Motor and visual influences on auditory neural processing during speaking and listening

Marc Sato

MOTOR AND VISUAL INFLUENCES ON AUDITORY NEURAL PROCESSING DURING SPEAKING AND LISTENING

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.

ACKNOWLEDGMENTS

The author thanks two anonymous reviewers for helpful comments and suggestions on an earlier draft of the manuscript. No part of the study procedures or analyses was pre-registered prior to the research being conducted. We report how we determined our sample size, all data exclusions (if any), all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study. Data and analyses pipelines are available in the project OSF repository at: https://osf.io/42qpz/.

Number of pages: 27, Figures: 4, Tables: 3

ABSTRACT

During speaking or listening, endogenous motor or exogenous visual processes have been shown to fine-tune the auditory neural processing of incoming acoustic speech signal. To compare the impact of these cross-modal effects on auditory evoked responses, two sets of speech production and perception tasks were contrasted using EEG. In a first set, participants produced vowels in a self-paced manner while listening to their auditory feedback. Following the production task, they passively listened to the entire recorded speech sequence. In a second set, the procedure was identical except that participants also watched online their own articulatory movements. While both endogenous motor and exogenous visual processes fine-tuned auditory neural processing, these cross-modal effects were found to act differentially on the amplitude and latency of auditory evoked responses. A reduced amplitude was observed on auditory evoked responses during speaking compared to listening, irrespective of the auditory or audiovisual feedback. Adding orofacial visual movements to the acoustic speech signal also speeded up the latency of auditory evoked responses, irrespective of the perception or production task. Taken together, these results suggest distinct motor and visual influences on auditory neural processing, possibly through different neural gating and predictive mechanisms.

KEYWORDS

Speech production, audiovisual speech perception, speaking-induced suppression, efference copy, corollary discharge, readiness potential, EEG.

INTRODUCTION

In the animal kingdom, the nervous system keeps track of motor commands and informs the sensory processing stream about movements that are to be produced by means of efference copy (that is, an internal copy of an efferent motor command; von Holst and Mittlestaedt 1950) and corollary discharge (that is, the expected sensation resulting from the motor command; Sperry, 1950). Across species and sensory domains, efference copy and corollary discharge take the form of suppressed sensory responses to self-generated action, thought to reflect a partial neural cancellation of the incoming sensory feedback (Crapse and Sommer, 2008; Straka et al., 2018). In the speech domain, suppressed auditory evoked responses to self-generated speech feedback, when compared with playback of the same speech signal, has been repeatedly observed using electroencephalography (EEG; Ford et al. 2001; Ford and Mathalon 2004; Heinks-Maldonado et al., 2005; Behroozmand and Larson, 2011; Sitek et al., 2013; Wang et al., 2014; Sato and Shiller, 2018), magnetoencephalography (MEG; Numminen and Curio, 1999; Numminen et al., 1999; Curio et al., 2000; Houde et al., 2002; Ventura et al., 2009; Niziolek et al., 2013; Franken et al., 2015) and direct cortical recordings (Creutzfeldt et al., 1989; Flinker et al., 2010; Chen et al., 2011; Chang et al., 2013). The so-called speaking-induced suppression (SIS) is observed on N1/M100 auditory evoked potentials (AEPs), with their sources mainly originating from the supratemporal plane of the auditory cortex in response to temporal, spectral and phonetic cues of an auditory stimulation (Näätänen and Picton, 1987; Woods, 1995). SIS is also thought to reflect the computation of an error signal, allowing talkers to adjust their speech motor output toward the auditory sensory target when the expected and actual auditory feedback do not match. SIS indeed appears reduced or even abolished in cases of online auditory feedback perturbation and associated compensatory vocal responses (Houde et al., 2002; Heinks-Maldonado et al., 2005; Behroozmand and Larson, 2011; Chang et al., 2013).

Although SIS is most often interpreted as a consequence of efference copy and corollary discharge acting on the auditory neural processing of incoming speech sounds, direct evidence linking auditory neural suppression to motor cortex activity during speaking is sparse (Chen et al., 2011; Chang et al., 2013; Wang et al., 2014). From this question, using EEG with anatomical MRI to facilitate source localization, Wang and colleagues (2014) showed that, 300 ms prior to speaking, movement-related cortical potentials (MRCPs) and premotor activity in the inferior frontal gyrus activity were associated with a reduced N1 amplitude, 100 ms following speech onset. Associated with a motor task and related to movement planning and execution, MRCPs are characterized by a slow negative deflection on fronto-central sites around 1000ms prior to the onset of a self-paced movement (i.e. Readiness Potential, RP, or Bereitschaftspotential, BP; e.g., Kornhuber and Deecke, 1965; Libet et al., 1983), reaching the maximum negativity near movement onset, and followed by a positive rebound (Birbaumer et al., 1990; Pereira et al., 2017). For the authors, the observed pre-speech activity in the inferior frontal gyrus likely reveals ”the accumulation and coordination

of neural computations related to action planning and preparing sensory systems for their expected consequences" (Wang et al., 2014).

The above-mentioned studies argue for a key role of endogenous motor-to-auditory cross-modal effects in speech motor control, facilitating the auditory neural processing of acoustic speech feedback. In the perceptual domain, a rich literature also demonstrates the impact of exogeneous visual-to-auditory cross-modal effects during audiovisual speech perception. It has been consistently shown that adding lip movements to auditory speech modulates activity early in the supratemporal auditory cortex, with the latency and amplitude of N1/M100 AEPs attenuated and speeded up during audiovisual compared to unimodal speech perception (Klucharev et al., 2003; Besle et al., 2004; van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007; Arnal et al., 2009; Huhn et al., 2009; Pilling, 2009; Vroomen and Stekelenburg, 2010; Winneke and Phillips, 2011; Frtusova et al., 2013; Schepers et al., 2013; Stekelenburg et al., 2013; Baart et al., 2014; Ganesh et al., 2014; Kaganovich and Schumaker, 2014; Treille et al., 2014a, 2014b, 2017, 2018; Baart and Samuel, 2015; Hisanaga et al., 2016; Paris et al., 2016; Pinto et al., 2019; for reviews, see van Wassenhove, 2013; Baart, 2016). Like SIS, visually induced suppression (VIS) is thought to help tuning auditory processing to the incoming speech sound, based on the available information from the speaker's articulatory movements that precede sound onset in these studies (Chandrasekaran et al., 2009; see also Schwartz and Savariaux, 2014). In addition, while SIS is thought to reflect the computation of an error signal when the expected and actual auditory feedback do not match, VIS has also been hypothesized to function as an error signal in case of a mismatch between visual and auditory inputs (Hertrich et al., 2007; Arnal et al., 2009).

Though motor-to-auditory and visual-to-auditory cross-modal effects appear clearly distinct by nature (from their sources, their underlying neural pathways, their temporal onsets and time-courses), ultimately, their common goal may be viewed as the fine-tuning of the sensory processing of endogenous and exogenous events to enhance perception. The goal of the present EEG study was to compare the impact of these motor-to-auditory and visual-to-auditory cross-modal effects on auditory neural processing and to determine whether visual feedback of one's own articulators during speaking further enhances auditory neural processing by reducing uncertainty of acoustic speech feedback (see Figure 1). To this aim, two sets of speech production and perception tasks were contrasted, leading to a two-by-two factorial design. In a first set, participants produced vowels in a self-paced manner while listening to their auditory feedback through earphones. Following the production task, they passively listened to the entire recorded speech sequence in a manner that was identical in timing and amplitude to the auditory feedback provided during the preceding production task. In a second set, the procedure was identical except that participants also watched their own articulatory movements displayed online on a computer screen during both the production and perception tasks. MRCPs, N1 and P2 amplitudes and latencies were computed in each condition to compare the impact of motor-to-auditory and visual-to-auditory cross-modal effects on auditory neural processing.

