frontiers in Human Neuroscience

BRIEF RESEARCH REPORT
published: 27 October 2020
doi: 10.3389/fnhum.2020.538619

Electrophysiological Dynamics of Visual Speech Processing and the Role of Orofacial Effectors for Cross-Modal Predictions

Maëva Michon[1,2]*, *Gonzalo Boncompte*[3] *and Vladimir López*[4]

[1] *Laboratorio de Neurociencia Cognitiva y Evolutiva, Escuela de Medicina, Pontificia Universidad Católica de Chile, Santiago, Chile,* [2] *Laboratorio de Neurociencia Cognitiva y Social, Facultad de Psicología, Universidad Diego Portales, Santiago, Chile,* [3] *Laboratorio de Neurodinámicas de la Cognición, Escuela de Medicina, Pontificia Universidad Católica de Chile, Santiago, Chile,* [4] *Laboratorio de Psicología Experimental, Escuela de Psicología, Pontificia Universidad Católica de Chile, Santiago, Chile*

The human brain generates predictions about future events. During face-to-face conversations, visemic information is used to predict upcoming auditory input. Recent studies suggest that the speech motor system plays a role in these cross-modal predictions, however, usually only audio-visual paradigms are employed. Here we tested whether speech sounds can be predicted on the basis of visemic information only, and to what extent interfering with orofacial articulatory effectors can affect these predictions. We registered EEG and employed N400 as an index of such predictions. Our results show that N400's amplitude was strongly modulated by visemic salience, coherent with cross-modal speech predictions. Additionally, N400 ceased to be evoked when syllables' visemes were presented backwards, suggesting that predictions occur only when the observed viseme matched an existing articuleme in the observer's speech motor system (i.e., the articulatory neural sequence required to produce a particular phoneme/viseme). Importantly, we found that interfering with the motor articulatory system strongly disrupted cross-modal predictions. We also observed a late P1000 that was evoked only for syllable-related visual stimuli, but whose amplitude was not modulated by interfering with the motor system. The present study provides further evidence of the importance of the speech production system for speech sounds predictions based on visemic information at the pre-lexical level. The implications of these results are discussed in the context of a hypothesized trimodal repertoire for speech, in which speech perception is conceived as a highly interactive process that involves not only your ears but also your eyes, lips and tongue.

Keywords: orofacial movements, place of articulation, ERPs, viseme, articuleme, speech motor system, cross-modal prediction

OPEN ACCESS

Edited by:
Xiaolin Zhou,
Peking University, China

Reviewed by:
Blake Warren Johnson,
Macquarie University, Australia
Kirrie J. Ballard,
The University of Sydney, Australia

***Correspondence:**
Maëva Michon
mmichon@uc.cl

Specialty section:
This article was submitted to Speech and Language, a section of the journal Frontiers in Human Neuroscience

Received: *06 March 2020*
Accepted: *29 September 2020*
Published: *27 October 2020*

Citation:
Michon M, Boncompte G and López V (2020) Electrophysiological Dynamics of Visual Speech Processing and the Role of Orofacial Effectors for Cross-Modal Predictions. Front. Hum. Neurosci. 14:538619. doi: 10.3389/fnhum.2020.538619

INTRODUCTION

Action-perception coupling has been the focus of extensive research in the field of cognitive neuroscience over the last decades. This research framework has led to a conception of the architecture of mind, that emphasizes the fact that behavior and neural dynamics are embedded in a body and situated in a context. It also reconsiders the importance of the agency and historicity

of living organisms in shaping behavior and cognition (Maturana and Varela, 1987; Thompson and Cosmelli, 2011; Gomez-Marin and Ghazanfar, 2019). In line with those ontological principles, considerable efforts have been made to rethink traditional, modular accounts of speech and language (Tremblay and Dick, 2016; Duffau, 2018). Increasingly robust findings from the field of psycholinguistics (Glenberg and Gallese, 2012; Gambi and Pickering, 2013; Pickering and Garrod, 2013), computational neuroscience (Pulvermüller and Fadiga, 2010; Pulvermüller et al., 2016) and cognitive neuroscience (D'Ausilio et al., 2009; Peelle, 2019) suggest that ecological human communication is achieved by means of highly interactive multi-modal processes and feedforward predictions.

Visemes-Phoneme Binding

While the association between speech sounds and articulatory representations of speech is well documented (Hickok and Poeppel, 2004, 2007; Okada et al., 2018), the interactions between visual and auditory forms of speech have only recently received considerable attention. Speech is not a purely auditory signal. A compelling illustration of the multisensorial integration of speech is the McGurk effect (McGurk and MacDonald, 1976). During ecological face-to-face interactions, perception of the speaker's orofacial articulatory movements offers critical complementary information for speech perception during infancy (Weikum et al., 2007; Lewkowicz and Hansen-Tift, 2012; Sebastián-Gallés et al., 2012; Tenenbaum et al., 2012), speech-in-noise perception (Sumby and Pollack, 1954; Ross et al., 2006), for non-native speech processing (Navarra and Soto-Faraco, 2005; Hirata and Kelly, 2010) and for people with hearing difficulties (Bernstein et al., 2000; Auer and Bernstein, 2007; Letourneau and Mitchell, 2013; Dole et al., 2017; Worster et al., 2017).

Imagine yourself, in a crowded party where the acoustic channel is overloaded by surrounding conversations, music and laughter. The perception of the articulatory movements of your friend's mouth would help you to cope with the challenging acoustic context, "perhaps by directing attentional resources to appropriate points in time when to-be-attended acoustic input is expected to arrive" (Golumbic et al., 2013, p. 1417). Visual information precedes auditory signals by 100-200 ms (Chandrasekaran et al., 2009). Thus, visual speech cues have the potential to serve a predictive function about the expected timing of upcoming auditory input (Arnal et al., 2009). For instance, if you see your friend opening her mouth, you would generate a prediction about her intention to initiate a conversation. This phenomenon, called predictive timing (Arnal and Giraud, 2012; van Wassenhove, 2013; Ten Oever et al., 2014), is especially relevant for turn-taking dynamics in human communication (Garrod and Pickering, 2015). In addition to providing temporal information about speech onset, visemes are particularly informative because the shape of the lips and/or the position of the tongue restrains the possible subsequent auditory input to a subset of possible phonemes. Seeing your friend pressuring her inferior and superior lips against each other would lead you to expect the upcoming sound to begin with a bilabial speech sound like /p/, /b/ or /m/. This phenomenon is known as predictive coding (van Wassenhove, 2007; Peelle and Sommers, 2015) and has been documented for both pre-lexical (Brunellière et al., 2013) and semantic (Økland et al., 2018) aspects of speech. Here, we will focus on cross-modal predictions in the context of speech perception, but it is important for the reader to be reminded that the neural feedforward processes taking place between auditory and visual modalities are not unique to speech, but rather rely on more domain-general dynamics of multisensory integration and error prediction (Kilner et al., 2007; Seth, 2013).

In electrophysiological studies, the effect of auditory facilitation, indexed by shorten latencies of the auditory evoked potential N1, is a well-documented consequence of cross-modal forward predictions [Shahin et al., 2018; also see Baart (2016) for a meta-analysis]. The N400, a component known for its responsiveness to semantic incongruence, has also been reported to be significantly enhanced in response to viseme-phoneme incongruence at the phonemic/syllabic level (Kaganovich et al., 2016; Kaganovich and Ancel, 2019). Interestingly, cross-modal facilitation and predictive coding have been shown to be modulated by visemic salience, with greater predictability for visemes with higher visual salience (van Wassenhove et al., 2005; Paris et al., 2013, 2017). Brunellière et al. (2013), for instance, reported an increase of late N400 component amplitude for visemes with highly salient visual cues (/p/) with respect to less salient visual cues (/k/).

