Do you see what I am saying? Electrophysiological dynamics of visual speech processing and the role of orofacial effectors.

Maëva Michon

Gonzalo Boncompte

Rodrigo Lagos

Vladimir López Hernández

Human speech perception has been prevalently studied focusing on auditory processing. However, visual and motor systems seem to play a more important role in speech perception than previously thought. The current study investigated the electrophysiological responses evoked by visual speech cues and other kind of orofacial movements and the role of automatic mimicry in speech versus non-speech visual perception. The results show that a) visual linguistic content and particularly the place of articulation of the syllables strongly modulated the electrophysiological responses and that b) this effect disappeared when automatic mimicry was interfered by asking the participants to hold an effector depressor between their teeth. These results support the idea that speech processing is multimodal and involves not only auditory but also visual and motor systems.

# 1. Introduction

Historically, the question of human language perception has been prevalently studied focusing on auditory processing of speech (Friederici, 2012). Recent investigations, however, postulate that language perception rather rely on an interactive multi-modal system, including not only auditory but also visual (Bernstein & Liebenthal, 2014) and motor systems (Pulvermüller & Fadiga, 2010; Glenberg & Gallese, 2012).

    The use of visual speech cues for language processing is present early in the ontogeny. Four month-old infants are capable to detect a switch from native to non-native language (and vice versa) in silent videos, suggesting that visual input alone is sufficient for language discrimination at these early ages (Sebastián-Gallés, Albareda-Castellot, Weikum, & Werker, 2012; Weikum et al., 2007). Further specializations seems to occur during the second half of the first year of life, when visual attention shifts from the eyes towards articulatory movements of the mouth, helping to construct a sensory-motor model for the emerging speech production (Tenenbaum, Shah, Sobel, Malle, & Morgan, 2012; Lewkowicz & Hansen-Tift, 2012). Studies with adults have demonstrated that having access to the visual information afforded by the interlocutor’s face can be especially advantageous in a noisy environment (Sumby & Pollack, 1954; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2006) and when hearing acuity is impaired (Bernstein, Tucker, & Demorest, 2000; Auer & Bernstein, 2007). Early audition deprivation leads to a greater dependence on vision during speech perception in deaf people, which is reflected behaviorally by a reorientation of visual attention in order to improve the perception of visual speech cues provided by orofacial movements (Dole, Méary, & Pascalis, 2017; Letourneau & Mitchell, 2013; Worster et al., 2017). On the other hand, it is known that adults often fail to hear the difference between certain non-native phonemic contrasts (when auditory only presented) but they do successfully distinguish these contrasts when presented audiovisually (Navarra & Soto-Faraco, 2005; Hirata & Kelly, 2010). Paris, Kim and Davis (2013) reported that the access to visual speech form speeds up the processing of auditory speech compared to when speech is presented in the auditory modality only. They argued that the temporal priority of visual speech may serve as a potential cue to predict aspects of up-coming auditory signal (Paris, Kim, & Davis, 2013).Interestingly, the more the articulatory movements are salient and predictive of a possible speech sound, the speediest auditory signal is processed. The authors propose that human adults possess “abstract internal representations” that link a specific visual form of the mouth to a restrained set of possible subsequent auditory input (van Wassenhove, Grant, & Poeppel, 2005). An alternative view to this abstract representational format would be emphasizing on the role of the motor system and the sensorimotor coupling as a mode of internal representation. The motor system seems to play an important role even in the most abstract forms of language (“Supplemental Material for Embodied Cognition: Is Activation of the Motor Cortex Essential for Understanding Action Verbs?”, 2017; Gallese & Cuccio, 2018; Kemmerer, 2014; Cardona et al., 2014). Abstract concepts activate the mouth motor representation in a way that has been interpreted as “a re-enactment of acquisition experience, or re-explanation of the word meaning, possibly through inner talk” (Borghi & Zarcone, 2016).