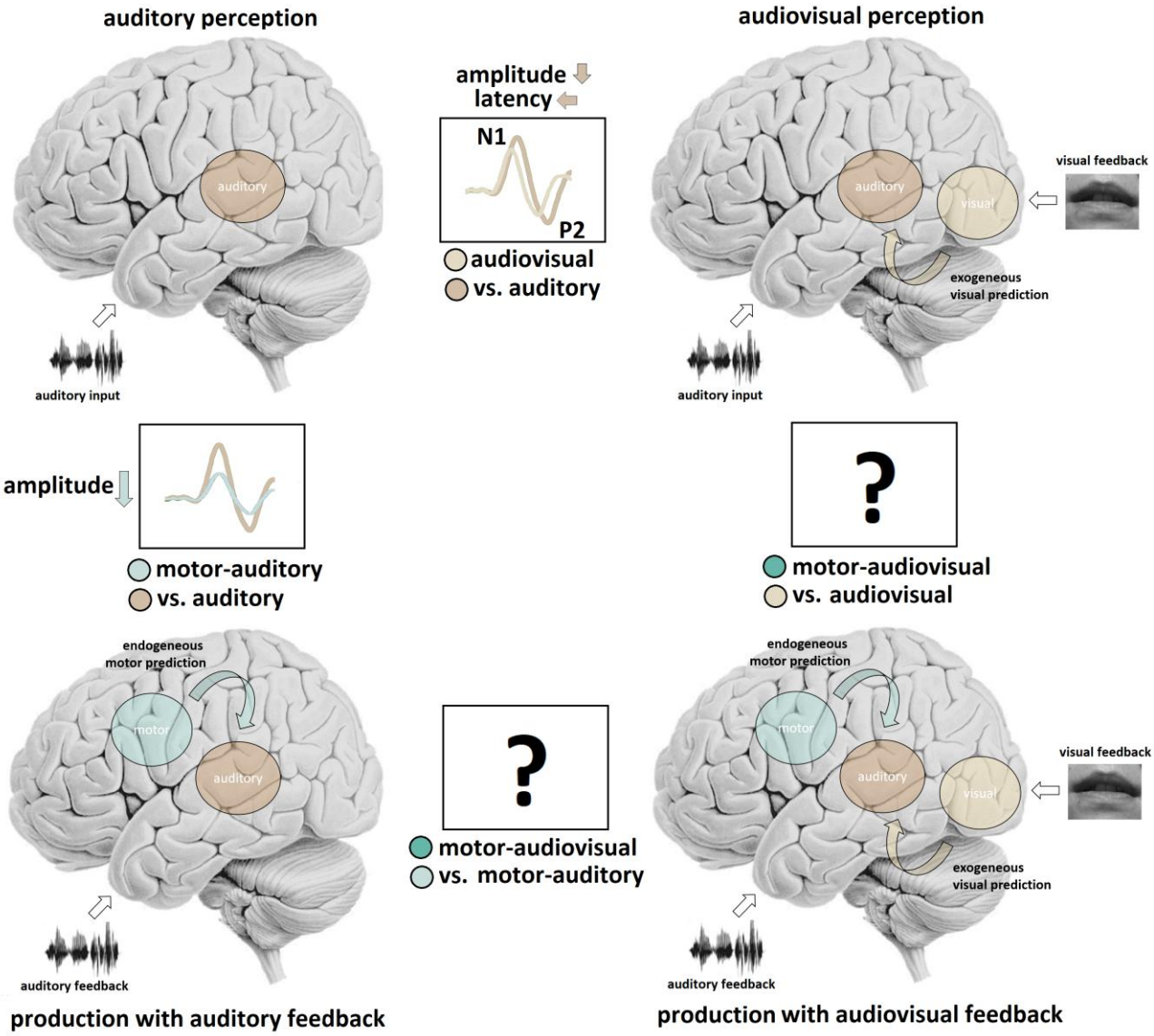

Figure 1. Schematic of motor-to-auditory and visual-to-auditory cross-modal effects during speech perception and production, and their known effect on the latency and amplitude of N1 and P2 auditory-evoked potentials. In past studies, a strongly reduced auditory response has been repeatedly observed during speech production compared to auditory speech perception, while an earlier and slightly reduced response has been observed during audiovisual compared to auditory speech perception. The goal of the present EEG study was to compare these effects during speech production with audiovisual feedback.

METHODS

Participants

Twenty healthy adults (12 females and 8 males), with a mean age of 27 years (±6 SD, range: 20-39 years), participated in the study after giving informed consent. All participants were native French speakers, with an average of 16 years of education (±2 SD, range: 11-20 years). All were right-handed according to the standard handedness inventory (Oldfield, 1971) with a mean score of 86% (±15 SD, range: 56-100 %), had normal or corrected-to-normal vision, and self-reported no history of hearing, speaking, language, neurological and/or neuropsychological disorders. The cognitive functioning of all participants was evaluated using the Montreal Cognitive Assessment scale (MoCA; Nasreddine et al., 2003, 2005), with a mean score of 29/30 (±1 SD, range: 26-30). The protocol was carried out in accordance with the ethical standards of the 1964 Declaration of Helsinki. Participants were compensated for the time spent in the study.

Experimental procedure

The experimental protocol was adapted from a well-defined vocal production and perception EEG protocol for studying corollary discharge (Ford et al., 2010). The experiment was carried out in a dimly lit sound-attenuated room and consisted of two consecutive sets of speech production and perception tasks. In one set, subjects first produced vowels in a self-paced manner for 3 minutes, while listening to their auditory feedback through earphones (motor with auditory feedback task, M-A). Following the production task, subjects passively listened to the entire recorded speech sequence in a manner that was identical in timing and amplitude to the auditory feedback provided during the preceding production task (auditory task, A). In another set, the procedure was identical except that participants also watched their own articulatory movements displayed online on a computer screen during both the production (motor with audiovisual feedback task, M-AV) and perception (audiovisual task, AV) tasks.

Regarding speech stimuli, although a number of previous studies on corollary discharge used a single vowel to limit articulatory artefacts on the EEG signal (e.g., /a/ vowel: Ford et al., 2010; Sitek et al., 2013; Wang et al., 2014), the three /a/, /ø/ and /e/ vowels were here selected to limit adaptation effects and to provide a more extended pattern of lip and jaw articulatory configurations and visual saliency. The three vowels differed in terms of height and/or roundedness phonetic features: the /a/ vowel being produced with the jaw opened and the lips unrounded, while the /ø/ vowel being produced with the jaw mid-opened and the lips rounded, and the /e/ vowel being produced with the jaw mid-opened and the lips unrounded and stretched back. Due to these distinct articulatory configurations, the three vowels were highly distinguishable from each other when produced in isolation (for examples, see Figure 2).

For the production tasks (M-A, M-AV), participants were asked to randomly produce one vowel at a time every 1-2s until asked to stop (i.e., after 3 minutes). In order to limit adaptation effects, they were also

asked not to produce the same vowel consecutively (e.g., /a/-/a/) and not to produce the same serie of three vowels through the entire sequence (e.g. /a/-/ø/-/e/-/a/-/ø/-/e/...). After being familiarized with EEG muscle artefacts (eye movements, eye blinks, articulatory movements), participants were asked to produce vowels in a natural manner but with minimal force/tension in the lip and jaw muscles. This was aided further by the instruction to produce vowels with a constant/natural intensity and duration, as well as to maintain a visually-neutral open mouth posture between each vowel. For the production task with audiovisual feedback (M-AV), they were also asked to carefully watch their articulatory movements during the entire sequence. For the perception tasks (A, AV), participants were asked to passively listen to (A) or to to passively listen to and watch (AV) the entire previously recorded speech sequence.

During the training session, prior to data collection, each subject practiced producing vowels to ensure that no visible artifact was present in the EEG signal and to confirm that they understood the production tasks. The order of the two sets of speech production and perception tasks was fully counterbalanced across participants (i.e., half of the participants first performed the production and perception tasks without visual feedback (M-A, A) while, in a second set, they performed the production and perception tasks with visual feedback (M-AV, AV)), and short breaks were offered between tasks.

Acoustic and visual setup

In the speech production tasks (without or with visual feedback, M-A and M-AV), all participants' productions were recorded using a microphone (NTG-2, Røde, Sydney, Australia) located approximately 25 cm from the mouth, with audio digitizing done at 48 kHz. In order to minimize the effects of bone conduction, the acoustic signal level played back through earphones (T205, JBL, Northridge, USA) was 10 dB greater than the input signal at the microphone (calibrated prior to testing using a 1000 Hz pure-tone). In addition, in the speech production task with visual feedback (M-AV), all participants' articulatory movements were recorded using a digital video camera (C922, Logitech, Lausanne, Switzerland) located approximately 50 cm from the head, with video digitizing done at 30 frames per second with a resolution of 1080 × 1920 pixels. The digital video camera was centered on the participant's full face (for examples, see Figure 2), with the visual signal horizontally flipped as a mirror image and played back on a 20 inch LCD monitor (E2009, DELL, Round Rock, USA). In the two speech perception tasks (without or with visual display, A and AV), the acoustic and visual signal recording and playback system was configured to ensure that both the auditory and visual presentation of stimuli during passive perception tasks was identical to that during live-feedback production tasks. The Capture software (Logitech, Lausanne, Switzerland; version 1.10) was used to record/digitize participant's productions and to control acoustic and visual presentation. The microphone and earphones were connected to a computer (Zbook 15 Workstation, Hewlett-Packard, Palo Alto, USA) equipped with 32 GB RAM and a 1 GB graphics card (K610m, Nvidia, Santa Clara, USA) through a USB audio interface (iO2, Alesis, Cumberland, USA). In addition, in all tasks, the acoustic signal delivered to the earphones was duplicated and sent to the EEG Biosemi system equipped with an auxiliar

connector for isolated sensor and synchronized with EEG recordings to determine offline the acoustical triggers for the EEG analyses (see below).

From the acoustic and visual setup, it should be noted that possible latency delays between production and auditory/visual feedback were not measured. For auditory feedback, since a purely hardware loop was used, a near-to-zero latency is quite likely. However, for visual feedback, the question of a visual-to-auditory lag remains. It should be noted, however, that no participant reported a visual delay, and the observed results, showing classical cross-modal interactions (i.e., shorter N1 latency observed for the visual modality during both the production and perception tasks), indirectly argue in favor of a minimal delay.

EEG setup

In all tasks, 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 as recommended by Ford et al., (2010), 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 experiment, the impedance of all electrodes was adjusted to get low offset voltages and stable DC.

Analyses

All statistical analyses were computed using the Statistica software (Statsoft, Tulsa, USA; version 10). 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, all post hoc analyses were conducted with Bonferroni corrections as implemented in the Statistica software.

Visual signal

Inspection of the visual signal in the production task with audiovisual feedback (M-AV) confirmed that all participants correctly performed the task, watching their own articulatory movements through the entire production sequence and maintaining a neutral opened mouth posture between vowels.

