did not impact accuracy but RT

In the lip-reading control experiment, a significant increase of RTs was observed between V_{240}, V_{120} and V_{40} stimuli, with the longest duration of visual prephonatory movements, the shortest RT. Since RTs were

computed from the acoustic onset, the observed difference can be easily explained by additional visual cues and processing time for the longer compared to shorter prephonatory movements. Interestingly, no significant difference in visual recognition rate was however observed between the three sets of visual stimuli. Since visual cues after the acoustic onset were the same in all three sets of syllables, this last result likely indicates that a single relevant frame of 40ms immediately preceding the acoustic onset was here sufficient for /pa/, /ta/ and /ka/ syllable discrimination.

N1/P2 auditory neural processing paralleled the timing of visual prephonatory movements

In the EEG experiment, results from the additive model appeared to be largely in agreement with previous EEG studies on audiovisual speech integration (Lebib et al., 2003; Klucharev et al., 2003; Besle et al., 2004; van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007; Winneke and Phillips, 2011; Treille et al., 2017, 2018; Pinto et al., 2019; Tremblay et al., 2021), with a reduced amplitude of P1 and P2 and a shorter latency of N1 and P2 for AV_{240} compared to $A+V_{240}$ (see supplementary materials).

More interestingly, the timing of visual gestures was found to modulate auditory neural processing differently for P1, N1 and P2 AEPs, presumably due to their respective roles in sensory gating, acoustic and phonetic decoding stages of auditory speech processing. While the higher P1 amplitude for AV_{120} is intriguing and may derive from complex multisensory gating processes (Klucharev et al., 2003), the significant increase observed from AV_{240} and AV_{120} to AV_{40} likely reflects the available processing time for multisensory gating mechanisms, with the shorter duration of visual prephonatory movements the longer latency. For N1 AEPs, a lower negative amplitude was observed for AV_{40} compared to AV_{240}. This result may derive from visual-to-auditory attentional processes known to induce a substantial increase of N1 but not preceding AEPs (Picton and Hillyard, 1974), with the closer the onset of visual prephonatory movements to the acoustic onset, the lower attentional load. Crucially, a significant increase of P2 amplitude was observed from AV_{240} to AV_{120} and AV_{40}. Previous EEG/MEG studies have shown that visual predictability induces a N1 and P2 facilitation of latency but not of amplitude, with the higher visual recognition rate of a syllable, the larger latency facilitation (van Wassenhove et al., 2005; Arnal et al., 2009). Since the viseme-specific facilitation of early auditory responses was not influenced by audiovisual incongruence, it has been hypothesized to depend on a direct corticocortical pathway from visual motion areas to auditory cortices and to reflect phase-resetting of auditory activity by visual syllables (Arnal et al., 2009). The present results appear to complement these findings and further clarify how audiovisual speech perception partly operates through visually-based predictions. Despite no difference in visual recognition rates here observed across the three sets of syllables, P2 amplitude classically involved in the acoustic-phonetic decoding stages of auditory speech processing was found to linearly parallel the timing of visual prephonatory gestures. This suggests that the amount of predictive visual information and related processing time before the acoustic speech signal linearly impact the phonetic neural processing of audiovisual speech perception.

REFERENCES

Arnal LH, Morillon B, Kell CA, Giraud AL (2009) Dual neural routing of visual facilitation in speech processing. The Journal of Neuroscience, 29(43):13445-13453.

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.

Benoît C, Mohamadi T, Kandel S (1994) Effects of phonetic context on audio–visual intelligibility of French speech in noise. J Speech Hear Res, 37:1195–1203

Besle J, Fort A, Delpuech C, Giard MH (2004) Bimodal speech: early suppressive visual effects in human auditory cortex. European journal of Neuroscience, 20:2225-2234.

Boersma P, Weenink D (2013) Praat: doing phonetics by computer. Computer program, Version 6.1., http://www.praat.org/.

