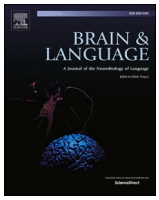
Brain & Language 247 (2023) 105359

Contents lists available at ScienceDirect

Brain and Language

journal homepage: www.elsevier.com/locate/b&l



Short communication

Competing influence of visual speech on auditory neural adaptation

Marc Sato

Laboratoire Parole et Langage, Centre National de la Recherche Scientifique, UMR 7309 CNRS & Aix-Marseille Université, Aix-Marseille Université, 5 avenue Pasteur, Aix-en-Provence, France

ARTICLE INFO

Keywords:
Auditory neural adaptation
Auditory memory
Audiovisual speech perception
EEG

ABSTRACT

Visual information from a speaker's face enhances auditory neural processing and speech recognition. To determine whether auditory memory can be influenced by visual speech, the degree of auditory neural adaptation of an auditory syllable preceded by an auditory, visual, or audiovisual syllable was examined using EEG. Consistent with previous findings and additional adaptation of auditory neurons tuned to acoustic features, stronger adaptation of N1, P2 and N2 auditory evoked responses was observed when the auditory syllable was preceded by an auditory compared to a visual syllable. However, although stronger than when preceded by a visual syllable, lower adaptation was observed when the auditory syllable was preceded by an audiovisual compared to an auditory syllable. In addition, longer N1 and P2 latencies were then observed. These results further demonstrate that visual speech acts on auditory memory but suggest competing visual influences in the case of audiovisual stimulation.

1. Introduction

Exploiting cross-modal regularities and complementarities between acoustic and visual speech signals is a key mechanism to help extract and decode phonetic cues in the continuous audiovisual speech stream (Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). Behaviorally, the high level of auditory and visual cross-predictability appears of tremendous benefit for speech perception (van Wassenhove, 2013; Bernstein & Liebenthal, 2014; Rosenblum, Dorsi, & Dias, 2016). Adding time-varying visual information from the speaker's face enhances sensitivity to acoustic speech cues by decreasing auditory detection threshold (Grant & Seitz, 2000; Schwartz, Berthommier, & Savariaux, 2004), speeds up and improves auditory speech recognition (Sumby & Pollack, 1954), enhances second language perception (Navarra & Soto-Faraco, 2005), and benefits hearing-impaired listeners (Grant, Walden, & Seitz, 1998). Adding to these behavioral findings, electro- and magnetoencephalography (EEG/MEG) studies have consistently reported that prephonatory visual movements before the acoustic speech onset modulates subsequent auditory processing. Specifically, the amplitude and latency of N1 and P2 auditory evoked potentials (AEPs) are attenuated and speeded up during audiovisual compared to unimodal auditory speech perception (Klucharev, Möttönen, & Sams, 2003; Besle, Fort, Delpuech, & Giard, 2004; van Wassenhove, Grant, & Poeppel, 2005; Arnal, Morillon, Kell, & Giraud, 2009; for a review, see Baart, 2016). These visual-to-auditory modulatory effects depend on the degree of visual salience, with the higher visual recognition the stronger latency facilitation (van Wassenhove et al., 2005; Arnal et al., 2009), and on the timing of visual prephonatory movements, with the longer duration the stronger amplitude suppression (Sato, 2022a).

To further understand visual influences on auditory speech processing, several EEG/MEG studies have focused on auditory memory and neural adaptation mechanisms. These studies were based on two classic and closely related phenomena (for a comparison, see Jääskeläinen et al., 2004a): an attenuation of AEPs by repeated auditory stimulation (repetition suppression, RS) and a negative deflection in AEPs by changes in the repetitive aspects of auditory stimulation (mismatch negativity, MMN). Using an irrelevant identification task, RS has classically been observed when a speech sound was repeated but also when it was preceded by its visual counterpart (Jääskeläinen et al., 2004b). Furthermore, in a variety of oddball paradigms, MMN was found to be triggered not only by infrequent compared to frequent auditory speech stimuli, but also by infrequent incongruent compared to frequent congruent audiovisual speech stimuli (Sams et al., 1991; Colin et al., 2002, 2004; Möttönen, Krause, Tiippana, & Sams, 2002; Hertrich, Mathiak, Lutzenberger, Menning, & Ackermann, 2007; Saint-Amour, De Sanctis, Molholm, Ritter, & Foxe, 2007; Stekelenburg, Keetels, & Vroomen, 2018). In all these studies, RS and MMN therefore appeared in the absence of any acoustic change. Taken together, they demonstrate that visual speech information acts on the memory traces of specific

E-mail address: marc.sato@cnrs.fr.

https://doi.org/10.1016/j.bandl.2023.105359
Received 19 April 2023; Received in revised form 25 September 2023; Accepted 6 November 2023

acoustic features and regularities in the auditory cortex.