Acoustic signal

Acoustic analyses were performed using the Praat software (Boersma and Weenink, 2013; version 5.3). For each participant and each production task (M-A, M-AV), a hybrid semi-manual procedure was first performed to determine the onset and offset of the recorded vowels (~5000 vowels). Using the Speech

Corpus Toolkit for Praat (Lennes, 2017; version 1.0), pauses between each vowel were automatically identified, based on minimal duration and low intensity energy parameters, with the vowel's boundaries set on that basis. All boundaries were then fine-tuned manually based on waveform and spectrogram information. Likely due to the neutral open mouth position between vowels, some occurrences began with a short "whispering" period (~10-20 ms) in which exhaled air passed directly through the restricted but open larynx, without vibration of the vocal folds. These unvoiced periods were characterized by the absence of fundamental frequency but typically exhibited 'formant-like' features (e.g., Fant, 1960; Thomas, 1969). This phenomenon also appeared at the end of vowel production. In some other cases, vowel onset occurred with transient glottal attacks. To minimize inter- and intra-variability, all vowel onsets and offsets were therefore strictly defined according to a continuous voicing period, without pause, based on the lowest frequency part of the wide band spectrogram (i.e., < 300-400 Hz; see Figure S1 in Supplementary Material). All vowels were then listened to and labeled. Low quality vowels (e.g., including hesitation, transient silent phonatory period, diphtongue) and/or including acoustic/electrical noise were removed from the acoustic and EEG analyses (on average, 5.5 % (±3 SD) and 4.7% (±3 SD) in the M-A and M-AV tasks, without significant difference between the tasks, $F(1,19) = .64$). Vowel onsets were saved as triggers, which were later used for EEG analysis (with vowel onsets matched with the acoustic signal recorded in the analog channel of EEG data; see Ford et al., 2010).

For each vowel, in order to select a stable, artefact-free period, the maximum peak intensity was calculated using parabolic interpolation. The fundamental frequency (f_0), F_1, F_2 and F_3 formant frequencies and intensity were averaged from a period defined as ±25 ms of the maximum peak intensity (Duckworth et al., 2011; see also Kent and Vorperian, 2018). f_0 was estimated using an autocorrelation procedure with a pitch range of 150-300 Hz for females and 75-200 Hz for males. F_1, F_2 and F_3 were estimated using LPC analysis (Linear Predictive Coding, Burg method), with LPC parameters adjusted on a per-subject basis in order to avoid/minimize the occurrence of spurious formant values. The intensity was computed using the mean energy averaging method.

For each participant, each task and each vowel, the number of occurences, the number of repetitions (i.e., the same vowel produced consecutively), the median intertrial duration, vowel duration, intensity, f_0, F_1, F_2, F_3 were calculated. To evaluate possible intra-individual variability difference between the two production tasks, SEM was also computed on the intertrial duration, vowel duration, intensity, f_0, F_1, F_2, F_3. Finally, the F_1-F_2-F_3 triangular /a/-/ø/-/e/ vowel space area (defined as the Pythagorean sum of the areas of the respective projections on the three principal planes) was calculated, as a quantitative index of articulatory working space (for a review, see Kent and Vorperian, 2018).

Two-way repeated-measures ANOVAs were performed separately on these measures with the task (M-A, M-AV) and the vowel (/a/, /ø/, /e/) as within-participant factors. In addition, a one-way repeated-

measures ANOVA was performed on F_1-F_2-F_3 triangular vowel space area with the task (M-A, M-AV) as a within-participant factor.

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 and each task (M-A, M-AV, A, AV), 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, speech-related 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.

To evaluate MRCPs and taking account their influence on AEPs, two analyses based on a distinct epoching procedure were performed. A first analysis was designed to evaluate N1/P2 AEPs considering/subtracting the temporally contingent influence of MRCPs on AEPS. To this aim, EEG data were segmented from −100 ms to 300 ms relative to the acoustic onset and corrected from a -100 ms to 0 ms baseline. A second analysis was designed to calculate N1/P2 AEPs in relation to a baseline supposed to be stable/equal for all tasks, as well as to further determine the time-course of MRCPs. Since RP has been shown to occur approximately 300 ms before vowel production (Wang et al., 2014), EEG data were here segmented from −1000 ms to 300 ms relative to the acoustic onset and corrected from a -1000 ms to -900 ms baseline.

First epoching procedure [-100 ms to 300 ms]

EEG data from /a/, /ø/ and /e/ vowels were averaged together (due to an insufficient number of trials per vowel for reliable EEG analyses) and segmented into 400 ms epochs (from −100 ms to 300 ms relative to the acoustic onset), corrected from a -100 ms to 0 ms baseline. 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 22% of epochs and left 93 epochs per task (for details, see Table 2).

As expected, visual inspection of EEG signals showed strongly reduced N1/P2 AEPs in the production compared to the perception tasks, but with their peaks sometimes ambiguous to detect. An individual peak detection procedure was therefore designed to avoid/minimize the detection of spurious N1/P2 peak values. For each participant, N1/P2 amplitude and latency were first computed on the EEG waveform

averaged over the four tasks, from a fixed temporal window of 40-120 ms for N1 and of 120-240 ms for P2. Clear and homogeneous N1 and P2 AEPs were observed for all but two participants, who were removed from the EEG analyses. For one of these two participants, no N1/P2 AEPs were observed in the EEG waveform, while, for the second participant, both the latency and amplitude were ±2 SD away from the mean (see Figure 3). On the remaining 18 participants, for each participant and each task, N1 and P2 amplitudes and latencies were automatically computed based on two fixed temporal windows defined as ±30 ms of the N1 and P2 peak latencies previously calculated from the individual participant waveform averaged over the four tasks (Ganesh et al., 2014; Treille et al., 2014b).

Two-way repeated-measures ANOVAs were performed separately on the number of rejected trials and on N1 and P2 amplitudes and latencies with the task modality (perception, production) and the sensory modality (auditory-only, audiovisual) as within-participant factors.

Second epoching procedure [-1000 ms to 300 ms]

As in the first analysis, EEG data from /a/, /ø/ and /e/ vowels were averaged together but here segmented into 1300 ms epochs (from −1000 ms to 300 ms relative to the acoustic onset), corrected from a -1000 ms to -900 ms baseline. 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 33% of epochs and left 80 epochs per task (for details, see Table 3).

For each participant and each task, the mean amplitude of the successive 100 ms periods from -900 ms to 0 ms prior to the acoustic onset were calculated to evaluate the time course of MRCPs. As previously, N1 and P2 amplitudes and latencies were computed based on two fixed temporal windows defined as ±30 ms of the N1 and P2 peak latencies calculated from the individual participant waveform averaged over the four tasks.

A three-way repeated-measures ANOVA was performed on the time course of MRCPs with the period ([-900 ms to -800 ms]…[-100 ms to 0 ms]), the task modality (perception, production) and the sensory modality (auditory-only, audiovisual) as within-participant factors. Two-way repeated-measures ANOVAs were performed separately on N1 and P2 amplitudes and latencies, with the task modality (perception, production) and the sensory modality (auditory-only, audiovisual) as within-participant factors.

RESULTS

Acoustic results (See Figure 2 and Table 1)

A first set of analyses confirmed that the two production tasks were correctly performed by the participants. First, a homogeneous distribution of /a/, /ø/ and /e/ vowels was observed in the two production tasks. On average, 40 occurrences were produced per vowel, without significant difference between the tasks ($F(1,19) = .3$) and the vowels ($F(2,38) = .7$), and no interaction ($F(2,38) = 1.3$). Furthermore, only one repetition was observed on average for each vowel, without significant difference between the tasks ($F(1,19) = .6$) and the vowels ($F(2,38) = .3$), and no interaction ($F(2,38) = 1.7$). The mean intertrial duration was 1454 ms without significant difference between the tasks ($F(1,19) = 1.4$), the vowels ($F(2,38) = 1.0$), and no interaction ($F(2,38) = .7$). Regarding vowel duration, although no difference was observed between the tasks ($F(1,19) = 1.1$), a significant vowel effect was observed ($F(2,38) = 17.7$, $p < .0001$, $p\eta^2 = .48$). Post hoc analyses showed that /a/ was significantly shorter than /e/, and /e/ shorter than /ø/ (on average, 211 ms vs. 218 ms. vs. 224 ms, respectively). In addition, a modest but significant interaction between task and vowel was found ($F(2,38) = 3.9$, $p = .03$, $p\eta^2 = .17$). Post hoc analyses showed that /a/ and /ø/, but not /e/, were longer in M-A compared to M-AV task (on average, 216 ms vs. 206 ms for /a/, 227 ms vs. 221 ms for /ø/, 220 ms vs. 216 ms for /e/).

	M-A			M-AV		
	/a/	**/ø/**	**/e/**	**/a/**	**/ø/**	**/e/**
Number of occurrences	39 (3)	41 (3)	40 (3)	41 (2)	41 (2)	41 (2)
Number of repetitions	1 (1)	1 (0)	1 (1)	1 (1)	1 (0)	1 (0)
Speech rate						
Intertrial interval (ms)	1470 (103)	1494 (107)	1483 (103)	1422 (98)	1428 (97)	1427 (99)
Vowel duration (ms)	216 (12)	227 (12)	220 (12)	206 (11)	221 (12)	216 (11)
Acoustic values						
Intensity (dB)	69 (1)	70 (1)	70 (1)	70 (1)	71 (1)	71 (1)
f_0 (Hz)	171 (10)	174 (11)	174 (11)	173 (11)	176 (11)	176 (11)
F_1 (Hz)	704 (32)	406 (17)	399 (16)	703 (27)	425 (21)	409 (18)
F_2 (Hz)	1388 (34)	1545 (35)	2208 (53)	1389 (34)	1546 (39)	2221 (49)
F_3 (Hz)	2702 (55)	2545 (47)	2893 (41)	2705 (54)	2550 (45)	2901 (40)
F_1-F_2-F_3 vowel space area (Hz^2)	151449 (17558)			144821 (16477)		
Individual variabilities						
Intertrial interval (ms)	46 (7)	45 (7)	40 (5)	53 (8)	51 (7)	48 (6)
Vowel duration (ms)	4 (0)	4 (0)	4 (0)	4 (0)	4 (0)	4 (0)
Intensity (dB)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
f_0 (Hz)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
F_1 (Hz)	12 (2)	5 (1)	4 (1)	10 (1)	5 (1)	4 (1)
F_2 (Hz)	14 (2)	16 (2)	23 (7)	13 (2)	16 (2)	19 (4)
F_3 (Hz)	22 (3)	22 (3)	20 (3)	21 (4)	22 (3)	18 (2)

Table 1. Mean vocal behaviors, acoustic values, and individual variabilities for /a/, /ø/ and /e/ vowels in the two production tasks with auditory and audiovisual feedback (M-A and M-AV; based on ~5000 occurrences; SEM are indicated).