It has been well-established by early fMRI studies that silent lip-reading produces an activation of auditory cortices (Sams et al., 1991; Calvert et al., 1997; Calvert and Campbell, 2003; Pekkola et al., 2005; Blank and von Kriegstein, 2013; Bernstein and Liebenthal, 2014). More recently, the analysis of oscillatory dynamics has consistently revealed that both auditory and visual speech perception induce neural entrainment at similar rhythms (Park et al., 2016, 2018; Assaneo and Poeppel, 2018; Poeppel and Assaneo, 2020). Importantly for the purpose of the current study, even in the absence of auditory input, the brain synthesizes the missing speech sounds based on visemic information. Silent lip-reading generates entrainment to the absent auditory speech at very slow frequencies (below 1Hz) in auditory cortices, even when participants do not know what the absent auditory signal should be (Bourguignon et al., 2018, 2020).

Articuleme: The Smallest Distinctive Unit of Speech Motor Repertoire

Both speech sounds and their visual counterparts are physical outcomes of a sequence of coordinated articulatory movements of the vocal tract and orofacial effectors. We will refer to the articulatory neural patterns of activity that give rise to particular phonemes and visemes as *articulemes*. In this line, articulemes are conceptualized as partially invariant and language-specific patterns of neural and motor activity that, when instantiated, produce contrastive and meaningful linguistic information. This concept was first introduced by the Russian neuropsychologist Luria (1965, 1973) to label the specific articulatory patterns required to produce a phoneme (Ardila et al., 2020). Although this terminology has gone mostly unused for decades, we believe the notion of articuleme could reduce ambiguity in many current debates (see Michon et al., 2019). In fact, depending on disciplinary and theoretical background, the terminology

used to refer to speech articulation varies (e.g., articulatory gesture, but also motor plan/program). We believe none of these terms is precise or clear enough to distinguish between the observed (visemes) and the produced (articuleme) language-specific orofacial gestures. Also, in light of recent evidence showing the relevance of the articulatory system in speech perception, we reintroduce the term *articuleme*. Here, we define it as the smallest unit of speech motor repertoire that can be isolated in the speech flux which produces meaningful and distinguishable elements of a given language.

Neuroimaging (Skipper et al., 2005; Pulvermüller et al., 2006; Correia et al., 2015; Archila-Meléndez et al., 2018) and TMS (Watkins et al., 2003; D'Ausilio et al., 2009; Sato et al., 2010; Swaminathan et al., 2013; Nuttall et al., 2018) studies have provided strong evidence of the participation of speech motor cortices during speech perception (but also see Stokes et al., 2019). Interestingly, using a variety of creative experimental procedures, a growing number of studies suggest that interfering with articulatory effectors negatively impacts speech perception. These sensorimotor influences on speech perception have been observed early in infancy: 6-month-old infants were unable to discriminate between non-native consonant contrasts when the relevant articulatory effector needed to produce the contrast was specifically restrained by a teething toy (Bruderer et al., 2015). Similarly, the ability of 7 years old children to recognize words by lipreading declined when they were holding a tongue depressor horizontally between their teeth (Bruderer et al., 2015). It has been suggested that language production system is required to elaborate predictions during speech perception (Pickering and Garrod, 2007). In this line, using fMRI Okada et al. (2018) showed that silent articulation of speech elicited greater activity than imagined speech in inferior frontal and premotor cortices and, although speech articulation was silent, in auditory cortex. The authors interpreted their results as an evidence of predictive coding, where activation of motor articulatory plans (here, articulemes) lead to predictions about the sensory consequences of those motor commands, which in turn serve to facilitate error monitoring and minimization. Martin et al. (2018) provided further evidence of the recruitment of language production systems during comprehension by demonstrating that the availability of the speech production system is necessary for generating lexico-semantic predictions, as indexed by greater amplitude of N400 when articulatory effectors were available vs. unavailable.

To summarize, the literature reviewed above suggests that visemes carry predictive information about forthcoming phonemes, with salient visemes producing stronger predictions of upcoming phonemes than less salient visemes. Strikingly, even in the absence of the auditory modality, the brain synthesizes the missing speech sounds on the basis of visemic content. Importantly, these cross-modal predictions between phonemic and visemic aspects of speech seem to depend on the speech motor system, and more specifically on the availability of the effectors required to generate particular speech sounds. This theory is often considered to be a modern and weaker version of Liberman's motor theory of speech perception (Liberman and Mattingly, 1985; Skipper et al., 2005; Massaro and Chen, 2008).

Although language perception and production have traditionally been studied as independent functional modules or "epicenters" (Tremblay and Dick, 2016), recent evidence points toward a highly interactive multimodal network that associates perceived orofacial movements with acoustic representation based on the motor sequences required to generate those movements. We recently proposed, based on this network, a trimodal repertoire of speech in which phonemes, visemes, and articulemes are bounded to achieve a more ecological, enactive and seamless perception of speech (Michon et al., 2019).

In contrast to the growing body of studies documenting the neuroanatomical circuits involved in audiovisual speech perception, the electrophysiological data available about the silent, visual processing of speech is still scarce. In the current study, two experiments were performed aiming to elucidate whether or not the linguistic content and the salience of visual speech cues modulates the electrophysiological responses elicited by perceiving silent orofacial movements and to what extent interfering with articulatory effectors can affect these responses.

METHODS

Stimuli

The stimuli consisted in a set of 120 silent video clips displaying either no facial movement (1- still faces), one of a variety of orofacial movements (2- forward syllables, 3- backward syllables, 4- non-linguistic movements) or the movement of a purely geometrical shape (5- geometric). Videos were rendered into 1,080 × 1,920 pixel clips, lasting ∼2 s (M = 2,052 ms and SD = 59 ms), with a frame rate of 29 frames per second. In the still faces condition (1), no mouth movements were produced (baseline). The forward syllable condition (2) contained videos of people producing consonant-vocal (CV) segments starting with phonemes that differed in their place of articulation (PoA) coarticulated with the vowel /a/. Three types of phonemes where included accordingly to their PoA: bilabial (/p/ or /b/), alveolar (/d/ or /t/) and velar (/g/ or /k/), which require lip, tongue-tip and tongue-back movements for their production, respectively. We chose these consonants because they have the common feature of being stop consonants, which means that they are articulated by closing the airway so as to impede the flow of air, then maintaining the airway closed thus generating a slight air pressure and finally generating the sound by opening the airway and releasing the airflow. Importantly, syllables with these three PoA have been reported to have different visual salience: syllables starting with bilabial movements are more salient than those starting with velar movements (van Wassenhove et al., 2007; Jesse and Massaro, 2010; Paris et al., 2013). In the backward syllables condition (3), the same videos described previously were reproduced backwards. Because of their particular motor sequence, CV formed with stop consonants cannot be pronounced backwards. In that sense, backward played syllables represent an ideal control condition because these kinds of articulatory movements are visually very similar to speech in their low-level features but at the same time they are not pronounceable, they are not present in our motor repertoire. In the non-linguistic condition