From these studies, one remaining issue is whether an auditory memory trace can be consolidated by an audiovisual compared to an auditory speech stimulation. The hypothesis of an enhanced representation of a speech sound by the addition of visual information seems consistent with the above-mentioned visual-to-auditory perceptual benefits and neurophysiological modulatory effects. To investigate this question, the goal of the present EEG study was to compare the degree of auditory neural adaptation of an auditory syllable following the presentation of either an auditory, visual, or audiovisual syllable. To this end, participants performed a syllable discrimination task on two successive same or different syllables. Crucially, while the first syllable was presented auditorily, visually, or audiovisually, the second syllable was always presented auditorily (see Fig. 1).

N1, P2 and N2 AEPs were compared to determine the impact of the modality of the first syllable on the subsequent auditory syllable. As mentioned above, early N1 and P2 AEPs are classically examined in audiovisual speech perception studies. Interestingly, these studies suggest two successive audiovisual interactions in association with speech recognition: a fast direct feedforward neural route from the visual cortex to the auditory cortex that helps tuning auditory processing depending upon visual motion temporal cues, and a slower and indirect feedback pathway from the associative posterior superior temporal sulcus that functions as a phonological error signal between visual prediction and auditory input (Hertrich et al., 2007; Arnal et al., 2009). In line with this proposal, suppression and speeding-up of N1 are unaffected by whether the auditory and visual information are phonologically congruent or incongruent, but crucially depend on whether the visual information contained anticipatory visual-to-auditory temporal information (Stekelenburg & Vroomen, 2007). In contrast, processing audiovisual congruency have been shown to start from P2 and later (Stekelenburg & Vroomen, 2007; Hertrich et al., 2007; Arnal et al., 2009). In addition to N1 and P2 AEPs, we therefore extended our investigation to the later N2 AEP, which is also commonly associated with detection of violations of regularities and changes in auditory memory (Näätänen, 1992).

Based on the above studies, it was hypothesized that neural auditory adaptation should be stronger when the auditory syllable was preceded by an auditory or audiovisual syllable rather than a visual one, reflecting additional adaptation of auditory neurons tuned to acoustic features. Crucially, stronger auditory neural adaptation due to a preceding audiovisual syllable compared to an auditory syllable would suggest that adding visual information from a speaker's face to an auditory speech sound enhances its auditory memory trace.

2. Methods

2.1. Participants

Twenty healthy adults (16 females and 4 males), with a mean age of 23 ± 4 years (range: 19–32 years), participated in the study after giving informed consent. All participants were native French speakers, with an average of 14 ± 2 years of education (range: 12–17 years). They were all right-handed according to the standard handedness inventory (Oldfield, 1971) with a mean score of 76 ± 15 % (range: 56–100 %), 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.

2.2. Stimuli

Multiple utterances of /pa/ and /ta/ syllables, starting with a visually neutral open mouth, were individually recorded by a female native French speaker in a soundproof room. These two syllables included an initial unvoiced stop consonant, allowing precise detection of the acoustic syllable onset for EEG analyses, and were highly discriminable visually from each other. Video digitizing (centered on the speaker's mouth; see Fig. 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.

Using Adobe Premiere (Adobe systems, San Jose, USA) and Praat (Boersma & Weenink, 2013), two clearly articulated /pa/ and /ta/ tokens were selected and edited based on acoustic and visual properties. The editing procedure ensured that the two selected 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 of the syllable. The acoustic intensity was normalized using a common maximal amplitude criterion.

For each syllable, the first frame corresponding to the neutral open mouth position was replicated before the prephonatory movements (9 frames, 360 ms) and after the phonatory movements (15 frames, 600 ms). A second auditory syllable (/pa/ or /ta/) was added 600 ms after the first one. With this procedure, AV-A stimuli (35 frames, 1400 ms) consisted of two successive audiovisual and auditory syllables (/pa/-/pa/, /pa/-/ta/, /ta/-/pa/ or /ta/-/ta/) with visual prephonatory and phonatory movements before and after the acoustic consonantal burst of the first audiovisual syllable. For V-A stimuli, visual prephonatory and phonatory movements of the first visual syllable were presented without the acoustic speech sound. For A-A stimuli, visual prephonatory and phonatory movements of the first auditory syllable were replaced by a visually neutral open mouth[1]. Importantly, for all stimuli, the 600 ms delay between the first and second syllables was fixed, and the second syllable was always presented auditorily, with a still image of a visually neutral open mouth.

2.3. 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 discrimination task. On each trial, they determined whether the two successive syllables were the same or different by pressing one of two 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 (1400 ms) was followed by a blank screen (600 ms) and then by a question mark (1000 ms), which served as cue for participants' responses (see Fig. 1).