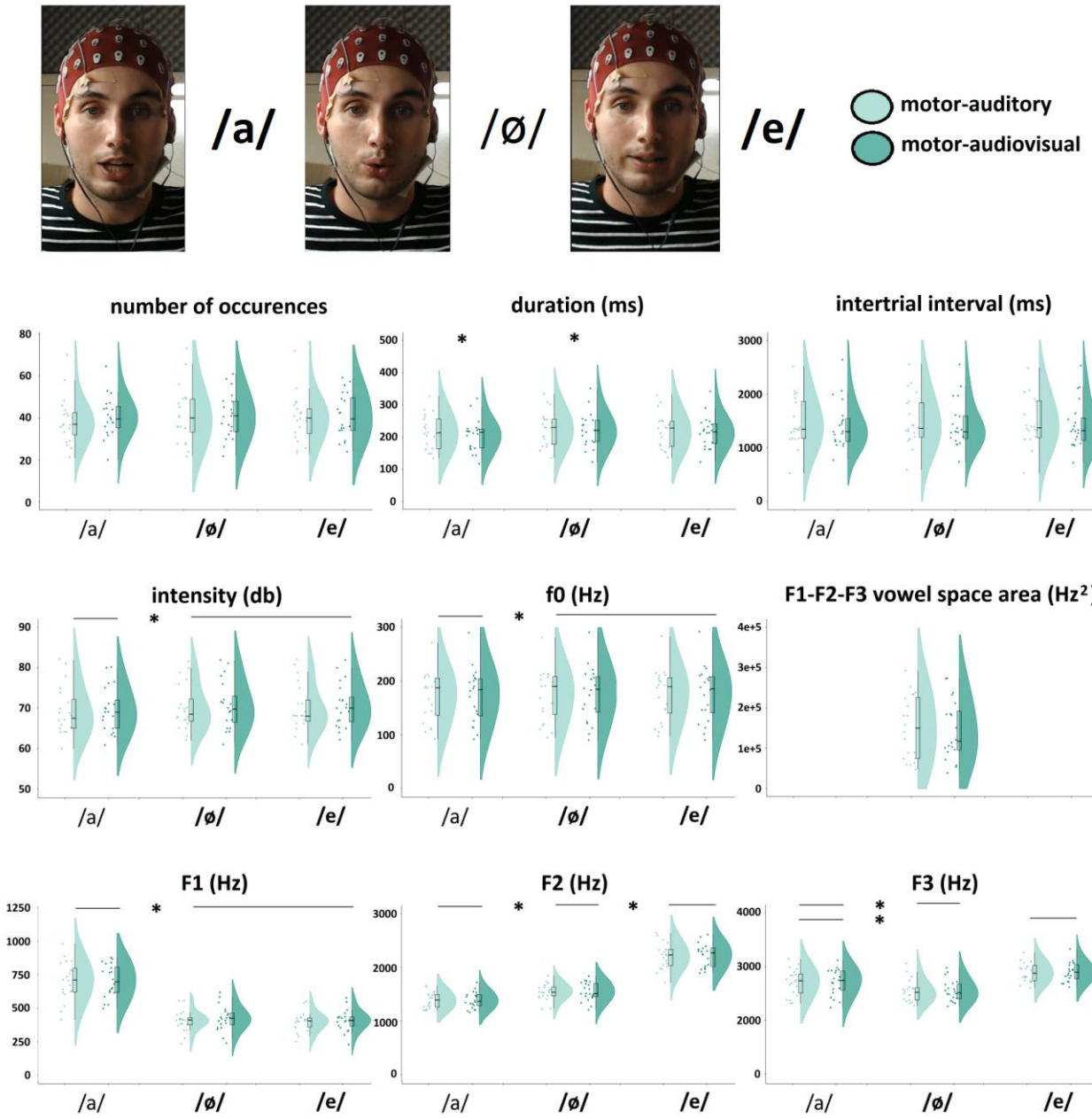

Figure 2. Mean vocal behaviors and acoustic values for /a/, /ø/ and /e/ vowels in the two production tasks with auditory and audiovisual feedback (M-A and M-AV; based on ~5000 occurrences). In the violin/whisker plots, the line across the box represents the median and the vertical bars represent the interquartile range. In each plot, individual data are shown as dots (n = 20). Significant contrasts are indicated.

A second set of analyses did not provide any evidence of acoustic differences between the two tasks. Regarding the intensity, no difference was observed between the tasks ($F(1,19) = 2.4$) and there was no task x vowel interaction ($F(2,38) = .4$). However, the intensity differed between vowels ($F(2,28) = 9.6$, $p < .0001$, $p\eta^2 = .34$). Post hoc analyses showed a lower intensity for /a/ compared to /ø/ and /e/ (on average, 69 dB vs. 70 dB vs. 70 dB, respectively). The same pattern was observed for f_0. No difference was found between the tasks ($F(1,19) = 2.7$) and there was no interaction ($F(2,38) = .5$). However, a vowel effect was found ($F(2,38) = 20.0$, $p < .0001$, $p\eta^2 = .51$). Post hoc analyses showed a lower f_0 for /a/ compared to /ø/ and /e/ (on average, 172 Hz vs. 175 Hz vs. 175 Hz, respectively). Regarding the formants, strong differences between vowels were observed. For F_1, no difference was found between the tasks ($F(1,19) = 2.4$) and

there was no task x vowel interaction (F(2,38) = .9). However, a vowel effect was observed (F(2,38) = 184.8, $p < .0001$, $p\eta^2 = .91$). Post hoc analyses showed a higher value for /a/ compared to /ø/ and /e/ (on average, 704 Hz vs. 415 HZ vs. 404 Hz, respectively). For F_2, no difference was found between the tasks (F(1,19) = 1.0) and there was no interaction (F(2,38) = .2). However, a vowel affect was found (F(2,38) = 209.5, $p < .0001$, $p\eta^2 = .92$). Post hoc analyses showed a lower value for /a/ compared to /ø/, and for /ø/ compared to /e/ (on average, 1389 Hz vs. 1545 Hz vs. 2215 Hz, respectively). For F_3, no difference was found between the tasks (F(1,19) = .2) and there was no interaction (F(2,38) = .0). However, a vowel effect was observed (F(2,38) = 80.0, $p < .0001$, $p\eta^2 = .81$). Post hoc analyses showed a lower value for /ø/ compared to /a/, and for /a/ compared to /e/ (on average, 2547 Hz vs. 2704 Hz vs. 2897 Hz, respectively). Finally, no difference was observed between the tasks for F_1-F_2-F_3 vowel space area (F(1,19) = .2), with a mean value of 148135 Hz^2.

Finally, analyses on individual variability showed that participants' productions were constant across tasks and vowels. No effect of the task (all F's < 3.1) and the vowel (all F's < 1.9), and no interaction (all F's < 1.6) were found for the intertrial interval, vowel duration, intensity, f_0, F_2 and F_3. A significant difference between vowels was only observed for F_1 (F(2,38) = 18.8, $p < .0001$, $p\eta^2 = .50$), although no difference was found between the tasks (F(1,19) = 2.2) and there was no interaction (F(2,38) = .1). Post hoc analyses showed a higher variability for /a/ compared to /ø/ and /e/ (on average, 11 Hz vs. 5 Hz vs. 4 Hz, respectively).

In sum, the two production tasks were correctly performed, with the expected distribution of f_0 and formant values for /a/, /ø/ and /e/ vowels (see Discussion). Crucially, apart from a modest but significant task x vowel interaction on vowel duration, no difference was observed when producing vowels with or without visual feedback.

EEG results

First epoching procedure [-100 ms to 300 ms] (See Figure 3 and Table 2)

A similar number of trials was observed across tasks, with on average 119 trials per task (F(1,17) = .5). As expected, likely due to articulatory movements, the EEG signal included more artefact and a higher number of rejected trials in the production compared to the perception tasks (on average, 33 vs. 19, respectively; F(1,17) = 6.7, p = .02, $p\eta^2 = .28$). There was no difference between the sensory modalities (F(1,17) = .3) and no interaction (F(1, 17) = .1).

Regarding AEP amplitudes, N1 response was reduced in the production compared to the perception tasks (on average, -3.43 µV vs. -5.20 µV, respectively; F(1,17) = 10.7, p = .004, $p\eta^2 = .39$). No difference was found between the sensory modalities (F(1,17) = .1) and there was no interaction (F(1, 17) = .7). For P2, the mean amplitude was 2.61 µV, without significant difference between the perception and the production tasks (F(1,17) = 3.3), the sensory modalities (F(1, 17) = .8) and no interaction (F(1,17) = .0).