    Consistently with behavioral studies, neuroimaging techniques revealed that silent lip-reading activate areas of the temporal auditory cortex that overlap considerably with those activated by auditory speech perception. Noteworthy, auditory cortex appear to be similarly activated by visual pseudospeech in contrast to mouth movements with no linguistic content. Considered as a central hub for multimodal integration, the left posterior superior temporal sulcus (pSTS) is thought to have a crucial role in predicting upcoming auditory speech on the basis of visual information that typically precede the acoustic signal in a natural face to face conversation. For instance, greater functional connectivity has been found between left pSTS and auditory-speech areas when visual cue mismatch upcoming auditory cue, suggesting the existence of predictive error signals (Blank & von Kriegstein, 2013). Skipper, Nusbaum and Small (2005) used fMRI to examine brain activity associated with the comprehension of short stories presented in three different conditions: audiovisual, auditory-only and visual-only. They reported several interesting results. First, the activity of pSTS is modulated by the saliency of articulatory movements, becoming more active as visemic content increase. Second, Broca’s area and particularly of pars opercularis (BA 44) are activated to greater extent in the visual-only condition compared to the audiovisual condition. Based on their shared functional properties and connectivity, the authors suggest that pSTS and pars opercularis work together to associate the sensory patterns of phonemes and/or visemes with the motor commands needed to produce them. Finally, the activity in dorsal precentral gyrus and sulcus (i.e., premotor and motor cortices), similarly to the pSTS, is modulated by the amount of visemic content. These areas are postulated to be involved in the encoding of motor plans of the specific articulatory effectors (e.g., lips, tongue, jaws) corresponding to the sensori-motor representation generated by the pSTS and pars opercularis (Skipper, Nusbaum, & Small, 2005). Despite the fact that Broca’s area, premotor and motor cortices have traditionally been associated with language production, it seems that they also are an important part of a highly interactive network that “translate” orofacial movements into phonetic representation based on the motor commands required to generate those movements. We propose this network to support the development of a trimodal repertoire in which phoneme, viseme and ‘*articuleme*’ are linked to achieve a more ecological and seamless perception of speech.

    Whereas evidence of the spatial organization of the brain is increasingly robust and consistent, the temporal dimension of visual speech processing and its electrophysiological correlates remain poorly understood. The temporal dimension is crucial for audiovisual processing, as illustrated by the effects of desynchronization between auditory and visual speech inputs, but also because visual speech cues are perceived first and have the potential to disambiguate the upcoming acoustic signal. The high temporal resolution of EEG techniques makes them especially suited to address such temporal dynamic questions. In the current study, two experiments were performed. The first experiment aimed to elucidate whether or not the linguistic content of visual speech cues modulates the electrophysiological response elicited by perceiving orofacial movements. We recorded participant’s EEG signal while they attentively observe or imitate different type of orofacial movements (a- still mouth, b-syllables, c-backward played syllables, d-non-linguistic movements) and non-biological movements displayed in short videos. The second experiment aimed to investigate to what extent interfering with automatic mimicry can affect the electrophysiological dynamics underlying orofacial movements processing. To do so, the very same experiment was run a second time but participants were asked to hold an effector depressor between their teeth while observing the videos.

# 2. Methods

## 2.1. Stimuli

The stimuli consisted in a set of 120 videos displaying different type of orofacial movements (1- still mouth, 2-syllables, 3-backward played syllables, 4-non-linguistic movements) or non-biological movements (5- non-human). In the first condition, no mouth movements were produced (Baseline). In the second, 3 type of syllables were produced differing in their place of articulation (PoA). Bilabial syllables ([pa] [ba]) requires lip movements whereas alveolar ([da] [ta]) and velar ([ga] [ka]) syllables require upper and lower tongue movements, respectively. These consonants have been chosen because they have the common characteristic of being stop consonants, which mean that they are articulated by closing the airway so as to impede the flow of air, by maintaining airway closed thus generating a slightly pressure because of accumulated air and finally by opening the airway and releasing the airflow producing in that way an audible sound. Importantly, these three kind of syllables have been reported to have different levels of visual salience (Jesse & Massaro, 2010; Paris, Kim, & Davis, 2013; van Wassenhove, Grant, & Poeppel, 2007). In the third condition, the same syllables were played backward. Because of their particular motoric sequence, stop syllables can not be pronounced backward. In that sense, backward played syllables represent an ideal control condition because this kind of articulatory movements are visually very similar to speech but at the same time they are not pronounceable, they do not belong to our hypothesized motor repertoire. In the forth condition, non-linguistic orofacial movements (e.g., tongue protusion, lip-smacking) were produced. This condition was introduced in order to control for the activity associated to the processing of orofacial movements with no linguistic content.  Finally, in order to control for general movement perception, independently of its biological and facial related nature, a fifth condition was added where movements of different geometrical figures (e.g., ovals, squares, triangles) were shown. These stimuli were generated using PsychoPy toolbox. Importantly, all the videos were silently displayed (i.e., audio removed) and only show the lower part of the actor’s face in order to ensure that his eyes movements could not interfere. Videos were 2 seconds long (M=2052ms and SD=59ms), they started with 10 frames displaying a closed, still mouth.