There were 6 experimental conditions related to the modality of

[1] With the exception of the first syllable including prephonatory movements in the visual and audiovisual conditions, a visually neutral open mouth was presented throughout each trial. Although it is unknown whether a neutral still face of a speaker can modulate AEPs associated with an auditory speech stimulation (compared to a blank screen or a fixation point), this procedure has been classically used in previous EEG studies of audiovisual speech perception to prevent the identity of the speech stimulus before the acoustic onset in the auditory condition and to balance as much as possible the visual attentional state of the participants between the auditory, visual and audiovisual conditions. Importantly, this procedure was similar to that used by Jääskeläinen et al. (2004b), who observed that RS caused by a preceding auditory syllable was stronger than that caused by a preceding visual syllable, with the mouth closed of a speaker shown on the screen throughout the experiment (with the exception of the first visual syllable).

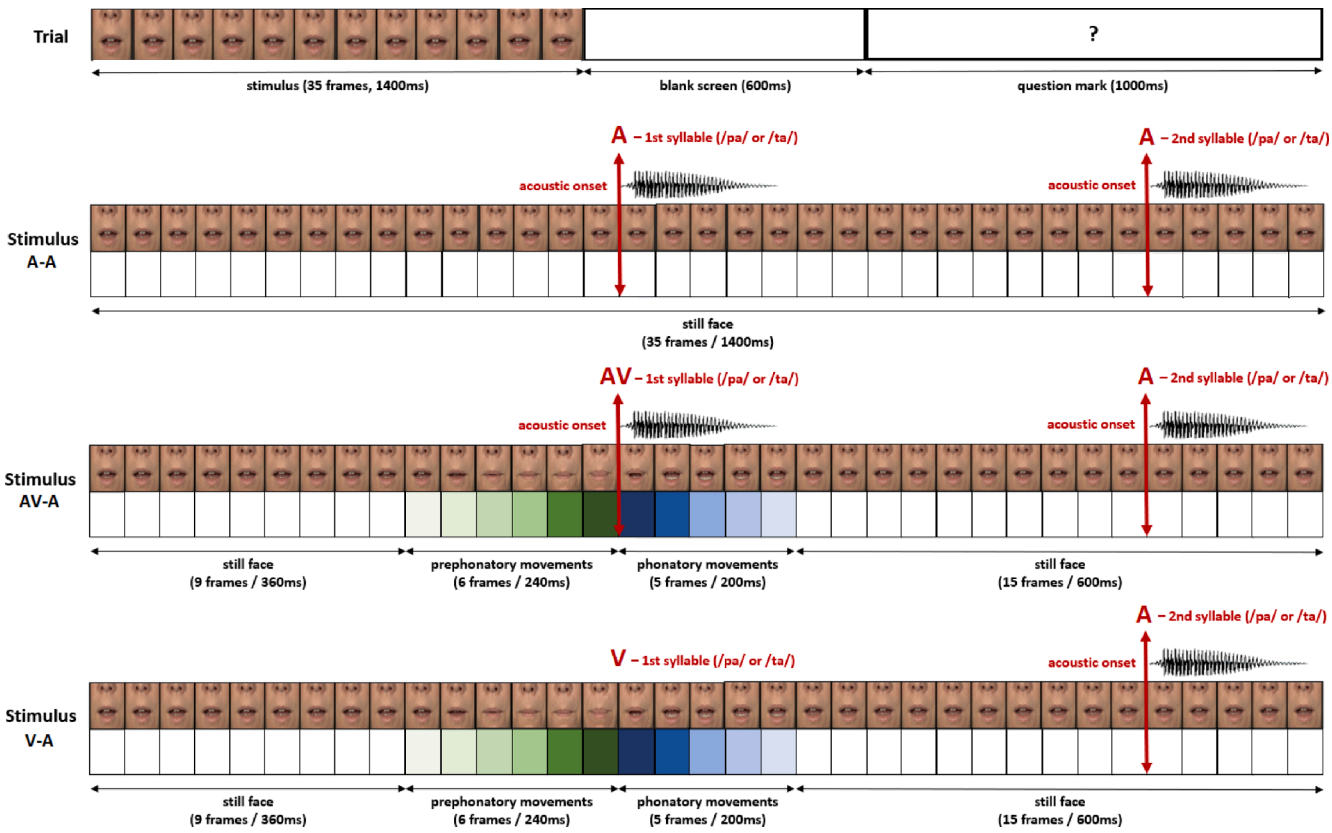


Fig. 1. Experimental design. Each trial consisted of a stimulus (1400 ms) followed by a blank screen (600 ms) and then by a question mark (1000 ms), which served as cue for participants' responses to the syllable discrimination task. A-A stimuli consisted of two successive auditory syllables (/pa/-/pa/, /pa/-/ta/, /ta/-/pa/ or /ta/-/ta/, inter-syllabic interval of 600 ms) presented with a visually neutral open mouth. For AV-A stimuli, visual prephonatory and phonatory movements were presented before and after the acoustic consonantal burst of the first syllable. For V-A stimuli, visual prephonatory and phonatory movements of the first syllable were presented without the acoustic speech sound. For all stimuli, the second auditory syllable was presented with a visually neutral open mouth.