As for AEP latencies, a shorter latency was observed on N1 for the audiovisual compared to the auditory modalities (on average, 78 ms vs. 84 ms, respectively; $F(1,17) = 7.2$, $p = .02$, $p\eta^2 = .30$). No difference was found between the production and perception tasks ($F(1,17) = 1.1$) and there was no interaction ($F(1, 17) = 2.3$). For P2, the mean latency was 165 ms, without difference between the perception and the production tasks ($F(1,17) = .6$), the sensory modalities ($F(1, 17) = 3.5$), and no interaction ($F(1,17) = .3$). It is to note that the latency difference on P2 between the two production tasks appears more pronounced on the average waveform than on the mean individual latency (see the blue lines on Figure 3). This is explained by a large inter-individual variability of P2 on these production tasks (see Table 2).

In sum, a SIS effect was observed on N1 amplitude, with a reduced response in the production compared to the perception tasks, irrespective of the sensory modality. In addition, the visual modality was found to speed up N1 latency, with a shorter latency for the visual compared to auditory modalities, irrespective of the task modality. However, these two effects were not found to interact.

	M-A	M-AV	A	AV
Number of trials	117 (8)	121 (7)	117 (8)	121 (7)
Number of rejected trials	34 (6)	33 (4)	20 (3)	18 (4)
Amplitudes (µV)				
N1	-3.35 (0.43)	-3.50 (0.43)	-5.37 (0.59)	-5.04 (0.54)
P2	2.15 (0.53)	1.83 (0.51)	3.42 (0.63)	3.04 (0.50)
Latencies (ms)				
N1	84 (3)	75 (3)	83 (3)	81 (3)
P2	165 (6)	161 (6)	170 (3)	163 (4)

Table 2: Mean N1/P2 amplitudes and latencies on fronto-central electrodes in the two production and perception tasks from the [-100 ms to 300 ms] epoching procedure (M-A, M-AV, A, AV; SEM are indicated).

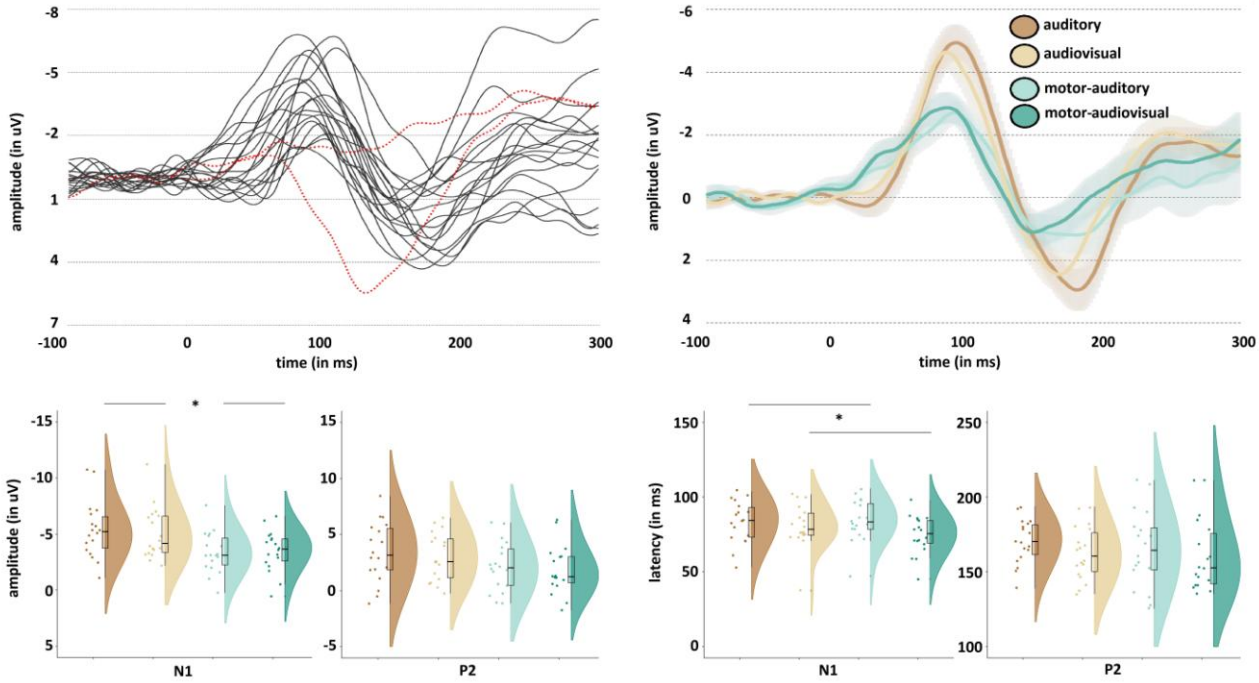

Figure 3. Left: Individual participant EEG waveforms on fronto-central electrodes averaged over the four tasks from the [-100 ms to 300 ms] epoching procedure (M-A, M-AV, A, AV; the red lines represented the two removed participants due to abnormal signal). Right: Average EEG waveform for each task (SEM are indicated in shadow). Bottom: Mean N1 and P2 AEP amplitudes and latencies. In the violin/whisker plots, the line across the box represents the median and the vertical bars represent the interquartile range. In each plot, individual data are shown as dots (n = 18). Significant contrasts are indicated.

Correlation between N1 latency and acoustic changes in response to the visual modality?

Previous EEG studies demonstrated that early neural auditory processing of vowels, notably indexed by N1 amplitude and latency, partly relies on their acoustic properties (i.e., fundamental and formant frequencies; Frank et al., 2020). Given the significant effect of the visual modality observed on N1 latency, linear regression analyses were therefore performed to determine whether these N1 latency changes were driven by (non-significant) acoustic differences between the audiovisual and auditory modalities. More specifically, linear regression analyses were performed separately on the perception and production tasks to test whether N1 latency change in response to the audiovisual compared to auditory modalities (i.e., individual latency differences between AV vs. A and between M-AV vs. M-A) correlated with the (non-significant) acoustic changes observed on the intertrial duration, vowel duration, intensity, f_0, F_1, F_2, F_3 and F_1-F_2-F_3 vowel space area (i.e., individual acoustic differences between M-AV vs. M-A).

For the perception tasks, no correlation was observed for any of the acoustic variables (all r's < ±.43). For the production tasks, although no correlation was observed for the intertrial duration, vowel duration, intensity, f_0, F_2 and F_3 changes (all r's < ±.30), N1 latency change was found to correlate with F_1 changes ($r = .50$, $F(1,16) = 5.4$, $p = .03$) and F_1-F_2-F_3 vowel space area changes ($r = -.53$, $F(1,16) = 6.2$, $p = .02$). However, using a Bonferroni correction and an adjusted alpha level, these two correlations were no longer significant.

In sum, no significant correlation between N1 latency and acoustic changes in response to the visual modality was found.

Second epoching procedure [-1000ms 300ms] (See Figure 4 and Table 3)

Compared to the first analysis, due to longer epochs, the EEG signal included more artefact, with a higher number of rejected trials in the production compared to the perception tasks (on average, 50 vs. 28, respectively; $F(1,17) = 13.0$, $p = .002$, $p\eta^2 = .44$). No difference was found between the sensory modalities ($F(1,17) = .4$) and there was interaction ($F(1, 17) = .2$).

Regarding MRCPs, a significant effect of the period ($F(8,136) = 11.8$, $p = .0001$, $p\eta^2 = .41$) and of the task ($F(1,17) = 5.1$, $p = .04$, $p\eta^2 = .23$) were found. Importantly, a significant period x task interaction was observed ($F(8,136) = 9.1$, $p = .0008$, $p\eta^2 = .35$). Post hoc analyses showed a higher negative amplitude in the motor compared to the perception tasks in the [-700 -600], [-600 -500], [-500 -400] and [-400 -300] periods but no significant difference between tasks in all other periods. No other effect or interactions reached significance (all F's < 1.6).

Regarding AEP amplitudes, N1 response was reduced in the production compared to the perception tasks (on average, -1.62 μV vs. -4.74 μV, respectively; $F(1,17) = 22.6$, $p = .0002$, $p\eta^2 = .57$). No difference was found between the sensory modalities ($F(1,17) = .5$) and there was no interaction ($F(1, 17) = 1.1$). For P2, the mean amplitude was 3.84 μV, without difference between the perception and the production tasks ($F(1,17) = .5$), the sensory modalities ($F(1, 17) = .1$), and no interaction ($F(1,17) = .3$).

As for AEP latencies, a shorter latency was observed on N1 for the audiovisual compared to the auditory modalities (on average, 78 ms vs. 84 ms, respectively; $F(1,17) = 6.4$, $p = .02$, $p\eta^2 = .27$). No difference was found between the production and perception tasks ($F(1,17) = .2$) and there was no interaction ($F(1, 17) = 1.5$). As in the previous analysis, the shorter latency on N1 was stronger for the M-AV vs. M-A production tasks than for the AV vs. A perception tasks (on average, -10 ms vs. -4 ms, respectively). For P2, a shorter latency was also observed for the audiovisual compared to the auditory modalities (on average, 158 ms vs. 167 ms, respectively; $F(1,17) = 10.7$, $p = .005$, $p\eta^2 = .39$). No difference was found between the perception and the production tasks ($F(1,17) = 1.0$), nor interaction ($F(1,17) = .1$).