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

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

Fisher CG (1968) Confusions among visually perceived consonants. Journal of Speech and Hearing Research, 11: 796-804.

Ganesh AC, Berthommier F, Vilain C, Sato M, Schwartz JL (2014) A possible neurophysiological correlate of audiovisual binding and unbinding in speech perception. Front. Psychol., 5: 1340.

Grant KW, Walden BE, Seitz PF (1998) Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition and auditory-visual integration. J. Acoust. Soc. Am., 103:2677–2690.

Grant KW, Seitz PFP (2000) The use of visible speech cues for improving auditory detection of spoken sentences. The Journal of the Acoustical Society of America, 108(3):1197–1208.

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

Lebib R, Papo D, de Bode S, Baudonnière PM (2003) Evidence of a visual-to-auditory cross-modal sensory gating phenomenon as reflected by the human P50 event-related brain potential modulation. Neuroscience Letters, 341(3):185-188.

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

Navarra J, Soto-Faraco S (2005) Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. Psychol. Res., 71(1):4–12.

Oldfield RC (1971) The Assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia, 9:97-113.

Picton TW, Hillyard SA (1974) Human auditory evoked potentials. II: Effects of attention. Electroencephalography and Clinical Neurophysiology, 36: 191-200.

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.

Rosenblum LD, Dorsi J, Dias JW (2016) The impact and status of Carol Fowler's supramodal theory of multisensory speech perception. Ecological Psychology, 28(4):262-294.

Sato M (2022). Motor and visual influences on auditory neural processing during speaking and listening. Cortex.

Scherg M, VonCramon D (1986) Evoked dipole source potentials of the human auditory cortex. Electroencephalogr. Clin. Neurol., 65:344–360.

Schwartz JL, Berthommier F, Savariaux C (2004) Seeing to hear better: evidence for early audio-visual interactions in speech identification. Cognition, 93:69-78.

Schwartz JL, Savariaux C (2014) No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. PLOS Computational Biology, 10(7): e1003743.

Stekelenburg JJ, Vroomen J (2007) Neural correlates of multisensory integration of ecologically valid audiovisual events. J Cogn Neurosci, 19:1964–1973.

Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. The Journal of the Acoustical Society of America, 26:212-215.

Summerfield QA (1987) Some preliminaries to a comprehensive account of audio-visual speech perception Hearing by Eye: The Psychology of LipReading (pp. 3-51). Londres: Erlbaum Associates.

Treille A, Cordeboeuf C, Vilain C, Sato M (2014a) Haptic and visual information speed up the neural processing of auditory speech in live dyadic interactions. Neuropsychologia, 57:71–77.

Treille A, Vilain C, Sato M (2014b) The sound of your lips: electrophysiological crossmodal interactions during hand-to-face and face-to-face speech perception. Front. Psychol., 5: 420.

Treille A, Vilain C, Kandel S, Sato M (2017) Electrophysiological evidence for a self-processing advantage during audiovisual speech integration. Experimental Brain Research.

Treille A, Vilain C, Schwartz J-L, Hueber T, Sato M (2018) Electrophysiological evidence for audio-visuo-lingual speech integration. Neuropsychologia, 109: 126-133

Tremblay P, Basirat A, Pinto S, Sato M (2021) Visual prediction cues can facilitate behavioural and neural speech processing in young and older adults. Neuropsychologia, 159: 107949.

van Wassenhove V, Grant KW, Poeppel D (2005) Visual speech speeds up the neural processing of auditory speech. Proc Natl Acad Sci USA, 102:1181–1186.

van Wassenhove V (2013) Speech through ears and eyes: interfacing the senses with the supramodal brain. Front. Psychol., 4: 1–17.

Vroomen J, Stekelenburg JJ (2010) Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. J Cogn Neurosci, 22:1583–1596.

Winneke AH, Phillips NA (2011) Does audiovisual speech offer a fountain of youth for old ears? An event-related brain potential study of age differences in audiovisual speech perception. Psychology and Aging, 26(2):427–438.