presentation of the first syllable (A: auditory, V: visual, AV: audiovisual) and to the matching between the two successive syllables (same: /pa/-/pa/ or /ta/-/ta/, different: /pa/-/ta/ or /ta/-/pa/): A-A_{same} A-$A_{different}$, AV-A_{same}, AV-$A_{different}$, V-A_{same}, V-$A_{different}$). The experiment consisted of 3 sessions of 144 trials (6 experimental conditions × 24 trials), each presented in a pseudo-randomized order (i.e., no more than one time the same experimental condition or the same two consecutive syllables). In total, each experimental condition included 72 trials and the total EEG recording lasted around 25 min with a short break between sessions.

2.4. 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 & Von-Cramon, 1986; Näätänen & Picton, 1987) and in line with previous EEG studies of audiovisual speech perception (Stekelenburg & Vroomen, 2007; Vroomen & Stekelenburg, 2010; Treille et al., 2014a, 2014b, 2017, 2018; Pinto, Tremblay, Basirat, & Sato, 2019; Tremblay, Basirat, Pinto, & Sato, 2021; Sato, 2022a,b), 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.

2.5. 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 Newman–Keuls tests for multiple comparisons.

2.5.1. Accuracy

The percentage of correct responses was determined for each participant and each experimental condition. For each experiment, a two-way repeated measure ANOVA was conducted with the modality of the first syllable (A-A, AV-A, V-A) and the syllable matching (same, different) as within-participant factors.

2.5.2. EEG signal

EEG data were processed using the EEGLAB software (Delorme & 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 (0.5–30 Hz)[2]. 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.

Since the syllable discrimination task was almost perfectly performed (mean proportion of correct responses of 95 %; see below), ERPs were computed across all trials for each experimental condition. For each modality (A-A, AV-A, V-A), each successive syllable (first, second) and each syllable matching (same, different), EEG data were segmented into 500 ms epochs, from −100 ms to 400 ms relative to the acoustic onset, corrected from a −100 ms to 0 ms baseline[3]. 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 13 % of epochs. A three-way repeated measure ANOVAs on the number of artifact-based rejected epochs was performed with the modality (A-A, AV-A, V-A), the syllable order (first, second) and the syllable matching (same, different) as within-participant factors. Only a significant effect of the syllable order was observed ($F(1,19) = 7.2$, $p = .01$, $p\eta^2 = 0.28$), with a lower number of artifact-based rejected epochs for the first compared to the second syllable (8% vs. 18%). No other main effect or interaction was found.

In order to determine the time windows of analysis for N1, P2 and N2 AEPs in an objective manner, N1, P2 and N2 peak latencies of the grand average waveform relative to all participants and all experimental conditions were first automatically determined from 50 ms to 150 ms, 150 ms to 250 ms and 250 ms to 350 ms, respectively (N1: 138 ms, P2: 216 ms, N2: 322; see Fig. 2A). For each participant and each experimental condition (except for the first visual syllabe in the V-A modality in which no AEPs were observed; see Fig. 2B), N1, P2 and N2 amplitudes and latencies were then automatically computed based on two fixed temporal windows defined as ± 20 ms of N1, P2 and N2 peak latencies previously calculated from the grand average waveform (Ganesh, Berthommier, Vilain, Sato, & Schwartz, 2014; Treille, Vilain, & Sato, 2014b; Sato, 2022a,b).

To determine whether the modality of presentation of the first syllable influenced AEPs of the second auditory syllable, two-way repeated measure ANOVAs on N1, P2 and N2 amplitudes and latencies of the second auditory syllables were performed with the modality of the first syllable (A, V, AV) and the syllable matching (same, different) as within-participant factors. To further evaluate the extent of auditory adaptation on the second compared to the first syllables, two-way repeated measure ANOVAs on N1-P2 peak-to-peak amplitudes were performed with the syllable (A, AV, (A)-A, (AV)-A, (V)-A) and the syllable matching (same, different) as within-participant factors. Since the results of this second analysis mainly confirm those of the first, they are described in supplementary materials (see SM1).

3. Results

3.1. Accuracy

The mean proportion of correct responses was 95 %, with a ceiling effect for all modalities except V-A (A-A_{same}: 97 %, A-$A_{different}$: 98 %, AV-A_{same}: 97 %, AV-$A_{different}$: 97 %, V-A_{same}: 91 %, V-$A_{different}$: 93 %). A strong effect of the modality of the first syllable was observed ($F(2,38) = 23.0$, $p < .000001$, $p\eta^2 = 0.55$), with a lower accuracy for V-A compared to A-A and AV-A ($p = .0001$ for both post hoc comparisons). The main effect of syllable matching did not reach significance ($F(1,19) = 1.1$, $p = .30$) nor the modality × syllable matching interaction ($F(2,38) = 0.9$, $p = .38$).