## 2.2. Participants

34 right-handed subjects (22 females) with normal or corrected-to-normal vision and hearing and without any history of psychiatric or neurological disorders performed the experiments. Participant’s age range from 18 to 36 years old (M=22,8 and SD 4,2 years). The experimental protocol was approved by the Ethics Committee of Pontifical Catholic University of Chile, School of Psychology. Before the experiment started each participant was explained the procedure and signed an informed consent form. Four participants were removed from final analysis because of poor signal-to-noise ratio.

## 2.3. Procedure

Participants sat at a distance of approximately 70 cm from the computer screen and were asked to attentively observe or imitate the movements shown in the videos.  Stimuli were displayed on the screen using PsychoPy toolbox. The trial started with a word lasting for 500ms that indicate the instruction, either “Observe” or “Imitate”. After 100 to 150ms,  a fixation cross appeared for 250ms. In the observation condition, the video was displayed one time, 1000 to 1500ms 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 seconds. In order to study the role of automatic mimicry the very same  experiment was repeated in Experiment 2 (B2) where participants were asked to hold an effector depressor horizontally between their teeth (i.e. in the imitation trials, participants were asked to remove the depressor when the word “Imitate” appeared so they can properly imitate). This procedure allowed to impede the automatic mimicry. The order of Experiment 1 and 2 was counterbalanced between participants.

Each of the 5 conditions [i.e., 1) still mouth, 2) syllables, 3)backward played syllables, 4) non-linguistic mouth movements and 5) non-human movements] consist in 3 repetitions of the 24 video-clips, leading to a total of 72 trials per condition (360 per experiment). The experimental design was an intra-subject 5 (type of movements) X 2 (without/with effector restriction) factorial design.

## 2.4.  Electroencephalographic recording parameters

Electrophysiological activity was register with a 64-channels (Ag-AgCl) EEG system (Biosemi ® ActiveTwo) with electrodes positioned according to the extended 10-20 international system. The sampling rate was of 2048 Hz (band-passed 0.1 to 100Hz). Four external electrodes were used to monitor eye movements. Two of them were placed in the outer canthi of the eyes in order 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 on bilateral mastoids for re-referenciation. Data pre-processing was performed using MatLab (The Mathworks, Inc.), EEGLAB (Delorme & Makeig, 2004) and ERPLAB toolbox (Lopez-Calderon & Luck, 2014). The signal was first down-sampled at a rate of 512 Hz, re-referenced to mastoides and band-passed between 0.1 and 40 Hz for ERP analysis. Then, EEG signal was segmented into epochs from -500ms to 1500ms respective to stimulus onset. Each epoch was visually inspected in order to reject large artifacts due to head movements and muscular artifacts. After Independant Component Analysis (ICA) decomposition and the rejection of components typically associated with eye-blinking, epochs exceeding maximum amplitude of ±100 μV were removed.    + ANALYSIS DE FREQ THETA

## 2.5. Statistical analysis

The ERP components of interest for statistical analyses were P2, N270, N400 and a positivity around 1000 ms. Using the ERP measurement tool in ERPLAB, mean amplitudes were calculated with respect to a 50 ms prestimulus baseline (un poco chico el baseline, pq tan chico?) for the following selected time windows: P2 [155-185 ms], N270 [245-295 ms], N400 [475-525 ms] and P1000 [975-1025 ms]. After mean amplitudes were extracted for each condition, data were analyzed using a mixed model for each experiement (1 and 2), with conditions as fixed factor and subjects as aleatory factor. Statistical analysis was performed using the *lme4* (Bates, Mächler, Bolker, & Walker, 2015) package of R. The pairwaise comparision within conditions was optained using *emmeans (Lenth, 2016)*package of R, that permit to compare slopes in a mix model.      + Porqué t distribution (cf. Rodrigo Lagos)