In sum, MRCPs/RPs were evident in the production tasks, but not in the perception tasks, and were characterized by a slow negative deflection on fronto-central sites from 700 ms to 300 ms prior to the vocalic onset. No reliable difference was observed on MCRPs/RPs between the two production tasks with or without visual feedback. As in the previous analysis, a SIS effect was observed on N1 amplitude, with a reduced response in the production compared to the perception tasks, irrespective the sensory modality. In addition, the visual modality was found to speed up N1 and P2 latencies, with a shorter latency for the visual compared to auditory modalities, whatever the task modality. Once again, these two effects were not found to interact.

	M-A	**M-AV**	**A**	**AV**
Number of trials	117 (8)	121 (7)	117 (8)	121 (7)
Number of rejected trials	52 (7)	48 (6)	29 (4)	27 (6)
Amplitudes (μV)				
-900 -800 ms	-0.88 (0.32)	0.04 (0.40)	-0.73 (0.25)	-0.15 (0.31)
-800 -700 ms	-1.61 (0.62)	-0.63 (0.56)	-0.64 (0.33)	-0.43 (0.32)
-700 -600 ms	-1.79 (0.65)	-1.62 (0.65)	-0.05 (0.28)	-0.14 (0.31)
-600 -500 ms	-1.63 (0.66)	-1.93 (0.57)	0.17 (0.35)	0.15 (0.28)
-500 -400 ms	-1.10 (0.72)	-1.67 (0.52)	0.47 (0.34)	0.27 (0.32)
-400 -300 ms	-1.04 (0.63)	-1.06 (0.52)	0.32 (0.33)	0.81 (0.31)
-300 -200 ms	-0.08 (0.57)	-0.49 (0.44)	0.51 (0.32)	1.00 (0.38)
-200 -100 ms	1.29 (0.63)	1.03 (0.51)	0.24 (0.29)	0.72 (0.37)
-100 0 ms	1.91 (0.74)	1.77 (0.71)	0.38 (0.35)	0.76 (0.33)
N1	-1.60 (0.72)	-1.64 (0.65)	-5.10 (0.57)	-4.38 (0.56)
P2	4.09 (0.74)	3.74 (0.64)	3.75 (0.68)	3.78 (0.46)
Latencies (ms)				
N1	85 (5)	75 (3)	83 (3)	80 (4)
P2	165 (5)	156 (5)	169 (4)	160 (5)

Table 3: MRCP amplitudes and N1/P2 amplitudes and latencies on fronto-central electrodes in the two production and perception tasks from the [-1000 ms to 300 ms] epoching procedure (M-A, M-AV, A, AV; SEM are indicated).

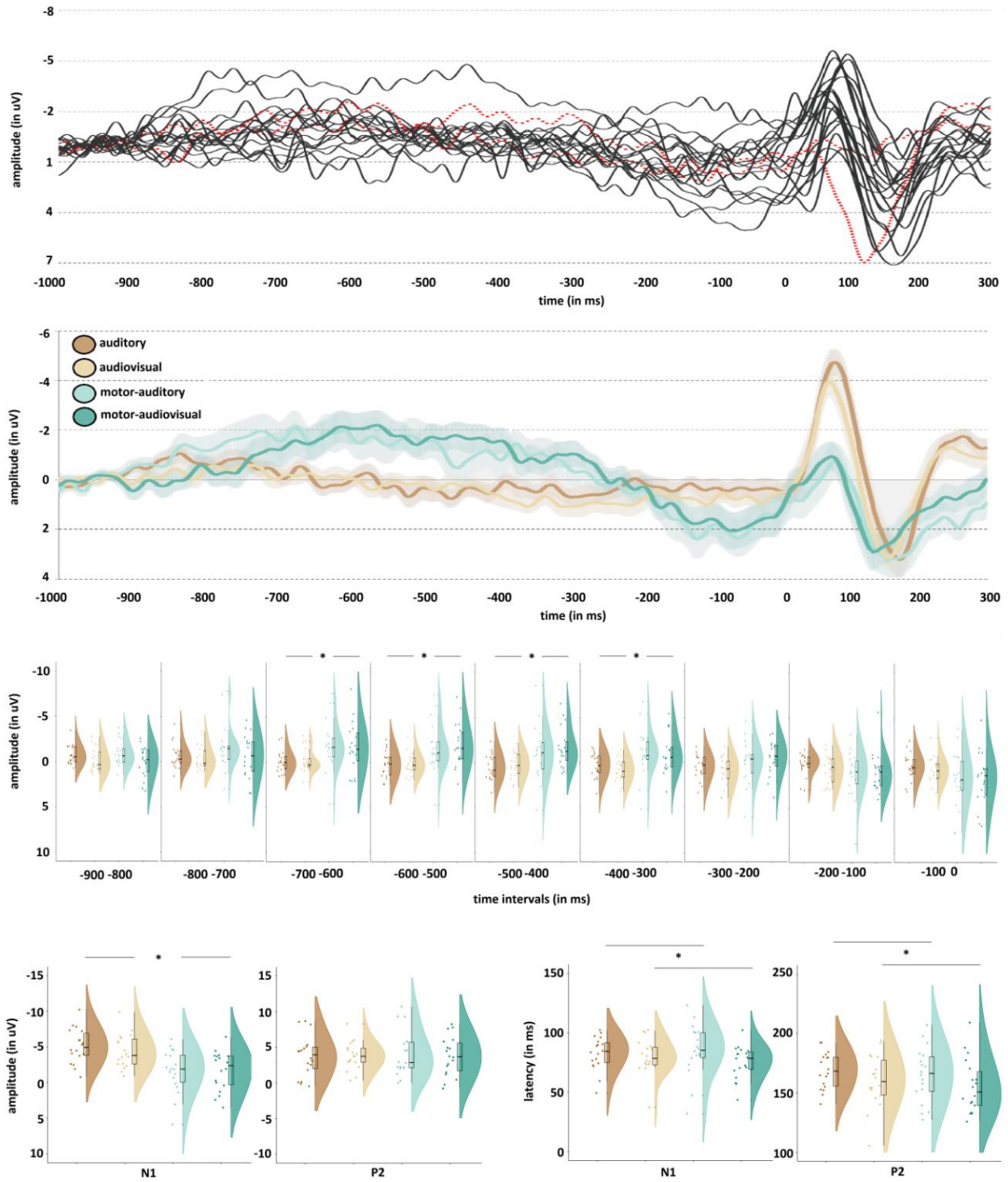

Figure 4. Top: Individual participant EEG waveforms on fronto-central electrodes averaged over the four tasks from the [-1000 ms to 300 ms] epoching procedure (M-A, M-AV, A, AV; the red lines represented the two removed participants due to abnormal signal). Middle-Top: Average EEG waveform for each task on fronto-central electrodes from the [-1000 ms to 300 ms] epoching procedure (SEM are indicated in shadow). Middle-Bottom: Mean amplitude of MRCPs. Bottom: Mean N1 and P2 AEP amplitudes and latencies. In the violin/whisker plots, the line across the box represents the median and the vertical bars represent the interquartile range. In each plot, individual data are shown as dots (n = 18). Significant contrasts are indicated.

DISCUSSION

Both endogenous motor and exogenous visual processes were found to act on the auditory neural processing of acoustic speech signal. As was expected, speaking-induced suppression was observed on N1 amplitude during speaking compared to listening, irrespective of the auditory or audiovisual feedback. Adding orofacial visual movements to the acoustic speech signal speeded up N1 latency, irrespective of the perception or production task. Importantly, these cross-modal effects were found to act differentially on N1 amplitude and latency. These results suggest distinct motor and visual influences on auditory neural processing, possibly through different neural gating and predictive mechanisms.

No acoustic difference when speaking with or without visual feedback

Before discussing these results, it should be noted that detailed analyses of the large corpus of recorded vowels did not show any significant acoustic/spectral difference during speaking, with or without visual feedback. In addition, analyses based on intra-individual acoustic variability did not show any difference between the two motor tasks either. The finding of similar acoustical realizations for vowels produced with or without visual feedback strengths the view that any difference in auditory neural processing during the motor tasks comes from motor-to-auditory and/or and visual-to-auditory cross-modal effects.

Well-acknowledged acoustic/spectral differences between vowels were however observed in both motor tasks. The distribution of F_1, F_2 and F_3 formant values indeed appeared exquisitely in line with those previously reported for French vowels (Calliope, 1989). As expected, vowel height was inversely correlated with F_1, while the relationship between F_2 and F_3 and vowel roundedness/backness appeared more complex (Schwartz et al., 1997a, 1997b; Ladefoged, 2006). The lower f_0 found for the open /a/ vowel compared to mid-close /ø/ and /e/ vowels has also been repeatedly observed in past phonetic studies and can be explained by additional air-pressure and biomechanical constraints acting on the rate of vocal fold vibration for open compared to close vowels (Ladefoged, 1964; Ohala, 1973). However, the lower intensity found for /a/ compared to /ø/ and /e/ appears inconsistent with previous studies showing that greater oral apertures lead to increased loudness (Fant, 1971; Lindblom and Sundberg, 1971). One possible explanation of this contradictatory finding may come from the neutral open mouth position between vowels, closer to /a/ than to /ø/ and /e/ articulatory configurations, that possibly may change respiratory drive among vowels. Finally, apart from a higher F_1 variability observed for /a/ compared to /ø/ and /e/, analyses on individual variability for all other acoustic parameters showed that participants' productions were constant across tasks and vowels.