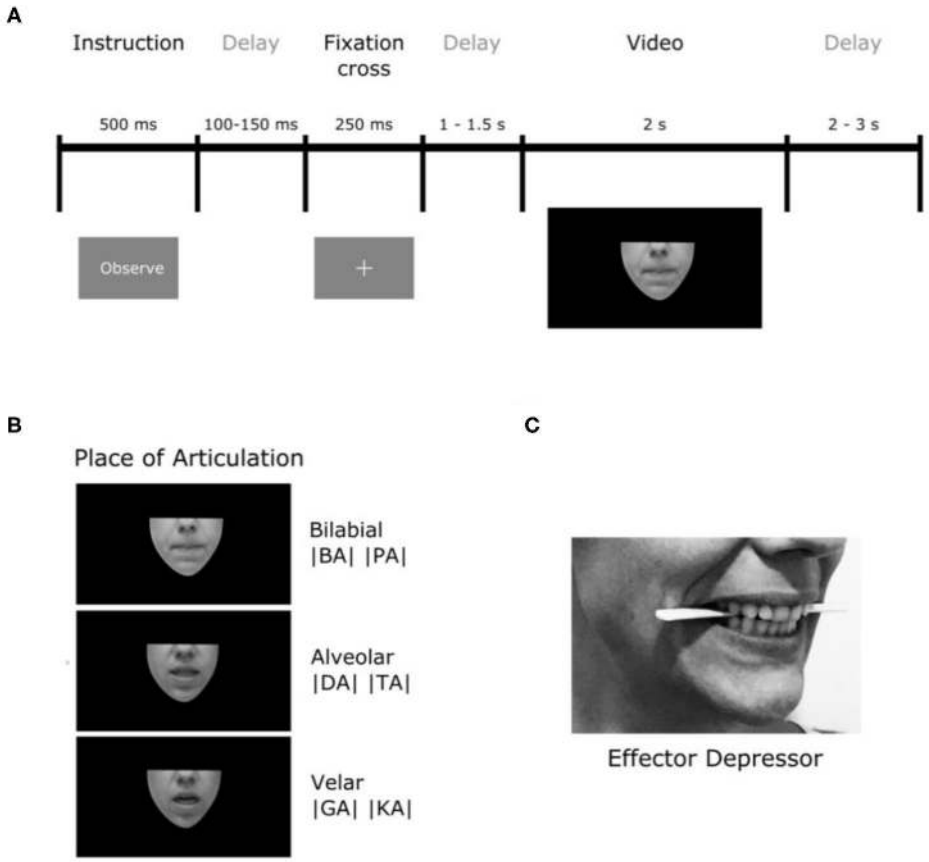

FIGURE 1 | Experimental procedure. **(A)** Timeline description of trials. **(B)** Illustration of the places of articulation (PoA) of the syllables used. **(C)** Depiction of the position of the effector depressor in participants' mouths.

(4), orofacial movements producing no audible sounds (e.g., tongue protrusion, lip-smacking) were presented. This condition was introduced to control the activity associated with the processing of orofacial movements that do not present linguistic meaning. Finally, to control for general movement perception, independently of its biological and face-related nature, a fifth condition was added where opening and closing movements of different geometrical figures (e.g., ovals, squares, triangles) were shown. These stimuli were generated and presented using PsychoPy (Peirce, 2009).

Importantly, all the videos were silently displayed (i.e., audio removed) and they only showed the lower part of the speaker's face (see **Figure 1**) in order to ensure that their eyes movements would not interfere. The software Adobe Premiere Pro CC 2017 (Adobe Systems) was used to edit the videos in a way that each began and ended with a still face (no mouth movements) or still geometrical shapes for condition 5 (sample videos for each condition are provided in the **Supplementary Material**).

Participants

Thirty-two right-handed subjects (20 females) with normal or corrected-to-normal vision and hearing and without any history of psychiatric or neurological disorders performed the experiments. Participants' ages ranged from 18 to 36 years old ($M = 22.8$, $SD = 4.2$ years). The experimental protocol was approved by the Ethics Committee of Pontificia Universidad Católica de Chile. The procedure was explained to every participant and written informed consent was obtained from each one before the experiment began. Four participants were removed from the final analysis because of incomplete EEG recording or poor signal-to-noise ratio.

Procedure

Participants sat at a distance of 70 cm from a computer screen and were asked to attentively observe or imitate the movements shown in the videos. The trial (see **Figure 1A**) started with a word lasting for 500 ms that indicate the instruction, either "Observe" (90% of the trials) or "Imitate" (10% of the trials). After 100 to 150 ms, a fixation cross appeared for 250 ms. In the observation condition, the video was displayed once, 1,000 to 1,500 ms after the white cross disappeared. In the imitation condition, the video was displayed a first time and participants were asked to attentively observe in order to co-imitate the orofacial gesture when the video was displayed for the second time. The onset of imitation was cued with a red fixation cross. After video offset, a new trial began within 2 to 3 s. The imitation condition

was included as a sham task for the participants to maintain their attention on the stimuli and the experiment in general. No imitation instruction was given for purely geometric stimuli. The data from imitation trials were not analyzed.

To study the role of speech the motor system in speech perception, the very same experiment was repeated (Experiment 2) with the difference that participants were asked to hold a wood tongue depressor horizontally between their premolars, just behind incisors (see **Figure 1C**). This strategy, which produced an unnatural skin stretching of cheeks and lips, was introduced with the objective of generating a local motor perturbation of articulatory effectors of interest (e.g., lips). It is important to notice that the object used is called a tongue depressor because it is generally used by physicians in clinical settings to lower the tongue so they can observe the patient's throat. However, its use here was different, and aimed to interfere with speech articulations in the upper vocal tract. More precisely, due to the position of the tongue depressor, the motor perturbation acted more on lips and tongue-tip movements than on tongue-back movements. For this reason, we will refer to this object as "effector depressor." In order to reduce muscle artifacts in the EEG signal, participants were asked not to squeeze their jaws, but to gently sustain the effector depressor between their premolars. For imitation trials, participants were asked to remove the depressor when the instruction "Imitate" appeared, so they could properly imitate.

Each of the 5 conditions (i.e., 1- still faces, 2- forward syllables, 3- backward syllables, 4- non-linguistic, and 5- geometric shape movements) consisted of 3 repetitions of 24 video-clips, leading to a total of 72 trials per condition (360 per experiment). The order of the conditions was pseudo-randomized across trials. For syllables conditions (2 and 3), an equal number of bilabial, alveolar, and velar syllables were used. The order of Experiment 1 and 2 was counterbalanced between participants.

Electroencephalographic Recording Parameters

Electrophysiological activity was registered with a 64-channel EEG system (Biosemi ® ActiveTwo) with electrodes positioned according to the extended 10–20 international system. The signal was acquired with a sampling rate of 2,048 Hz and an online band-pass filter (0.1 to 100 Hz). Four external electrodes were used to monitor eye movements. Two of them were placed in the outer canthi of the eyes to record horizontal EOG and the other two were positioned above and below the right eye to record vertical EOG. Two additional external electrodes were placed bilaterally on the mastoids for re-referencing. Data pre-processing was performed using MATLAB (The Mathworks, Inc.) with EEGLAB (Delorme and Makeig, 2004) and ERPLAB toolboxes (Lopez-Calderon and Luck, 2014). Afterwards, the signal was down sampled to 512 Hz, re-referenced to mastoids and band-pass filtered between 0.1 and 40 Hz for ERP analysis. The 40 Hz low-pass filter ensured that muscle and 50 Hz AC current artifacts were removed. The EEG signal was then segmented into epochs from −500 to 1,500 ms respect to stimulus onset. Each epoch was visually inspected to reject large artifacts caused by head movements, electrode drifts or any amplitude changes exceeding ±100 μV. Then, Independent Component Analysis (ICA) decomposition was performed using the "binica" function (EEGLAB), the components typically associated with eye-blinking and the remaining artifacts were rejected using the MARA ("Multiple Artifact Rejection Algorithm") plugin of EEGLAB.

Statistical Analyses

The ERP components of interest for statistical analyses were N400 and a positivity around 1,000 ms (P1000). Using the ERP measurement tool in ERPLAB, mean amplitudes were calculated with respect to a 500 ms pre-stimulus baseline for the following selected time windows: N400 [475-525 ms] and P1000 [975-1025 ms]. These time windows were chosen based on the peak amplitude of each component. After mean amplitudes of those time windows were extracted for each condition and subject, data were analyzed using repeated measures ANOVAs. For comparisons relative to the PoA effect, 3-way (effector depressor × forward/backward × electrode) repeated measures ANOVAs were performed independently for the three types of syllables (bilabial, alveolar, and velar). When main effects were significant, simple main effect analysis was performed analyzing the difference of means between the levels of a single way of the ANOVA (e.g., comparison between bilabial | Forwards syllables | With v/s Without effector depressor). The reported p-values correspond to the significance of comparisons after Bonferroni corrections. For all statistical analyses involving more than two levels, the sphericity assumption was checked using Mauchly's test. In cases of a violation of the sphericity assumption, the Greenhouse-Geisser adjusted p-values were used to determine significance. Effect sizes are reported for all significant repeated-measures ANOVAs using the partial eta squared statistic (η_p^2). All statistical analyses were performed using JASP software (Version 0.12.2; JASP Team., 2020).