**Agregar Theta como indicador**

**+ d de cohen reportar 0.5 see G-power**

# 3. Results

## 3.1. Electrophysiological results

### 3.1.1. Experiment 1

The evoked responses of all conditions can be observed in Figure 1a. It shows that the evocation of the components P2, N270, N400 and P1000. Upon closer inspection of these waveforms it can be seen that the amplitude of the late positive potential observed around 1000ms differentiate between linguistic and non-linguistic visual stimuli. Particularly, the amplitude in alveolar syllables was significantly higher compared with still-mouth and orofacial gestures in electrode FCz (t=-3.392, p=0.0148, **ETA** and  t=3.593, p=0.0076, **ETA** respectively).  Moreover, the topographical representations for syllables and backward syllables notably differ from those of still-mouth and non-linguistic gestures. The former ones were associated with a centro-frontal activity while the activity of the latter ones were more located in posterior-occipital regions (Fig.1b).



ERPs across all conditions (electrode Fcz)

###

Since ERPs responses were strongly modulate by the PoA of the syllables, we decided to treat bilabial, alveolar and velar syllables independently for further analysis, rather than collapsing them into the syllable condition. Bilabial syllables elicited greater amplitude in N270 compared with velar syllables (t= -3.993, p=0.0018 **ETA**) and in N400 compared with velar syllables (t= -4,804, p = 0.0001 **ETA**) and alveolar syllables (t= -3,174, p=0.0290 **ETA**). No significant differences were found for components P2 and P1000 (Fig.2).



ERPs of the syllables as a function of their PoA (electrode FCz).

### 3.1.2. Experiment 2

The amplitude of the N400 component was significantly greater for the bilabial syllables in experiment 1 compared to experiment 2 (t= 2.186, p= 0.037 **ETA**) ), whereas the amplitude of N270 didn’t differ between the two experiments. No significant differences were found between experiment 1 and 2 for the other type of syllables nor for non-linguistic movements.



The effect of effector depression on bilabial syllables processing(electrode FCz).

# 4. Discussion

     The present study was intended to determine whether or not the electrophysiological dynamics underlying perceptual processing of orofacial movements are modulated by the linguistic content of visual speech cues and to what extent interfering with automatic mimicry can affect this process. We reported two main electrophysiological findings. First, early ERPs amplitudes were clearly modulated by linguistic content and more specifically by the PoA of the syllables. Second, the effect of the PoA was significantly reduced by the effector depression.

    The first component modulated by the PoA was N270. That negative deflection peaking around 270 ms have previously been associated with conflicts in audiovisual integration. For instance, Wang et al. (2002b) reported a modulation of this component in a task where the gender of a visually presented face mismatched the gender of a voice pronouncing a vowel. The amplitude of N270 increased in response to audiovisual incongruity between face and voice gender (Wang, Wang, Cui, Tian, & Zhang, 2002). Another study reported that the N270 was elicited in the presence of an audiovisual incongruity independently of its relevance for task solving. The authors concluded that “this component reflects the activity of a conflict detection process of automatic nature, rather resistant to top-down influences from voluntary attention networks” (ORTEGA, LÓPEZ, & ABOITIZ, 2008). More recently, Chennu et al. (2016) reported a negative deflection similar to the mismatch negativity effect in response to omitted sounds (i.e., the omission effect) indicating the presence of top-down attentional processes that strenghtens the brain’s prediction of future events (Chennu et al., 2016). In the current study, we interpret the N270 as marker of audiovisual inconsistency, in the sense that participants perceived a mouth saliently articulating a syllable but they never heard the corresponding speech sound. Congruently with this interpretation, it has been reported that, “during the processing of silently played lip movements, the visual cortex tracks the missing acoustic speech information when played forward as compared to backward” (Hauswald, Lithari, Collignon, Leonardelli, & Weisz, 2018). Interestingly, the most significant differences observed in this study were between bilabial and velar syllables. Those syllables have very different PoA, the former being performed with a clear movement of the lips and the latter being performed by a nearly imperceptible movement of the lower tongue. In that sense,  the greater amplitude of N270 observed for bilabial syllables compared to other syllables may be attributable to their different degrees of visual salience. The absence of differences in the amplitude of N270 between experiments 1 and 2 suggest that the restrained mobility of the upper articulatory system (i.e., lips and tongue) does not impair the detection of the crossmodal conflict induced by the omission of the auditory couterpart.