Apart from acoustic/spectral parameters, a significant task x vowel interaction was observed on vowel duration, with /a/ (+10 ms) and /ø/ (+6 ms), but not /e/, slightly longer without than with visual feedback. These subtle duration differences for /a/ and /ø/ vowels may derive for more precise visual monitoring due to their visually salient openness and roundedness configurations (compared to /e/ and the neutral open

mouth position). Importantly, it is quite unlikely, if not impossible, that these subtle duration differences could have induced changes in N1 AEPs. Indeed, N1 peaks occurred well before the very end of vowels in which duration changes occurred (i.e., 75-85 ms vs. 206-227 ms).

Electrophysiological evidence for motor-to-auditory and visual-to-auditory cross-modal effects

Motor-to-auditory influences on auditory neural processing were characterized in the production tasks, but not the perception tasks, by MRCPs/RPs and by SIS occurring 700 ms to 300 ms prior to and 100 ms after speech onset, respectively. Importantly, no difference was observed on MCRPs/RPs and SIS between the two production tasks, with or without visual feedback.

Regarding MRCPs/RPs, a slow negative deflection was observed on fronto-central sites from 700 ms to 300 ms before the acoustic onset. This is in line with the litterature on MRCPs/RPs and their reported temporal profiles (Kornhuber and Deecke, 1965; Libet et al., 1983; Birbaumer et al., 1990; Wang et al., 2014). In addition to MRCPs/RPs, a classical SIS effect was observed on N1 amplitude, with a reduced response in the production compared to the perception tasks, irrespective of the sensory modality. This result is in line with previous studies EEG/MEG studies on efference copy and corollary discharge during speech production (Numminen and Curio, 1999; Numminen et al., 1999; Curio et al., 2000; Ford et al. 2001; Houde et al., 2002; Ford and Mathalon 2004; Heinks-Maldonado et al., 2005; Ventura et al., 2009; Behroozmand and Larson, 2011; Niziolek et al., 2013; Sitek et al., 2013; Wang et al., 2014; Franken et al., 2015; Sato and Shiller, 2018). Intriguingly, the SIS effect did not speed up N1 AEP, with no latency difference between the production and perception tasks. It is noteworthy that no consensus however emerges from the literature as to whether self-generated sounds can speed up N1 AEP: many studies did not report N1 latency (Ford et al., 2001; Ford and Mathalon, 2004; Heinks-Maldonado et al., 2005; Behroozmand and Larson, 2011; Sitek et al., 2013; Wang et al., 2014; Franken et al., 2015; Sato and Shiller, 2018), some studies showed shorter N1 latency (Numminen and Curio, 1999; Curio et al., 2000) or, rather, longer N1 latency (Houde et al., 2002) during speaking, and some studies reported no N1 latency difference between speaking and listening (Ventura et al., 2009; Niziolek et al., 2013).

Beside motor influences on auditory neural processing, visual-to-auditory cross-modal effects were characterized by a shorter N1 latency in the audiovisual compared to auditory modalities, irrespective of the perception or production task. This result is in line with previous EEG/MEG studies showing that adding lip movements to auditory speech speeds up N1/M100 during audiovisual compared to unimodal speech perception (Klucharev et al., 2003; Besle et al., 2004; van Wassenhove et al., 2005; Stekelenburg and Vroomen, 2007; Arnal et al., 2009; Huhn et al., 2009; Pilling, 2009; Vroomen and Stekelenburg, 2010; Winneke and Phillips, 2011; Frtusova et al., 2013; Schepers et al., 2013; Stekelenburg et al., 2013; Baart et al., 2014; Ganesh et al., 2014; Kaganovich and Schumaker, 2014; Treille et al., 2014a, 2014b; Baart and Samuel, 2015; Hisanaga et al., 2016; Paris et al., 2016; Treille et al., 2017, 2018; Pinto et al., 2019). However, compared to previous audiovisual speech perception studies, the observed N1 amplitude

reduction in the audiovisual compared to auditory modalities was not significant. This null finding is most probably due to the specific experimental procedure used here that might act on N1 amplitude: for example, the task order (each perception task being always performed following its related production task, using the same recorded speech sequence), the speech stimuli (with previous audiovisual studies almost always using consonant-vowel monosyllables of different visemic saliency and time-course; see Fisher, 1968; Summerfield, 1987), the absence of explicit and overt categorization/decision processes (with previous audiovisual studies often using manual responses), the perception of self-generated speech and the absence of an additive-model to truly test audiovisual integration (i.e., AV ≠ A +V; although audiovisual integration appears not to be significantly modulated by whether or not the visual EEG signal is subtracted from the audiovisual EEG ones; see Baart, 2016).

Two complementary cross-modal influences?

During speaking, both endogenous motor-to-auditory and exogenous visual-to-auditory processes fine-tuned the neural auditory processing of one's speech feedback. However, these two cross-modal effects were found to differentially act on N1 amplitude and latency (classically described as reflecting the size of neural population and activation synchrony and the time to process auditory events during the component generation, respectively; Näätänen and Picton, 1987; Woods, 1995).

Regarding N1 amplitude, a classical and strong SIS effect was observed in the speaking compared to the listening tasks while the visually induced amplitude reduction in the audiovisual compared to auditory-only modalities was not reliable. In line with past studies on efference copy and corollary discharge, the large amplitude reduction during speaking can be explained simply by neural auditory cancellation to self-generated speech feedback. More speculatively, it may also partly reflect speech-specific predictive mechanisms and an accurate expectation of the forthcoming acoustic speech feedback. From this hypothesis, previous studies have demonstrated that while SIS is reduced or even abolished when the expected and actual auditory feedback do not match (in cases of online auditory feedback perturbation; i.e., pitch-shifted voice, noise masking or "alien" voice; Behroozmand and Larson, 2011; Chang et al., 2013; Heinks-Maldonado et al., 2006; Houde et al., 2002), SIS is maximal for higher compared to lower prototypical speech occurrences (Niziolek et al., 2013; Sitek et al., 2013), and for simpler compared to more complex speech production tasks, the latter implying a higher acoustic variability (Ventura et al., 2009). The optimal degree of predictability is here indirectly supported by the simplicity of the production tasks (participants had to produce vowels in a natural manner and at a comfortable rhythm) and stimuli (with overlearned vowels as elementary speech units, poorly contaminated by complex coarticulation effects), and the resulting very low intra-variability of the acoustic, spectral and duration values of the produced vowels (mean SEM between 0% and 2%; see Table 1).

As for N1 latency, while a classical shorter N1 latency was observed in the audiovisual compared to auditory-only modalities, the SIS effect did not speed up N1 auditory processing during the speaking

compared to listening tasks. In past EEG/MEG studies on audiovisual speech perception, visual-to-auditory crossmodal effects resulted in articulatory-specific temporal facilitation, systematically depending on the degree to which the visual signal predicted auditory targets (van Wassenhove et al., 2005; Arnal et al., 2009). In line with these studies, the observed latency facilitation during audiovisual speech perception may likely reflect speech-specific predictive mechanisms (van Wassenhove et al., 2005; Arnal et al., 2009; van Wassenhove, 2013). Although available empirical evidence as to whether N1 is speeded up during self-generated sounds appears largely equivocal in past studies, the reason why motor processes do no speed up auditory neural processing while visual processes do remains intriguing. One first possibility is that SIS only reflects non-specific neural auditory cancellation/gating to self-generated speech feedback, thus explaining the amplitude reduction but the absence of latency facilitation during speaking. However, as discussed above, SIS may also partly reflect speech-specific predictive mechanisms together with non-specific sensory cancellation/gating mechanisms (Press et al., 2019). From this latter hypothesis, the absence of neural auditory facilitation during speaking could stem from the measurement itself, with N1 peak latency in the motor tasks reflecting different underlying neural generators as well as a less prominent peak due to strongly suppressed auditory evoked responses, possibly blurring a predictive facilitatory effect.

Hence, according to the literature, while the observed N1 latency facilitation during audiovisual compared to auditory speech perception is likely to reflect speech-specific predictive and integrative mechanisms, the large N1 amplitude reduction during speaking may primarily derive from non-specific neural auditory cancellation to self-generated speech feedback. Importantly, the shorter latency and reduced amplitude of auditory evoked responses during speaking with audiovisual feedback also suggest that motor and visual influences on auditory neural processing might operate through different neural predictive and non-specific cancellation/gating mechanisms.

At first sight, the present results have very little to tell about speech motor control. In daily-life, the use of visual feedback during speech production appears anecdotal. However, the common goal of motor-to-auditory and visual-to-auditory cross-modal effects can be seen as the fine-tuning of sensory processing to enhance perception. From that view, the observed results in the present study argue for distinct motor and visual influences on auditory neural processing, possibly through different neural predictive and non-specific cancellation/gating mechanisms. From a broader perspective, the hypothesis that motor and visual processes differentially act on neural auditory processing open new perspectives as to whether the motor and sensory systems interact synergistically in action goal (de)coding and speech motor control (Guenther, 1995, 2015; Houde and Nagarajan, 2011; Parell et al., 2019).

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, Stekelenburg JJ, Vroomen J (2014) Electrophysiological evidence for speech-specific audiovisual integration. Neuropsychologia, 65:115–211.

Baart M, Samuel AG (2015) Turning a blind eye to the lexicon: ERPs show no cross-talk between lip-read and lexical context during speech sound processing. Journal of Memory and Language, 85:42–59.

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.

Behroozmand R, Larson CR (2011) Error-dependent modulation of speech-induced auditory suppression for pitch-shifted voice feedback. BMC Neurosci., 12:1–10.

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.

Birbaumer N, Elbert T, Canavan AG, Rockstroh B (1990) Slow potentials of the cerebral cortex and behavior. Physiol. Rev., 70:1–41.

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

Calliope (1989) Calliope, La parole et son traitement automatique. Masson, Paris.