In sum, the syllable discrimination task was almost perfectly performed, although more challenging when the first syllable had to be lipread in the V-A modality.

3.2. N1, P2 and N1 AEPs of the second auditory syllable

For N1 amplitude, a strong effect of the modality of the first syllable on the second auditory syllable was observed ($F(2,38) = 40.3$, $p < .000001$, $p\eta^2 = 0.68$), with a lower negative amplitude for (A)-A compared to (AV)-A, and for (AV)-A compared to (V)-A ((A)-A: −3.42 uV, (AV)-A: −5.15 uV, (V)-A: −7.30 uV). The main effect of syllable matching did not reach significance ($F(1,19) = 1.0$, $p = .32$) nor the modality × syllable matching interaction ($F(2,38) = 1.2$, $p = .31$).

For P2 amplitude, a significant effect of the modality of the first syllable on the second auditory syllable was observed ($F(2,38) = 6.0$, $p = .005$, $p\eta^2 = 0.24$), with a lower positive amplitude for (A)-A and (AV)-A compared to (V)-A ((A)-A: 2.15 uV, (AV)-A: 1.41 uV, (V)-A: 3.36 uV). The main effect of syllable matching did not reach significance ($F(1,19) = 0.0$, $p = .97$) nor the modality × syllable matching interaction ($F(2,38) = 0.9$, $p = .42$).

For N2 amplitude, a strong effect of the modality of the first syllable on the second auditory syllable was observed ($F(2,38) = 47.6$, $p < .000001$, $p\eta^2 = 0.71$), with a lower negative amplitude for (A)-A compared to (AV)-A, and for (AV)-A compared to (V)-A ((A)-A: −3.02 uV, (AV)-A: −4.55 uV, (V)-A: −8.43 uV). The main effect of syllable matching was also significant ($F(1,19) = 6.5$, $p = .02$, $p\eta^2 = 0.26$), with a lower negative amplitude for the same compared to different successive syllables (same: −4.76 uV, different: −5.91 uV). The modality × syllable

[2] Although the 0.5Hz high pass filter used here may carry the risk of distorting EEG data (Tanner, Morgan-Short, & Luck, 2015), a two-way least-square FIR filtering of 0.5–30 Hz was here applied for two main reasons. First, most previous EEG studies that examined audiovisual speech perception and N1/P2 AEPs used similar filtering (e.g., Besle et al., 2004: 1-30Hz; Ganesh et al., 2014: 2-20Hz; Klucharev et al., 2003: 1-25Hz; Pinto et al., 2019: 3-30Hz; Stekelenburg & Vroomen, 2007: 0.5-30Hz; Treille, Cordeboeuf, Vilain, & Sato, 2014: 1-20Hz; Tremblay et al., 2021: 3-30Hz; van Wassenhove et al., 2005: 1-55Hz; Vroomen & Stekelenburg, 2010: 0.5-30Hz). Second, the syllable discrimination task may have induced task-related neural activity common to all experimental conditions and characterized by a slow deflection on fronto-central sites (Teder-Sälejärvi, McDonald, Di Russo, & Hillyard, 2002). Since slow deflection is well known to induce a substantial increase of artefact-based rejected epochs, a 0.5Hz high-pass filter was applied on the EEG data to minimize the contribution of slow potentials and related artefact-based rejected epochs. However, the entire EEG data set was reanalyzed using the same pipeline but based on a two-way least-square FIR filtering of 0.01–30 Hz, also excluding trials with incorrect responses. Since the results of this second analysis confirm those of the first, with the notable exception of a substantial increase in rejected trials and a few significant differences on P2 amplitude and latency, they are described in supplementary materials (see SM2).

[3] The entire EEG data set was also reanalyzed with a fixed baseline that does not imply any visual differences between the experimental conditions. To this aim, EEG data were segmented into 1340 ms epochs, from −340 ms to 400 ms relative to the acoustic onset of the second syllable, corrected from a −340 ms to −240 ms baseline (with a neutral open mouth in all conditions; see Fig. 1). However, this new baseline appears to be contaminated by vertex positive potentials likely elicited by the start of the trial and the appearance of the speaker's face. More importantly, not only did this new analysis substantially increase related artefact-based rejected epochs, but marked differences between experimental conditions in slow potentials before the acoustic onset of the second auditory syllable did not allow a strict comparison of peak amplitudes of AEPs. Results of this third analysis are described in supplementary materials (see SM3).