    The second component modulated by the PoA was N400. This negativity is traditionally associated with semantic incongruity processing. Interestingly, a recent study reported that when visual speech (i.e., silent articulations) was incongruent with preceding auditory words a significantly larger N400 was elicited compared to congruent conditions, suggesting the detection of the auditory-articulatory mismatch (Kaganovich, Schumaker, & Rowland, 2016). In our study however, the stimuli were visually and silently displayed. Thus, rather than an auditory-articulatory mismatch, the larger amplitude of N400 could be interpreted as  a response to the conflict caused by the missing auditory counterpart of syllables articulation. Supporting this interpretation, no N270 or N400 components were elicited in response to backward syllables probably because they are not pronounceable so they lack auditory and motoric counterparts. An alternative interpretation can be formulated on the basis of studies suggesting that rather than being an index of semantic incongruity,  N400 reflects errors in speech prediction. A recent study showed that its amplitude increases in response to sentences containing unexpected target nouns compared to expected nouns. Importantly, the effect expectation violation as indexed by N400 amplitude was not observable when speech production system was not available (i.e., when articulators were suppressed). The latter suggest that the availability orofacial articulators is necessary for lexical prediction during reading (Martin, Branzi, & Bar, 2018). In line with these results, we observed a significant difference in the amplitude of N400 for bilabial syllables between Experiment 1 and Experiment 2 when speech effectors, mostly the lips, were blocked. Bilabial articulatory movements are more visible and salient, as a consequence the subsequent auditory cues is more predictable. So, when it is not perceived, the effect of expectation violation is greater.In contrast, when speech articulators were restrained this effect was not observed. In that sense, our results support the idea of Martin et al. (2018) that speech effectors play a critical role in generating speech predictions. But, because we used syllables and not words, the results of the current study demonstrate that speech predictions are generated as early as the pre-lexical level.

    We hypothesize that the mechanism underlying the ability to make speech predictions on the basis of articulatory movements is automatic imitation. Several authors have proposed that listeners covertly imitate speaker’s orofacial gestures during face to face interactions allowing them to construct up-dating foward model and make predictions about the up-coming speech (Pickering & Garrod, 2013; Gambi & Pickering, 2013; Brass & Heyes, 2005). Although the absence of prediction error effect **AND THETA** in Experiment 2 strongly suggests that orofacial movements have a key role in speech perception, the data analysis performed in this study are not directly investigating the involement of motor systems. Further studies analyzing time-frequency domain or electromyographic responses of orofacial muscles could be more conclusive in that respect.

# Acknowledgements

This research was founded by a PhD fellowship from the Consejo Nacional de Investigación Científica y Tecnológica (CONICYT) from the chilean governement.

# References

Friederici, A. D. (2012). The cortical language circuit: from auditory perception to sentence comprehension. *Trends in Cognitive Sciences*, *16*(5), 262–268. <http://doi.org/10.1016/j.tics.2012.04.001>

Bernstein, L. E., & Liebenthal, E. (2014). Neural pathways for visual speech perception. *Frontiers in Neuroscience*, *8*. <http://doi.org/10.3389/fnins.2014.00386>

Pulvermüller, F., & Fadiga, L. (2010). Active perception: sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience*, *11*(5), 351–360. <http://doi.org/10.1038/nrn2811>

Glenberg, A. M., & Gallese, V. (2012). Action-based language: A theory of language acquisition comprehension, and production. *Cortex*, *48*(7), 905–922. <http://doi.org/10.1016/j.cortex.2011.04.010>

Sebastián-Gallés, N., Albareda-Castellot, B., Weikum, W. M., & Werker, J. F. (2012). A Bilingual Advantage in Visual Language Discrimination in Infancy. *Psychological Science*, *23*(9), 994–999. <http://doi.org/10.1177/0956797612436817>