RESULTS

N400

N400 has been found to peak over fronto-parietal electrodes but also to be lateralized in similar linguistic settings (Kutas and Hillyard, 1980). In this line, to assess the effect that PoA could have in the amplitude of the N400 component, we conducted a repeated measures two-way ANOVA, with PoA and electrode (F3, Fz, and F4) as ways, and the N400 amplitude as the dependent variable, for forward syllables in Experiment 1. This analysis showed a significant main effect of PoA [$F(2,54) = 4.576$, $p = 0.022$, $\eta_p^2 = 0.145$; see **Figure 2A**]. More specifically, *post-hoc t*-tests comparisons (with Bonferroni correction) indicated that N400 amplitude was significantly greater for bilabial than for velar syllables across electrodes (mean difference = −2.037 μV, $p = 0.012$) whereas no significant differences were found between other PoAs. In this analysis, no significant main effect of electrode was found [$F(2,54) = 0.657$, $p = 0.439$, $\eta^2_p = 0.024$], however, the topological distribution of our N400 component is consistent with the literature (**Figure 2B**). Interestingly, the same analysis was run for Experiment 2, revealing no main effect of PoA [$F(2,54) = 0.170$, $p = 0.844$, $\eta^2_p = 0.006$] and no significant difference between bilabial and velar syllables (all corrected $p > 0.05$). This suggests that the effector depressor

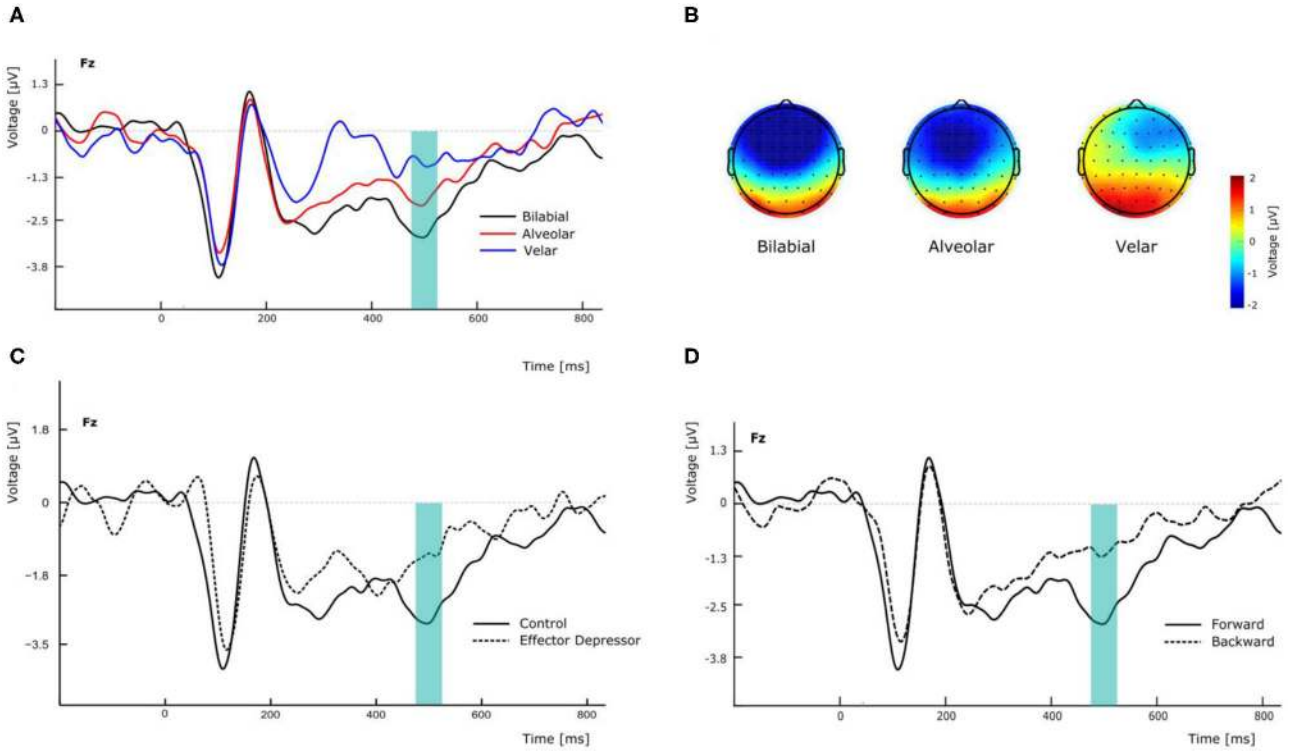

FIGURE 2 | Effects of experimental manipulation on N400 component. **(A)** Effect of the different PoA of forward syllables in electrode Fz. **(B)** Topographical maps of the N400 according the PoA at the peak amplitude latency (500 ms). **(C)** Effect of effector depressor on the perception of forward bilabial syllables in electrode Fz. **(D)** Effect of forward versus backward syllables on the perception of bilabial syllables without effector depressor in electrode Fz.

had an important impact on the elicitation of the N400 ERP component.

To better assess the differential effect of the effector depressor in bilabial CVs, we conducted a three-way repeated measure ANOVA for syllables with a bilabial PoA using Experiment, forward/backward and electrode as ways. This analysis revealed a significant interaction between Experiment (presence or absence of effector depressor) and forward/backward [$F(1,27) = 10.219$, $p = 0.004$, $\eta^2_p = 0.275$]. Simple main effect analysis showed that forward bilabial syllables elicited significantly greater N400 in the Experiment 1, in which participants freely observed the stimuli (control), compared to Experiment 2, where orofacial articulatory movements of the participants were restrained (effector depressor; **Figure 2C**). This effect was significant for all electrodes tested (F3, Fz, and F4; see **Table 1**), indicating the importance of the availability of motor effectors for the elicitation of N400. Importantly, this effect of the effector depressor was not observed for backward bilabial syllables. We then analyzed the simple main effects of bilabial syllables presented forward vs. backward. This analysis showed that forward syllables elicited greater N400 than backward syllables (**Table 2**). This effect was significant in all electrodes tested (F3, Fz, and F4 for Experiment 1, see **Figure 2D**) but were not significant for Experiment 2.

To further investigate the visemic modulation of N400, an additional two-way ANOVA was performed for electrode Fz in Experiment 1 with PoA and forward/backward as ways, eliciting a significant interaction [$F(2,54) = 7.337$, $p = 0.002$, $\eta^2_p = 0.214$]. More specifically, *post-hoc* t-test comparisons (with Bonferroni correction) indicated that N400 amplitude was significantly greater for forward bilabial CVs compared to backward bilabial CVs (mean difference $= -1.648$ μV, $p = 0.005$).