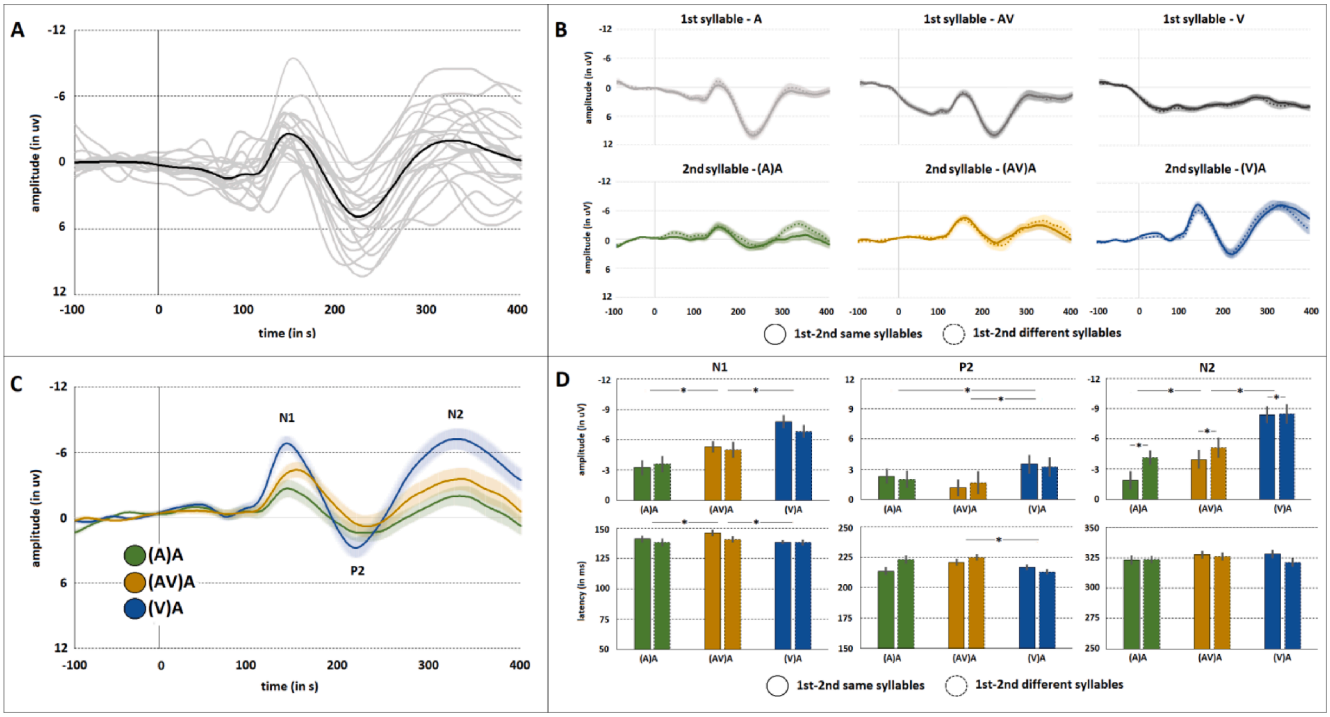


Fig. 2. A. Individual EEG waveforms averaged over all experimental conditions (in grey) and grand average EEG waveform averaged over all participants and all experimental conditions (in black) on fronto-central electrodes. B. Grand average EEG waveform for each modality (A-A, AV-A, V-A), each successive syllable (first, second) and each syllable matching (same, different). C. Grand average EEG waveforms for the second auditory syllable related to the modality of presentation of the first syllable. D. Mean N1, P1 and N2 AEP amplitudes and latencies for the second auditory syllable related to the modality of presentation of the first syllable and the matching between the two successive syllables (the error bars represent the standard error of the mean).

matching interaction did not reach significance ($F(2,38) = 2.8$, $p = .07$).

For N1 latency, a significant effect of the modality of the first syllable on the second auditory syllable was observed ($F(2,38) = 5.1$, $p = .01$, $p\eta^2 = 0.21$), with a longer latency for (AV)-A compared to (A)-A and (V)A ((A)-A: 140 ms, (AV)-A: 144 ms, (V)-A: 139 ms). The main effect of syllable matching did not reach significance ($F(1,19) = 2.8$, $p = .11$) nor the modality × syllable matching interaction ($F(2,38) = 1.2$, $p = .31$).

For P2 latency, a significant effect of the modality of the first syllable on the second auditory syllable was observed ($F(2,38) = 5.1$, $p = .01$, $p\eta^2 = 0.21$), with a longer latency for (AV)-A compared to (V)A ((A)-A: 218 ms, (AV)-A: 223 ms, (V)-A: 215 ms). The main effect of syllable matching did not reach significance ($F(1,19) = 4.0$, $p = .06$) nor the modality × syllable matching interaction ($F(2,38) = 3.0$, $p = .06$).