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastian-Galles, N., & Werker, J. F. (2007). Visual Language Discrimination in Infancy. *Science*, *316*(5828), 1159–1159. <http://doi.org/10.1126/science.1137686>

Tenenbaum, E. J., Shah, R. J., Sobel, D. M., Malle, B. F., & Morgan, J. L. (2012). Increased Focus on the Mouth Among Infants in the First Year of Life: A Longitudinal Eye-Tracking Study. *Infancy*, *18*(4), 534–553. <http://doi.org/10.1111/j.1532-7078.2012.00135.x>

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences*, *109*(5), 1431–1436. <http://doi.org/10.1073/pnas.1114783109>

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, *26*(2), 212–215. <http://doi.org/10.1121/1.1907309>

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2006). Do You See What I Am Saying? Exploring Visual Enhancement of Speech Comprehension in Noisy Environments. *Cerebral Cortex*, *17*(5), 1147–1153. <http://doi.org/10.1093/cercor/bhl024>

Bernstein, L. E., Tucker, P. E., & Demorest, M. E. (2000). Speech perception without hearing. *Perception & Psychophysics*, *62*(2), 233–252. <http://doi.org/10.3758/bf03205546>

Auer, E. T., & Bernstein, L. E. (2007). Enhanced Visual Speech Perception in Individuals With Early-Onset Hearing Impairment. *Journal of Speech Language and Hearing Research*, *50*(5), 1157. [http://doi.org/10.1044/1092-4388(2007/080)](http://doi.org/10.1044/1092-4388%282007/080%29)

Dole, M., Méary, D., & Pascalis, O. (2017). Modifications of Visual Field Asymmetries for Face Categorization in Early Deaf Adults: A Study With Chimeric Faces. *Frontiers in Psychology*, *8*. <http://doi.org/10.3389/fpsyg.2017.00030>

Letourneau, S. M., & Mitchell, T. V. (2013). Visual field bias in hearing and deaf adults during judgments of facial expression and identity. *Frontiers in Psychology*, *4*. <http://doi.org/10.3389/fpsyg.2013.00319>

Worster, E., Pimperton, H., Ralph-Lewis, A., Monroy, L., Hulme, C., & MacSweeney, M. (2017). Eye Movements During Visual Speech Perception in Deaf and Hearing Children. *Language Learning*, *68*, 159–179. <http://doi.org/10.1111/lang.12264>

Navarra, J., & Soto-Faraco, S. (2005). Hearing lips in a second language: visual articulatory information enables the perception of second language sounds. *Psychological Research*, *71*(1), 4–12. <http://doi.org/10.1007/s00426-005-0031-5>

Hirata, Y., & Kelly, S. D. (2010). Effects of Lips and Hands on Auditory Learning of Second-Language Speech Sounds. *Journal of Speech Language, and Hearing Research*, *53*(2), 298–310. [http://doi.org/10.1044/1092-4388(2009/08-0243)](http://doi.org/10.1044/1092-4388%282009/08-0243%29)

Paris, T., Kim, J., & Davis, C. (2013). Visual speech form influences the speed of auditory speech processing. *Brain and Language*, *126*(3), 350–356. <http://doi.org/10.1016/j.bandl.2013.06.008>

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences*, *102*(4), 1181–1186. <http://doi.org/10.1073/pnas.0408949102>

(2017). *Journal of Experimental Psychology: Learning Memory, and Cognition*. <http://doi.org/10.1037/xlm0000451.supp>

Gallese, V., & Cuccio, V. (2018). The neural exploitation hypothesis and its implications for an embodied approach to language and cognition: Insights from the study of action verbs processing and motor disorders in Parkinsons disease. *Cortex*, *100*, 215–225. <http://doi.org/10.1016/j.cortex.2018.01.010>

Kemmerer, D. (2014). Visual and Motor Features of the Meanings of Action Verbs: A Cognitive Neuroscience Perspective. In *Cognitive Science Perspectives on Verb Representation and Processing* (pp. 189–212). Springer International Publishing. <http://doi.org/10.1007/978-3-319-10112-5_9>