P1000

As illustrated in **Figure 3A**, the perception of visual forward and backward syllables, independently of their PoA, produced a late positivity with a peak amplitude latency around 1,000 ms, which was absent in the control conditions (still Face, non-linguistic, and geometrical shape). A three-way repeated measures ANOVA (condition, Experiment, and electrode as ways) was conducted for the amplitude of this late positivity revealing a significant main effect of condition [$F(4,108) = 9.684$, $p < 0.001$, $\eta^2_p = 0.264$]. *Post-hoc* pairwise t-tests comparisons (with Bonferroni correction) indicated that the amplitude was significantly greater for forward syllables and backward syllables respect to still faces (mean difference $= -1.085$ μV, $p = 0.01$ and mean difference $= 0.873$ μV, $p = 0.009$ for forward and backward syllables, respectively), to non-linguistic orofacial movements (mean difference $= -1.040$ μV, $p = 0.001$ and mean difference $= 0.828$, $p = 0.004$ for forward and backward syllables, respectively), and to geometrical shapes (mean difference $= -1.025$ μV, $p = 0.014$ and mean difference $= 0.813$ μV, $p = 0.0019$ for forward and backward syllables, respectively). Importantly, as illustrated in **Figure 3B**, a remarkable topographic difference was observed

TABLE 1 | Simple main effects of effector depressor on bilabial syllables.

Experiment 1 vs. experiment 2						
FvsB	**Electrode**	**Sum of squares**	**df**	**Mean square**	***F***	***p***
Forward	F3	23.582	1	23.582	4.233	**0.049***
	Fz	35.970	1	35.970	5.335	**0.029***
	F4	31.175	1	31.175	5.808	**0.023***
Backward	F3	2.230	1	2.230	0.548	0.466
	Fz	2.195	1	2.195	0.405	0.530
	F4	4.277	1	4.277	1.043	3.316

*$p < 0.05$.

TABLE 2 | Simple main effects of forward vs. backward displaying of bilabial syllables.

Forward vs. backward						
Effector depressor	**Electrode**	**Sum of squares**	**df**	**Mean square**	***F***	***p***
Experiment 1	F3	34.864	1	34.864	10.480	**0.003****
	Fz	38.014	1	38.014	9.140	**0.005****
	F4	39.363	1	39.363	8.393	**0.007****
Experiment 2	F3	0.198	1	0.198	0.110	0.743
	Fz	1.726	1	1.726	0.640	0.431
	F4	1.898	1	1.898	0.607	0.443

**$p < 0.01$.

between syllables and non-syllabic stimuli, the former presenting a robust P1000 component over fronto-central regions.

DISCUSSION

The present study attempted to elucidate (1) whether or not the electrophysiological dynamics underlying perceptual processing of orofacial movements are modulated by the linguistic content and visemic salience of visual speech information, and (2) to what extent interfering with orofacial articulatory effectors can affect this process. In line with the reviewed literature, we formulated three rationales. First, if the missing speech sounds can be synthesized on the basis of visemic information only, different patterns of electrophysiological responses should be observed for visual speech cues vs. non-speech orofacial movements, since the former, but not the latter, have an associated auditory counterpart (e.g., backward syllables are not pronounceable and thus should lack an associated articuleme). The latter would advocate for cross-modal predictions during speech perception. Second, in line with previous results, within speech related movements, those with greater visual salience should show a greater effect of cross modal prediction (e.g., bilabial vs. velar syllables). Third, if it is the case that, additionally to visemic and phonemic dimensions of speech, articulatory motor patterns also play an important role in speech processing, the effect of cross-modal predictions should be disrupted by the interference of articulatory effectors introduced in Experiment 2. This would support the hypothesis of a trimodal repertoire of speech perception.

We observed that the N400 component was elicited for the syllable condition. In line with previous studies, we interpret the modulation of N400 amplitude as indexing predictive coding during speech perception. Since stimuli were silent, the increasing amplitude of N400 for increasingly salient visemes (**Figure 2A**) could reflect the error in prediction caused by the absence of the corresponding phonemes. Supporting this interpretation, Chennu et al. (2016) reported a negative deflection similar to the mismatch negativity in response to omitted sounds (i.e., the omission effect), indicating the presence of top-down attentional processes that strengthens the brain's prediction of future events (Chennu et al., 2016). Congruently, in our data the conditions in which no auditory counterpart was expected (e.g., still faces) did not elicited N400. This cross-modal facilitation and predictive coding have also been shown by other experiments manipulating visemic salience, which show that greater predictability is evoked by visemes with higher visual salience (van Wassenhove et al., 2005; Paris et al., 2013, 2017). Additionally, Bourguignon et al. (2018, 2020) have shown that silent lip-reading generates neural entrainment to an absent auditory speech, even when participants do not know what the absent auditory signal should be. In this context, our results support the link between visemic and phonemic dimensions of speech in terms of predictive coding during perception.

Additional evidence for this comes from the presence of N400 elicitation for forward syllables, which have an expected auditory counterpart, but not for backward syllables (**Figure 2D**). Importantly, we only employed stop syllables, which cannot be uttered backwards. Thus, backward syllables in our experiment lacked an auditory counterpart. This is consistent with the lack of N400 evoked in this condition. Consistently, it has been reported that, "during the processing of silently played lip movements, the visual cortex tracks

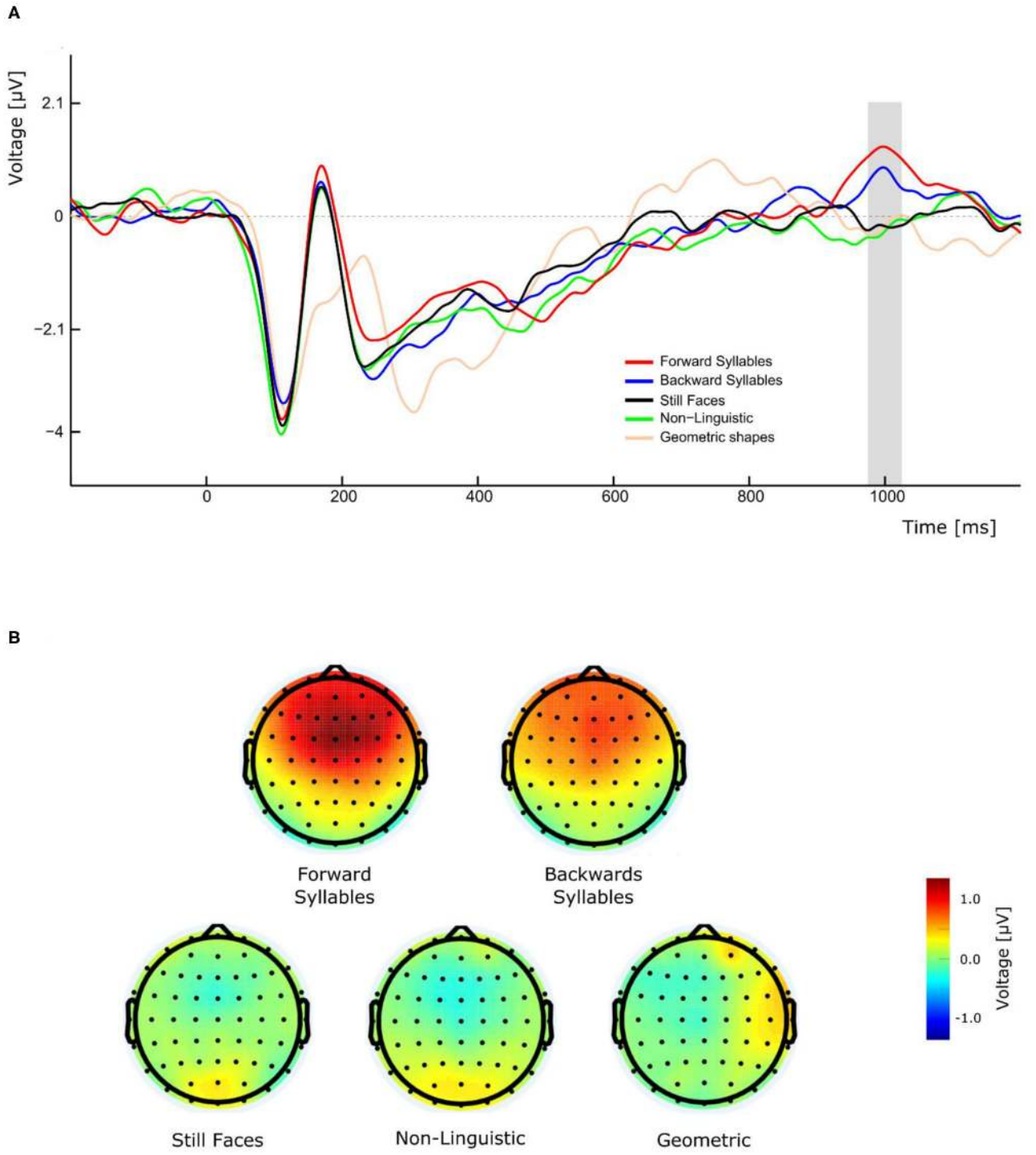

FIGURE 3 | Effect of conditions on P1000 component in experiment 1. **(A)** The gray bar represents the time window in which significant differences in amplitude were found. **(B)** Forward and backward syllables (top maps) and non-syllabic conditions (bottom maps) showed different topographic representations.