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

Chang EF, Niziolek CA, Knight RT, Nagarajan SS, Houde JF (2013) Human cortical sensorimotor network underlying feedback control of vocal pitch. Proceedings of the National Academy of Sciences of the United States of America, 110(7):2653–2658.

Chen C-MA, Mathalon DH, Roach, BJ, Calvus I, Spencer DD, Ford JM (2011) The corollary discharge in humans is related to synchronous neural oscillations. J Cogn Neurosci, 23(10):2892–2904.

Crapse TB, Sommer MA (2008) Corollary discharge across the animal kingdom. Nat. Rev. Neurosci., 9:587–600.

Creutzfeldt O, Ojemann G, Lettich E (1989) Neuronal activity in the human lateral temporal lobe. II. Responses to the subjects own voice. Exp Brain Res, 77:476–489.

Curio G, Neuloh G, Numminen J, Jousmaki V, Hari R (2000) Speaking modifies voice-evoked activity in the human auditory cortex. Hum Brain Mapp, 9:183–191.

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

Duckworth M, McDougall K, de Jong G, Schockey L (2011) Improving the consistency of formant measurement. International Journal of Speech, Language and the Law, 18:35–51.

Fant G (1960) Acoustic Theory of Speech Production. The Hague: Mouton.

Flinker A, Chang EF, Kirsch HE, Barbaro NM, Crone NE, Knight RT (2010) Single-trial speech suppression of auditory cortex activity in humans. Journal of Neuroscience, 30(49):16643–16650.

Ford JM, Mathalon DH, Heinks T, Kalba S, Faustman WO, Roth WT (2001) Neurophysiological evidence of corollary discharge dysfunction in schizophrenia. Am J Psychiatry, 158:2069–2071.

Ford JM, Mathalon DH (2004) Electrophysiological evidence of corollary discharge dysfunction in schizophrenia during talking and thinking. J Psychiatr Res, 38:37–46.

Ford JM, Roach BJ, Mathalon DH (2010) Assessing corollary discharge in humans using noninvasive neurophysiological methods. Nature Protocols, 5:1160–1168.

Fant G (1971) Acoustic theory of speech production. The Hague, the Netherlands: Mouton

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

Frank M, Muhlack B, Zebe F, Scharinger M (2020) Contributions of pitch and spectral information to cortical vowel categorization. Journal of Phonetics, 79:100963.

Franken MK, Hagoort P, Acheson DJ (2015) Modulations of the auditory M100 in an imitation task. Brain & Language, 142:18–23.

Frtusova JB, Winneke AH, Phillips NA (2013) ERP evidence that auditory–visual speech facilitates working memory in younger and older adults. Psychology and Aging, 28(2):481–494.

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.

Guenther FH (2015) Neural control of speech. Cambridge, MA: The MIT Press.

Guenther FH (1995) Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. Psychological Review, 102:594–621.

Heinks-Maldonado TH, Nagarajan SS, Houde JF (2006) Magnetoencephalographic evidence for a precise forward model in speech production. Neuroreport, 17:1375–1379.

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.

Hisanaga S, Sekiyama K, Igasaki T, Murayama N (2016) Language/culture modulates brain and gaze processes in audiovisual speech perception. Scientific Reports, 6:35265.

Houde JF, Nagarajan SS, Sekihara K, Merzenich MM (2002) Modulation of the auditory cortex during speech: an MEG study. J Cogn Neurosci, 14:1125–1138.

Houde JF, Nagarajan SS (2011). Speech production as state feedback control. Front Hum Neurosci., 5:82.

Huhn Z, Szirtes G, Lörincz A, Csépe V (2009) Perception based method for the investigation of audiovisual integration of speech. Neuroscience Letters, 465(3):204–209.

Kaganovich N, Schumaker J (2014) Audiovisual integration for speech during mid-childhood: Electrophysiological evidence. Brain and Language, 139:36–48.

Kent RD, Vorperian HK (2018) Static measurements of vowel formant frequencies and bandwidths: a review. Journal of Communication Disorders, 74:74-97.

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.

Kornhuber HH, Deecke L (1965) Hirnpotentialänderungen bei Willkürbewegungen und passiven Bewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale. Pflügers Arch., 284:1–17.

Ladefoged P (1964) A phonetic study of West African languages: an auditory-instrumental survey. Cambridge: Cambridge University Press.

Ladefoged P (2006). A course in phonetics (5th ed.). Boston, MA: Thomson Wadsworth.

Lennes M (2017) SpeCT - The Speech Corpus Toolkit for Praat (v1.0.0). First release on GitHub. Zenodo. http://doi.org/10.5281/zenodo.375923.

Libet B, Gleason CA, Wright EW, Pearl DK (1983) Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). Brain, 106(3):623–642.

Lindblom B (1967) Vowel duration and a model of lip mandible coordination. Quart. Progr. Status Rep., 4:1-29.

Lindblom B, Sundberg, JEF (1971) Acoustical consequences of lip, tongue, jaw, and larynx movement. The Journal of the Acoustical Society of America, 50:1166–1179.

Nasreddine ZS, Chertkow H, Phillips N, Bergman H, Whitehead V (2003) Sensitivity and specificity of the Montreal Cognitive Assessment (MoCA) for detection of mild cognitive deficits. Can J Neurol Sci, 30(30).

Nasreddine ZS, Phillips NA, Bedirian V, Charbonneau S, Whitehead V, Collin I, Chertkow, H (2005) The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc, 53(4):695-699.

Niziolek CA, Nagarajan SS, Houde JF (2013) What does motor efference copy represent? Evidence from speech production. The Journal of Neuroscience, 33(41):16110–16116.

Numminen J, Curio G (1999) Differential effects of overt, covert and replayed speech on vowel-evoked responses of the human auditory cortex. Neurosci Lett, 272:29–32.

Numminen J, Salmelin R, Hari R (2000) Subject's own speech reduces reactivity of the human auditory cortex. Neurosci Lett, 265:119–122.

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.

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

Ohala JJ (1973) Explanations for the intrinsic pitch of vowels. Monthly Internal Memorandum, Phonology Laboratory University of California, Berkeley, 9-26.

Paris T, Kim J, Davis C (2016) Using EEG and stimulus context to probe the modelling of auditory-visual speech. Cortex, 75:220–230.

Parell B, Ramanarayanan V, Nagarajan S, Houde JF (2019) The FACTS model of speech motor control: fusing state estimation and task-based control. Plos Comput. Biol., 15(9): e1007321.

Pereira J, Ofner P, Schwarz A, Sburlea AI, Müller-Putz GR (2017) EEG neural correlates of goal-directed movement intention. NeuroImage, 149:129-140.

Perkell JS (1969) Physiology of speech production: results and implications of a quantitative cineradiographic study. Research Monograph No. 53, MIT, Cambridge.

Pilling M (2009) Auditory event-related potentials (ERPs) in audiovisual speech perception. J Speech Lang Hear Res, 52(4):1073–1081.

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.

Press C, Kok P, Yon D (2019) The perceptual prediction paradox. Trends in Cognitive Sciences, 24(1):13-24.

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.

Sams M, Mottonen R, Sihvonen T (2005) Seeing and hearing others and oneself talk. Cognitive Brain Research, 23:429-435.

Sato M, Shiller DM (2018) Auditory prediction during speaking and listening. Brain and Language, 187:92-103.

Schepers IM, Schneider TR, Hipp JF, Engel AK, Senkowski D (2013) Noise alters beta-band activity in superior temporal cortex during audiovisual speech processing. NeuroImage, 70:101–112.

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

Schwartz JL, Boë, LJ, Vallée N, Abry C (1997a) Major trends in vowel system inventories. Journal of Phonetics, 25:233–254.

Schwartz, JL, Boë LJ, Vallée N, Abry C (1997b) The dispersion-focalization theory of vowel systems. Journal of Phonetics, 25:255–286.

Schwartz JL, Ménard L, Basirat A, Sato M (2012) The Perception for Action Control Theory (PACT): a perceptuo-motor theory of speech perception. Journal of Neurolinguistics, 25(5):336-354.

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.

Sitek K, Mathalon DH, Roach BJ, Houde JF, Niziolek CA, Ford JM (2013) Auditory cortex processes variation in our own speech. PLoS ONE, 8(12), e82925.

Sperry R (1950) Neural basis of the spontaneous optokinetic response produced by visual inversion. J. Comp. Physiol. Psychol., 43:482–489.

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

Stekelenburg JJ, Maes JP, van Gool AR, Sitskoorn M, Vroomen J (2013) Deficient multisensory integration in schizophrenia: An event-related potential study. Schizophrenia Research, 147:253–261.

Straka H, Simmers J (2018) A new perspective on predictive motor signaling. Current Biology, 28:R232-R243.

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.

Thomas IB (1969) Perceived pitch of whispered vowels. Journal of the Acoustical Society of America, 46:468–470.

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 S, Shiller DM, Ostry DJ (2003) Somatosensory basis of speech production. Nature, 423: 866–869.

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.

Wang J, Mathalon DH, Roach BJ, Reilly J, Keedy SK, Sweeney JA, Ford JM (2014) Action planning and predictive coding when speaking. NeuroImage, 9:91–98.

Ventura MI, Nagarajan SS, Houde JF (2009) Speech target modulates speaking induced suppression in auditory cortex. BMC Neurosci, 10:58.

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.

von Holst E, Mittelstaedt H (1950) The reafference principle. Naturwissenschaften 37, 464–467.

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

Woods D (1995) The component structure of the N1 wave of the human auditory evoked potential. Electroencephalography and Clinical Neurophysiology, 44:102–109.