Cardona, J. F., Kargieman, L., Sinay, V., Gershanik, O., Gelormini, C., Amoruso, L., … Ibáñez, A. (2014). How embodied is action language? Neurological evidence from motor diseases. *Cognition*, *131*(2), 311–322. <http://doi.org/10.1016/j.cognition.2014.02.001>

Borghi, A. M., & Zarcone, E. (2016). Grounding Abstractness: Abstract Concepts and the Activation of the Mouth. *Frontiers in Psychology*, *7*. <http://doi.org/10.3389/fpsyg.2016.01498>

Blank, H., & von Kriegstein, K. (2013). Mechanisms of enhancing visualspeech recognition by prior auditory information. *NeuroImage*, *65*, 109–118. <http://doi.org/10.1016/j.neuroimage.2012.09.047>

Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: motor cortical activation during speech perception. *NeuroImage*, *25*(1), 76–89. <http://doi.org/10.1016/j.neuroimage.2004.11.006>

Jesse, A., & Massaro, D. W. (2010). The temporal distribution of information in audiovisual spoken-word identification. *Attention Perception, & Psychophysics*, *72*(1), 209–225. <http://doi.org/10.3758/app.72.1.209>

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, *45*(3), 598–607. <http://doi.org/10.1016/j.neuropsychologia.2006.01.001>

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. <http://doi.org/10.1016/j.jneumeth.2003.10.009>

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*. <http://doi.org/10.3389/fnhum.2014.00213>

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Usinglme4. *Journal of Statistical Software*, *67*(1). <http://doi.org/10.18637/jss.v067.i01>

Lenth, R. V. (2016). Least-Squares Means: TheRPackagelsmeans. *Journal of Statistical Software*, *69*(1). <http://doi.org/10.18637/jss.v069.i01>

Wang, Y., Wang, H., Cui, L., Tian, S., & Zhang, Y. (2002). The N270 component of the event-related potential reflects supramodal conflict processing in humans. *Neuroscience Letters*, *332*(1), 25–28. [http://doi.org/10.1016/s0304-3940(02)00906-0](http://doi.org/10.1016/s0304-3940%2802%2900906-0)

ORTEGA, R. O. D. R. I. G. O., LÓPEZ, V. L. A. D. I. M. I. R., & ABOITIZ, F. R. A. N. C. I. S. C. O. (2008). Voluntary modulations of attention in a semantic auditory-visual matching Task: an ERP study. *Biological Research*, *41*(4). <http://doi.org/10.4067/s0716-97602008000400010>

Chennu, S., Noreika, V., Gueorguiev, D., Shtyrov, Y., Bekinschtein, T. A., & Henson, R. (2016). Silent Expectations: Dynamic Causal Modeling of Cortical Prediction and Attention to Sounds That Werent. *Journal of Neuroscience*, *36*(32), 8305–8316. <http://doi.org/10.1523/jneurosci.1125-16.2016>

Hauswald, A., Lithari, C., Collignon, O., Leonardelli, E., & Weisz, N. (2018). A Visual Cortical Network for Deriving Phonological Information from Intelligible Lip Movements. *Current Biology*, *28*(9), 1453–1459.e3. <http://doi.org/10.1016/j.cub.2018.03.044>

Kaganovich, N., Schumaker, J., & Rowland, C. (2016). Matching heard and seen speech: An ERP study of audiovisual word recognition. *Brain and Language*, *157-158*, 14–24. <http://doi.org/10.1016/j.bandl.2016.04.010>

Martin, C. D., Branzi, F. M., & Bar, M. (2018). Prediction is Production: The missing link between language production and comprehension. *Scientific Reports*, *8*(1). <http://doi.org/10.1038/s41598-018-19499-4>

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(04), 329–347. <http://doi.org/10.1017/s0140525x12001495>

Gambi, C., & Pickering, M. J. (2013). Prediction and imitation in speech. *Frontiers in Psychology*, *4*. <http://doi.org/10.3389/fpsyg.2013.00340>

Brass, M., & Heyes, C. (2005). Imitation: is cognitive neuroscience solving the correspondence problem?. *Trends in Cognitive Sciences*, *9*(10), 489–495. <http://doi.org/10.1016/j.tics.2005.08.007>