the missing acoustic speech information when played forward as compared to backward" (Hauswald et al., 2018). Our results strongly suggest that the visual perception of backward CVs that cannot be produced do not generate expectations about an associated speech sound. This result provides preliminary support for our hypothesized trimodal repertoire for speech, since the perceived motor articulations do not belong to the participants' motor repertoire, they

are not identifiable as articulemes and therefore, they lack audiovisual binding.

Directly in this line, another compelling argument supporting the trimodal network hypothesis is the effect of effector depressor on cross-modal speech prediction in our study. In Experiment 2, when orofacial effectors movements were restrained, the effect of cross-modal predictions was not observed (**Figure 2C**). Specifically, forward bilabial CVs ceased to elicit an N400 when the related motor effectors were disrupted by the effector depressor (**Table 2**). In a recent study (Martin et al., 2018), N400 amplitude was shown to increase in response to sentences containing unexpected target nouns compared to expected nouns, but the effect of expectation violation was not observable when speech production system was not available (i.e., when articulators were involved in a secondary task). The latter suggests that the availability of orofacial articulators is necessary for lexical prediction during lip-reading. The results of Experiment 2 support the idea of Martin et al. (2018) that speech effectors are important in generating speech predictions. However, since we used syllables and not words, the results of the current study further extend these results, suggesting that the motor involvement in speech predictions occur not only at the lexico-semantic level but also the pre-lexical level.

In addition to the expected N400, we also observed a late positive ERP component peaking around 1,000 ms, which was evoked only during the presentation of syllables (forward and backward) but not during any other condition. This effect is clearly illustrated by the topographical maps of the different conditions (**Figure 3B**). Importantly, this late positivity was not affected by the introduction of the effector depressor in Experiment 2. In the context of semantically unexpected sentence continuations, Van Petten and Luka (2012) reported that, following the N400, late positivities were elicited by low-plausible word completion. Similarly to the P1000 observed in the current study, these post-N400 positivities (PNP) are topographically distributed over frontal region and observed in time windows ranging from 600 to 1,200 ms after stimulus onset (deLong and Kutas, 2020). These anterior PNPs have recently been interpreted as a reintegration of the incorrectly predicted information in order to reach a new high-level interpretation (Kuperberg et al., 2020). In the context of our study, the elicitation of P1000 for speech related orofacial movements could be attributed to the reintegration and recuperation of the missing speech sounds. Even though this interpretation is more challenging to account for the presence of P1000 in response to backward CVs, it is possible that those stimuli have been re-interpreted as VCs (e.g., backward/ba/ being reinterpreted as/ab/). The latter, however, is more speculative and further studies are needed to clarify the functional significance of P1000 in this context. For instance, future research including a control experiment with full audio-visual stimuli could be helpful to disentangle this point.

To summarize, here we show that (1) electrophysiological dynamics underlying the perception of orofacial movements are modulated by the visemic salience of speech information, (2) when visemic salience is high (e.g., bilabial CVs) cross-modal prediction effects occur from visual to auditory modalities and (3) interfering with orofacial articulatory effectors can disrupt these feedforward processes. The current study, among an increasing body of evidence from the cognitive neuroscience literature on audiovisual speech processing and motor control of speech, points toward the necessity to rethink ecological speech perception beyond the auditory modality and include visual and motor systems in mechanistic explanations and neurobiological models of language.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of Pontifical Catholic University of Chile, School of Psychology. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

VL and MM conceived and planned the experiments. MM carried out the experiments. GB and MM performed the data analyses. VL, GB, and MM contributed to the interpretation of the results. MM took the lead in writing the manuscript. Nevertheless, all authors provided critical feedback and helped shape the research, analysis, and manuscript.

FUNDING

This research was supported by a post-doctoral fellowship from the Agencia Nacional de Investigación y Desarrollo (ANID) from the Chilean government to MM (Grant No. 3201057) and GB (Grant No. 3200248) and by a regular grant to VL (Grant No. 1150241).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnhum.2020.538619/full#supplementary-material

Supplementary Video 1 | Still face.

Supplementary Video 2 | Forward bilabial.

Supplementary Video 3 | Forward alveolar.

Supplementary Video 4 | Forward velar.

Supplementary Video 5 | Backward bilabial.

Supplementary Video 6 | Backward alveolar.

Supplementary Video 7 | Backward velar.

Supplementary Video 8 | Non-linguistic tongue.

Supplementary Video 9 | Non-linguistic lips.

Supplementary Video 10 | Geometrical shape.

REFERENCES

Archila-Meléndez, M. E., Valente, G., Correia, J. M., Rouhl, R. P., van Kranen-Mastenbroek, V. H., and Jansma, B. M. (2018). Sensorimotor representation of speech perception. Cross-decoding of place of articulation features during selective attention to syllables in 7T fMRI. *eNeuro* 5, 1–12. doi: 10.1523/ENEURO.0252-17.2018

Ardila, A., Akhutina, T. V., and Mikadze, Y. V. (2020). AR Luria's contribution to studies of the brain organization of language. *Neurol Neuropsychiatr Psychosomat.* 12, 4–12. doi: 10.14412/2074-2711-2020-1-4-12

Arnal, L. H., and Giraud, A. L. (2012). Cortical oscillations and sensory predictions. *Trends Cogn. Sci.* 16, 390–398. doi: 10.1016/j.tics.2012.05.003

Arnal, L. H., Morillon, B., Kell, C. A., and Giraud, A. L. (2009). Dual neural routing of visual facilitation in speech processing. *J. Neurosci.* 29, 13445–13453. doi: 10.1523/JNEUROSCI.3194-09.2009

Assaneo, M. F., and Poeppel, D. (2018). The coupling between auditory and motor cortices is rate-restricted: evidence for an intrinsic speech-motor rhythm. *Sci. Adv.* 4:eaao3842. doi: 10.1126/sciadv.aao3842

Auer, E. T., and Bernstein, L. E. (2007). Enhanced visual speech perception in individuals with early-onset hearing impairment. *J. Speech Lang. Hear. Res.* 50, 1157–65. doi: 10.1044/1092-4388(2007/080)

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

Bernstein, L. E., and Liebenthal, E. (2014). Neural pathways for visual speech perception. *Front. Neurosci.* 8:386. doi: 10.3389/fnins.2014.00386

Bernstein, L. E., Tucker, P. E., and Demorest, M. E. (2000). Speech perception without hearing. *Percept. Psychophys.* 62, 233–252. doi: 10.3758/BF03205546

Blank, H., and von Kriegstein, K. (2013). Mechanisms of enhancing visual speech recognition by prior auditory information. *Neuroimage* 65, 109–118. doi: 10.1016/j.neuroimage.2012.09.047

Bourguignon, M., Baart, M., Kapnoula, E. C., and Molinaro, N. (2018). Hearing through lip-reading: the brain synthesizes features of absent speech. *bioRxiv* 395483. doi: 10.1101/395483