For N2 latency, the main effect of the modality ($F(2,38) = 1.0$, $p = .39$), the syllable matching ($F(1,19) = 1.3$, $p = .27$) and the modality × syllable matching interaction ($F(2,38) = 1.0$, $p = .39$) were not reliable.

In sum, the modality of presentation of the first syllable had a strong impact on N1 and N2 amplitudes of the second auditory syllable, with a lower negative amplitude for (A)-A compared to (AV)-A, and for (AV)-A compared to (V)-A. For P2 amplitude, a similar but more moderate effect was observed, with a lower positive amplitude for (A)-A and (AV)-A compared to (V)-A. In addition, the syllable matching also had a significant effect on N2, with a lower negative amplitude for the same compared to different successive syllables. Finally, the modality of presentation of the first syllable had a significant effect on N1 and P2 latencies of the second syllable, with a longer latency for (AV)-A compared to (A)-A and/or (V)-A.

4. Discussion

In line with previous findings and consistent with additional adaptation of auditory neurons tuned to acoustic features, stronger neural adaptation on N1, P2 and N2 AEPs was observed when the auditory syllable was preceded by an auditory or audiovisual compared to a visual syllable. However, contrary to our hypothesis, weaker neural adaptation was observed when the auditory syllable was preceded by an audiovisual compared to an auditory syllable. In addition, longer N1 and P2 latencies were then observed.

Before discussing these findings, it is worth noting that, behaviorally, the syllable discrimination task was equally and perfectly performed when the auditory syllable was preceded by an auditory or audiovisual syllable. Only the discrimination accuracy was slightly lower in case of lipreading, that is when the auditory syllable was preceded by a visual one. As in a previous MEG study by Jääskeläinen and colleagues (2004b), it is also important to note that no significant syllable-specific adaptation effects were observed on N1 nor on P2. For the authors, this can be explained by shared acoustic and phonetic properties of the speech stimuli (with in the present study all syllables differing in terms of bilabial /p/ and alveolar /t/ consonantal bursts but sharing the /a/ vowel). In contrast, coherent with its role in active discrimination, syllable matching did act on N2, with a smaller negative amplitude for the same successive syllables than for different syllables.

Consistent with additional adaptation of auditory neurons to acoustic features, and in agreement with Jääskeläinen et al. (2004b), RS caused by a preceding auditory syllable was stronger than that caused by a preceding visual syllable. This appears also consistent with the stronger RS related to a preceding audiovisual compared to a visual syllable. Compared to lip reading, given the respective roles of N1 and P2 in the acoustic/phonetic decoding stages of auditory speech processing and of N2 in detecting changes in auditory memory (Näätänen, 1992), auditory and audiovisual perception of the first syllable would have led to an improved representation of the speech sound in auditory memory and, consequently, to stronger adaptation during the presentation of the second auditory syllable. However, since acoustic features of the auditory and audiovisual syllables were strictly identical, neural auditory adaptation cannot explain the weaker RS caused by a preceding

audiovisual syllable compared to an auditory syllable. Nor it can explain the observed longer N1 and P2 latencies. Given the previously mentioned visual-to-auditory perceptual benefits and neurophysiological modulatory effects, a weaker auditory memory trace appears quite unlikely in the case of an audiovisual stimulation. Several alternative hypotheses can be put forward, namely refractoriness, visual attentional process, and task-based strategy. A first hypothesis is that the observed RS would primarily be the result of a refractory period or recovery cycle of auditory neurons (Rossburg et al., 2022) after the presentation of the first syllable. This appears coherent with the amplitude differences in AEPs observed for the first auditory, audiovisual and visual syllables (i.e., A > AV > V; see supplementary materials), with the higher AEP for the first syllable the stronger refractoriness and lower AEP for the second syllable. However, this hypothesis cannot explain the longer N1 and P2 latencies for the second auditory syllable only observed in the case of a preceding audiovisual syllable. A second hypothesis is that of a higher attentional degree in case of visual stimulation. This appears consistent with the slightly lower performance when the auditory syllable was preceded by a visual syllable. However, as for the refractoriness hypothesis, this cannot explain the longer N1 and P2 latencies only observed in the case of a preceding audiovisual syllable, not in the case of a preceding visual syllable. Alternatively, the greater difficulty in recognizing visual syllables might have prompted a task-based strategy. During the task, the combination of auditory and visual speech cues for audiovisual syllables would then have been a way to establish a better match between visual and auditory speech cues. From this task-based strategy, a higher degree of attention and the matching between visual and auditory speech representations could explain both the lower reduction in N1 and N2 amplitudes but longer N1 and P2 latencies in the case of a preceding audiovisual syllable.

Taken together, the present results again demonstrate that visual speech acts on auditory memory but suggest competing visual influences in the case of a preceding audiovisual stimulation. From a broader perspective, they raise the old but debated issue of how visual speech is represented in memory (Bernstein & Liebenthal, 2014).