Bourguignon, M., Baart, M., Kapnoula, E. C., and Molinaro, N. (2020). Lip-reading enables the brain to synthesize auditory features of unknown silent speech. *J. Neurosci.* 40, 1053–1065. doi: 10.1523/JNEUROSCI.1101-19.2019

Bruderer, A. G., Danielson, D. K., Kandhadai, P., and Werker, J. F. (2015). Sensorimotor influences on speech perception in infancy. *Proc. Natl. Acad. Sci. U.S.A.* 112, 13531–13536. doi: 10.1073/pnas.1508631112

Brunellière, A., Sánchez-García, C., Ikumi, N., and Soto-Faraco, S. (2013). Visual information constrains early and late stages of spoken-word recognition in sentence context. *Int. J. Psychophysiol.* 89, 136–147. doi: 10.1016/j.ijpsycho.2013.06.016

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., et al. (1997). Activation of auditory cortex during silent lipreading. *Science* 276, 593–596. doi: 10.1126/science.276.5312.593

Calvert, G. A., and Campbell, R. (2003). Reading speech from still and moving faces: the neural substrates of visible speech. *J. Cogn. Neurosci.* 15, 57–70. doi: 10.1162/089892903321107828

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., and Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Comput. Biol.* 5:e1000436. doi: 10.1371/journal.pcbi.1000436

Chennu, S., Noreika, V., Gueorguiev, D., Shtyrov, Y., Bekinschtein, T. A., and Henson, R. (2016). Silent expectations: dynamic causal modeling of cortical prediction and attention to sounds that weren't. *J. Neurosci.* 36, 8305–8316. doi: 10.1523/JNEUROSCI.1125-16.2016

Correia, J. M., Jansma, B. M., and Bonte, M. (2015). Decoding articulatory features from fMRI responses in dorsal speech regions. *J. Neurosci.* 35, 15015–15025. doi: 10.1523/JNEUROSCI.0977-15.2015

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., and Fadiga, L. (2009). The motor somatotopy of speech perception. *Curr. Biol.* 19, 381–385. doi: 10.1016/j.cub.2009.01.017

deLong, K. A., and Kutas, M. (2020). Comprehending surprising sentences: sensitivity of post-N400 positivities to contextual congruity and semantic relatedness. *Lang. Cogn. Neurosci.* 35, 1044–1063. doi: 10.1080/23273798.2019.1708960

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134, 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Dole, M., Méary, D., and Pascalis, O. (2017). Modifications of visual field asymmetries for face categorization in early deaf adults: a study with chimeric faces. *Front. Psychol.* 8:30. doi: 10.3389/fpsyg.2017.00030

Duffau, H. (2018). The error of broca: from the traditional localizationist concept to a connectomal anatomy of human brain. *J. Chem. Neuroanat.* 89, 73–81. doi: 10.1016/j.jchemneu.2017.04.003

Gambi, C., and Pickering, M. J. (2013). Prediction and imitation in speech. *Front. Psychol.* 4:340. doi: 10.3389/fpsyg.2013.00340

Garrod, S., and Pickering, M. J. (2015). The use of content and timing to predict turn transitions. *Front. Psychol.* 6:751. doi: 10.3389/fpsyg.2015.00751

Glenberg, A. M., and Gallese, V. (2012). Action-based language: a theory of language acquisition comprehension, and production. *Cortex* 48, 905–922. doi: 10.1016/j.cortex.2011.04.010

Golumbic, E. Z., Cogan, G. B., Schroeder, C. E., and Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". *J. Neurosci.* 33, 1417–1426. doi: 10.1523/JNEUROSCI.3675-12.2013

Gomez-Marin, A., and Ghazanfar, A. A. (2019). The life of behavior. *Neuron* 104, 25–36. doi: 10.1016/j.neuron.2019.09.017

Hauswald, A., Lithari, C., Collignon, O., Leonardelli, E., and Weisz, N. (2018). A visual cortical network for deriving phonological information from intelligible lip movements. *Curr. Biol.* 28, 1453–1459.e3. doi: 10.1016/j.cub.2018.03.044

Hickok, G., and Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition* 92, 67–99. doi: 10.1016/j.cognition.2003.10.011

Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. *Nat. Rev. Neurosci.* 8, 393–402. doi: 10.1038/nrn2113

Hirata, Y., and Kelly, S. D. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *J. Speech Lang. Hear. Res.* 53, 298–310. doi: 10.1044/1092-4388(2009/08-0243)

JASP Team. (2020). *JASP (Version 0.12.2)[Computer software].*

Jesse, A., and Massaro, D. W. (2010). The temporal distribution of information in audiovisual spoken-word identification. *Atten. Percept. Psychophys.* 72, 209–225. doi: 10.3758/APP.72.1.209

Kaganovich, N., and Ancel, E. (2019). Different neural processes underlie visual speech perception in school-age children and adults: An event-related potentials study. *J. Exp. Child Psychol.* 184, 98–122. doi: 10.1016/j.jecp.2019.03.009

Kaganovich, N., Schumaker, J., and Rowland, C. (2016). Matching heard and seen speech: an ERP study of audiovisual word recognition. *Brain Lang.* 157–158, 14–24. doi: 10.1016/j.bandl.2016.04.010

Kilner, J. M., Friston, K. J., and Frith, C. D. (2007). Predictive coding: an account of the mirror neuron system. *Cogn. Process* 8, 159–166. doi: 10.1007/s10339-007-0170-2

Kuperberg, G. R., Brothers, T., and Wlotko, E. W. (2020). A tale of two positivities and the N400: distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *J. Cogn. Neurosci.* 32, 12–35. doi: 10.1162/jocn_a_01465

Kutas, M., and Hillyard, S. A. (1980). Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biol. Psychol.* 11, 99–116. doi: 10.1016/0301-0511(80)90046-0

Letourneau, S. M., and Mitchell, T. V. (2013). Visual field bias in hearing and deaf adults during judgments of facial expression and identity. *Front. Psychol.* 4:319. doi: 10.3389/fpsyg.2013.00319

Lewkowicz, D. J., and Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proc. Natl. Acad. Sci. U.S.A.* 109, 1431–1436. doi: 10.1073/pnas.1114783109

Liberman, A. M., and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition* 21, 1–36. doi: 10.1016/0010-0277(85)90021-6

Lopez-Calderon, J., and Luck, S. J. (2014). ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Front. Hum. Neurosci.* 8:213. doi: 10.3389/fnhum.2014.00213

Luria, A. R. (1965). Aspects of aphasia. *J. Neurol. Sci.* 2, 278–287. doi: 10.1016/0022-510X(65)90112-7

Luria, A. R. (1973). Towards the mechanisms of naming disturbance. *Neuropsychologia* 11, 417–421. doi: 10.1016/0028-3932(73)90028-6

Martin, C. D., Branzi, F. M., and Bar, M. (2018). Prediction is production: the missing link between language production and comprehension. *Sci. Rep.* 8:1079. doi: 10.1038/s41598-018-19499-4

Massaro, D. W., and Chen, T. H. (2008). The motor theory of speech perception revisited. *Psychon. Bull. Rev.* 15, 453–457. doi: 10.3758/PBR.15.2.453

Maturana, H. R., and Varela, F. J. (1987). *The Tree of Knowledge: The Biological Roots of Human Understanding*. Boston, MA: New Science Library/Shambhala Publications.

McGurk, H., and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature* 264, 746–748. doi: 10.1038/264746a0

Michon, M., López, V., and Aboitiz, F. (2019). Origin and evolution of human speech: emergence from a trimodal auditory, visual and vocal network. *Prog. Brain Res.* 250, 345–371. doi: 10.1016/bs.pbr.2019.01.005

Navarra, J., and Soto-Faraco, S. (2005). Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. *Psychol. Res.* 71, 4–12. doi: 10.1007/s00426-005-0031-5

Nuttall, H. E., Kennedy-Higgins, D., Devlin, J. T., and Adank, P. (2018). Modulation of intra-and inter-hemispheric connectivity between primary and premotor cortex during speech perception. *Brain Lang.* 187, 74–82. doi: 10.1016/j.bandl.2017.12.002

Okada, K., Matchin, W., and Hickok, G. (2018). Neural evidence for predictive coding in auditory cortex during speech production. *Psychon. Bull. Rev.* 25, 423–430. doi: 10.3758/s13423-017-1284-x

Økland, H. S., Todorović, A., Lüttke, C. S., McQueen, J. M., and De Lange, F. P. (2018). Predicting audiovisual speech: early combined effects of sentential and visual constraints. *BioRxiv* 360578. doi: 10.1101/360578

Paris, T., Kim, J., and Davis, C. (2013). Visual speech form influences the speed of auditory speech processing. *Brain Lang.* 126, 350–356. doi: 10.1016/j.bandl.2013.06.008

Paris, T., Kim, J., and Davis, C. (2017). Visual form predictions facilitate auditory processing at the N1. *Neuroscience* 343, 157–164. doi: 10.1016/j.neuroscience.2016.09.023

Park, H., Kayser, C., Thut, G., and Gross, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *Elife* 5:e14521. doi: 10.7554/eLife.14521.018

Park, H., Thut, G., and Gross, J. (2018). Predictive entrainment of natural speech through two fronto-motor top-down channels. *Lang. Cogn. Neurosci.* 35, 739–751. doi: 10.1080/23273798.2018.1506589

Peelle, J. E. (2019). "*The neural basis for auditory and audiovisual speech* perception," in *The Routledge Handbook of Phonetics,* eds Katz and Assmann (New York, NY: Routledge).

Peelle, J. E., and Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex* 68, 169–181. doi: 10.1016/j.cortex.2015.03.006

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Front. Neuroinform.* 2:10. doi: 10.3389/neuro.11.010.2008

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., Tarkiainen, A., et al. (2005). Primary auditory cortex activation by visual speech: an fMRI study at 3 T. *Neuroreport* 16, 125–128. doi: 10.1097/00001756-200502080-00010

Pickering, M. J., and Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends Cogn. Sci.* 11, 105–110. doi: 10.1016/j.tics.2006.12.002

Pickering, M. J., and Garrod, S. (2013). An integrated theory of language production and comprehension. *Behav. Brain Sci.* 36, 329–347. doi: 10.1017/S0140525X12001495

Poeppel, D., and Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nat. Rev. Neurosci.* 21, 322–334. doi: 10.1038/s41583-020-0304-4

Pulvermüller, F., and Fadiga, L. (2010). Active perception: sensorimotor circuits as a cortical basis for language. *Nat. Rev. Neurosci.* 11, 351–360. doi: 10.1038/nrn2811

Pulvermüller, F., and Fadiga, L. (2016). "Brain language mechanisms built on action and perception," in *Neurobiology of Language* (Academic Press; Elsevier), 311–324. doi: 10.1016/C2011-0-07351-9

Pulvermüller, F., Huss, M., Kherif, F., del Prado Martin, F. M., Hauk, O., and Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proc. Natl. Acad. Sci. U.S.A.* 103, 7865–7870. doi: 10.1073/pnas.0509989103

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., and Foxe, J. J. (2006). Do you see what i am saying? *Exploring visual enhancement of speech comprehension in noisy environments. Cereb. Cortex* 17, 1147–1153. doi: 10.1093/cercor/bhl024

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. *Neurosci. Lett.* 127, 141–145. doi: 10.1016/0304-3940(91)90914-F

Sato, M., Buccino, G., Gentilucci, M., and Cattaneo, L. (2010). On the tip of the tongue: modulation of the primary motor cortex during audiovisual speech perception. *Speech Commun.* 52, 533–541. doi: 10.1016/j.specom.2009.12.004

Sebastián-Gallés, N., Albareda-Castellot, B., Weikum, W. M., and Werker, J. F. (2012). A bilingual advantage in visual language discrimination in infancy. *Psychol. Sci.* 23, 994–999. doi: 10.1177/0956797612436817

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends Cogn. Sci.* 17, 565–573. doi: 10.1016/j.tics.2013.09.007

Shahin, A. J., Backer, K. C., Rosenblum, L. D., and Kerlin, J. R. (2018). Neural mechanisms underlying cross-modal phonetic encoding. *J. Neurosci.* 38, 1835–1849. doi: 10.1523/JNEUROSCI.1566-17.2017

Skipper, J. I., Nusbaum, H. C., and Small, S. L. (2005). Listening to talking faces: motor cortical activation during speech perception. *Neuroimage* 25, 76–89. doi: 10.1016/j.neuroimage.2004.11.006

Stokes, R. C., Venezia, J. H., and Hickok, G. (2019). The motor system's [modest] contribution to speech perception. *Psychon. Bull. Rev.* 26, 1354–1366. doi: 10.3758/s13423-019-01580-2

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215. doi: 10.1121/1.1907309

Swaminathan, S., MacSweeney, M., Boyles, R., Waters, D., Watkins, K. E., and Möttönen, R. (2013). Motor excitability during visual perception of known and unknown spoken languages. *Brain Lang.* 126, 1–7. doi: 10.1016/j.bandl.2013.03.002

Ten Oever, S., Schroeder, C. E., Poeppel, D., van Atteveldt, N., and Zion-Golumbic, E. (2014). Rhythmicity and cross-modal temporal cues facilitate detection. *Neuropsychologia* 63, 43–50. doi: 10.1016/j.neuropsychologia.2014.08.008

Tenenbaum, E. J., Shah, R. J., Sobel, D. M., Malle, B. F., and Morgan, J. L. (2012). Increased focus on the mouth among infants in the first year of life: a longitudinal eye-tracking study. *Infancy* 18, 534–553. doi: 10.1111/j.1532-7078.2012.00135.x

Thompson, E., and Cosmelli, D. (2011). Brain in a vat or body in a world? Brainbound versus enactive views of experience. *Philos. Topics* 39, 163–180. doi: 10.5840/philtopics201139119

Tremblay, P., and Dick, A. S. (2016). Broca and Wernicke are dead, or moving past the classic model of language neurobiology. *Brain Lang.* 162, 60–71. doi: 10.1016/j.bandl.2016.08.004

Van Petten, C., and Luka, B. J. (2012). Prediction during language comprehension: benefits, costs, and ERP components. *Int. J. Psychophysiol.* 83, 176–190. doi: 10.1016/j.ijpsycho.2011.09.015

van Wassenhove, V. (2007). "Analysis-by-synthesis in auditory-visual speech perception: multi-sensory motor interfacing," in *Proceeding of 16th ICPhS*.

van Wassenhove, V. (2013). Speech through ears and eyes: interfacing the senses with the supramodal brain. *Front. Psychol.* 4:388. doi: 10.3389/fpsyg.2013.00388

van Wassenhove, V., Grant, K. W., and Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proc. Natl. Acad. Sci. U.S.A.* 102, 1181–1186. doi: 10.1073/pnas.0408949102

van Wassenhove, V., Grant, K. W., and Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia* 45, 598–607. doi: 10.1016/j.neuropsychologia.2006.01.001

Watkins, K. E., Strafella, A. P., and Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia* 41, 989–994. doi: 10.1016/S0028-3932(02)00316-0

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastian-Galles, N., and Werker, J. F. (2007). Visual language discrimination in infancy. *Science* 316, 1159–1159. doi: 10.1126/science.1137686

Worster, E., Pimperton, H., Ralph-Lewis, A., Monroy, L., Hulme, C., and MacSweeney, M. (2017). Eye movements during visual speech perception in deaf and hearing children. *Lang. Learn.* 68, 159–179. doi: 10.1111/lang.12264

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.