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bandl.2023.105359.

References

Arnal, L. H., Morillon, B., Kell, C. A., & Giraud, A. L. (2009). Dual neural routing of visual facilitation in speech processing. *The Journal of Neuroscience, 29*(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.

Bernstein, L. E., & Liebenthal, E. (2014). Neural pathways for visual speech perception. *Frontiers in Neuroscience, 8*, 386.

Besle, J., Fort, A., Delpuech, C., & Giard, M. H. (2004). Bimodal speech: Early suppressive visual effects in human auditory cortex. *European journal of Neuroscience, 20*, 2225–2234.

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 Computational Biology, 5*, e1000436.

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

Colin, C., Radeau, M., Soquet, A., & Deltenre, P. (2004). Generalization of the generation of an MMN by illusory McGurk percepts: Voiceless consonants. *Clinical Neurophysiology, 115*(9), 1989–2000.

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

Ganesh, A. C., Berthommier, F., Vilain, C., Sato, M., & Schwartz, J. L. (2014). A possible neurophysiological correlate of audiovisual binding and unbinding in speech perception. *Frontiers in Psychology, 5*, 1340.

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition and auditory-visual integration. *The Journal of the Acoustical Society of America, 103*, 2677–2690.

Grant, K. W., & Seitz, P. F. P. (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.

Hertrich, I., Mathiak, K., Lutzenberger, W., Menning, H., & Ackermann, H. (2007). Sequential audiovisual interactions during speech perception: A whole-head MEG study. *Neuropsychologia, 45*(6), 1342–1354.

Jääskeläinen, I. P., Ahveninen, J., Bonmassar, G., Dale, A. M., Ilmoniemi, R. J., Levänen, S., et al. (2004a). Human posterior auditory cortex gates novel sounds to consciousness. *Proceedings of the National Academy of Sciences, 101*, 6809–6814.

Jääskeläinen, I. P., Ojanen, V., Ahveninen, J., Auranen, T., Levänen, S., Möttönen, R., et al. (2004b). Adaptation of neuromagnetic N1 responses to phonetic stimuli by visual speech in humans. *Neuroreport, 15*(18), 2741–2744.

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

Möttönen, R., Krause, C. M., Tiippana, K., & Sams. (2002). Processing of changes in visual speech in the human auditory cortex. *Cognitive Brain Research, 13*(3), 417–425.

Näätänen, R., & Picton, T. W. (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.

Näätänen, R. (1992). *Attention and Brain Function*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Navarra, J., & Soto-Faraco, S. (2005). Hearing lips in a second language: Visual articulatory information enables the perception of second language sounds. *Psychological Research, 71*(1), 4–12.

Oldfield, R. C. (1971). The Assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9*, 97–113.

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, L. D., Dorsi, J., & Dias, J. W. (2016). The impact and status of Carol Fowler's supramodal theory of multisensory speech perception. *Ecological Psychology, 28*(4), 262–294.

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

Sams, M., Aulanko, R., Hämäläinen, M., Hari, R., Lounasmaa, O. V., Lu, S. T., et al. (1991). Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience Letters, 127*(1), 141–145.

Sato, M. (2022a). The timing of visual speech modulates auditory neural processing. *Brain and Language, 235*, Article 105196.

Sato, M. (2022b). Motor and visual influences on auditory neural processing during speaking and listening. *Cortex, 152*, 21–35.

Scherg, M., & VonCramon, D. (1986). Evoked dipole source potentials of the human auditory cortex. *Electroencephalography and Clinical Neurophyiology, 65*, 344–360.

Schwartz, J. L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition, 93*, 69–78.

Stekelenburg, J. J., Keetels, M., & Vroomen, J. (2018). Multisensory integration of speech sounds with letters vs.visual speech: Only visual speech induces the mismatch negativity. *European Journal of Neuroscience, 47*, 1135–1145.

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience, 19*, 1964–1973.

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

Tanner, D., Morgan-Short, K., & Luck, S. J. (2015). How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. *Psychophysiology, 52*(8), 997–1009.

Teder-Sälejärvi, W. A., McDonald, J. J., Di Russo, F., & Hillyard, S. A. (2002). An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings. *Cognitive Brain Research, 14*(1), 106–114.

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., Kandel, S., & Sato, M. (2017). Electrophysiological evidence for a self-processing advantage during audiovisual speech integration. *Experimental Brain Research, 235*(9), 2867–2876.

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. *Frontiers in Psychology, 5*, 420.

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*, Article 107949.
van Wassenhove, V. (2013). Speech through ears and eyes: Interfacing the senses with the supramodal brain. *Frontiers in Psychology, 4*, 1–17.
van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences, 102*, 1181–1186.
Vroomen, J., & Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *Journal of Cognitive Neuroscience, 22*, 1583–1596.