FACULTAD DE CIENCIAS SOCIALES
ESCUELA DE PSICOLOGÍA

DO YOU SEE WHAT I AM SAYING?

ELECTROPHYSIOLOGICAL DYNAMICS OF VISUAL SPEECH PROCESSING AND THE ROLE OF OROFACIAL EFFECTORS FOR CROSS-MODAL PREDICTIONS.

POR

MAËVA MICHON DESBIEY

Tesis presentada a la Escuela de Psicología de la Pontificia Universidad Católica de Chile para optar al grado académico de Doctor en Psicología

Director de Tesis: Vladimir López Hernández
Comité de Tesis: Francisco Aboitiz, Edmundo Kronmüller, David Huepe

Junio 2019
Santiago, Chile

A mi musa, Noa.

AKNOWLEDGMENTS

In the first place, I would like to address my deepest gratitude to my advisor Vladimir López who have supported me unconditionally from the very beginning of this adventure and well before that. I feel so lucky for having the opportunity to work with you over the past 4 years. I have tried to take the greatest benefits of your expertise and of your human qualities. I really enjoyed every meeting, every conversation with you and your endless anecdotes. I am glad to say that, thank to you, I close this chapter of my life with the sensation to be a better scientist, a better professor and a better person.

I am also indebted to my thesis committee, Edmundo Kronmüller, David Huepe and especially, Francisco Aboitiz. I really appreciated your insightful suggestions and contributions to this project. You all showed a sincere enthusiasm respect to a central and audacious concept of my work, the *articuleme*. Francisco, I want you to know that I am sincerely grateful for the opportunities you gave me to introduce my ideas to the scientific community and to give the trimodal repertoire for speech a chance to mature and gain robustness in your lab.

Gonzalo Boncompte, I clearly own you a special mention. How many coffees, how many hours, how many error line 139? Thanks to you now I can say I am a multilingual person. I speak French, Spanish, English and Matlab. More seriously, I admire your patience and your good disposition to help me dealing with the programing, the statistics, the emotional storms and existential crises. I will eternally be grateful to you my dear friend.

I thank my family, my parents and sisters, for their support in every moment and despite the distance. Hernán, my love, thank you for being there in the ups and downs, for believing in me no matter what and to make sure I don't solely feed myself with eggs and cheese. Last but not least, thank you Noa, my son, my sunshine, my inspiration, thank you for understanding and making me feel the best mum anyway.

Table of Contents

Illustration Index

Abstract

The need of a comprehensive model to account for the neurobiology of language has quite a long history. New empirical data have challenged classical models. A new framework is needed to foster our understanding of brain mechanisms underlying speech perception and production. In this dissertation, we postulate the existence of a trimodal network for speech perception that emphasizes the importance of visual processing of speech related orofacial movements and its representation in motor cortices. In that sense, we hypothesize that auditory (phonemes), visual (visemes) and motor (articulemes) aspects of speech are bonded in a trimodal repertoire.

In order to test our hypothesis, we recorded EEG signal while participants were attentively observing different type of linguistic and non-linguistic orofacial movements in two conditions: under normal observation and observation holding a speech effector depressor horizontally between their teeth.

ERPs analyses of the signal provide evidence of cross-modal predictions indexed by the N270 and the N400-like components. The amplitude of these components was specifically modulated by the visual salience of visual speech cues; the more salient the more predictable. Interestingly, when orofacial effectors were restricted, the amplitude of N400 was significantly reduced, suggesting that language production system is recruited for predictions. The time-frequency analysis, on the other hand, demonstrated the involvement of motor cortices for visual speech perception. More specifically, a significant difference in the μ-suppression was observed between linguistic and no-linguistic orofacial movements. The power of the μ-suppression was modulated by visual salience but diminished for the more salient visual speech cues when the participants orofacial effectors were blocked.

The results reported in this dissertation represent preliminary evidence of the existence of the proposed trimodal network and, in particular, of the articuleme. Undoubtedly, further research using complementary neuroimaging techniques are

required to better understand this multimodal interplay during language perception and production.

List of Abbreviations

AF: Arcuate Fascicule

aVLPFC: anterior VentroLateral PreFrontal Cortex

BA: Broadmann Area

DTI: Diffusion Tensor Imaging

ECoG: ElectroCorticalGraphy

EEG: Electroencephalography

EOG: Electrooculography

ERPs: Event-Related Potentials

FFT: Fast Fourier Transform

fMRI: functional Magnetic Resonance Imaging

ICA: Independent Component Analysis

IFG: Inferior Frontal Gyrus

IPL: Inferior Parietal Lobe

LMC: Laryngeal Motor Cortex

MEG: Magnetoencephalography

MNS: Mirror Neuron System

MTG: Medial Temporal Gyrus

NHP: Non-Human Primate

PoA: Place of Articulation

pSTS: posterior Superior Temporal Sulcus

SLF: Superior Longitudinal Fascicule

STG: Superior Temporal Gyrus

STS: Superior Temporal Sulcus

TF : Time-Frequency

V5/MT: movement-sensitive visual cortex

VLPFC: VentroLateral PreFrontal Cortex

vPMC: ventral PreMotor Cortex

WLG: Wernicke-Lichtheim-Geschwind

Chapter I

Introduction

« On ne peut réduire la langue au son, ni détacher le son de l'articulation buccale. Réciproquement, on ne peut pas définir les mouvements des organes vocaux si l'on fait abstraction de l'impression acoustique. »

"You can not reduce the language to the sound, or detach the sound of the oral articulation. Reciprocally, we can not define the movements of the vocal organs if we ignore the acoustic impression"

Ferdinand de Saussure. *Cours de Linguistique Générale* (1916)
Personal translation

1. General introduction

Language is one of the most sophisticated outcomes in the evolution of the human brain. As social and communicative beings, humans differed from other non-human primates in their capacity to construct, develop and manipulate a vast repertoire of symbols (i.e., spoken words but also written or signed symbols) that are associated to objects or concepts in an arbitrary way. Such a repertoire has afforded the man with a unique and powerful ability to communicate complex ideas, contributing in turn to the emergence of organized societies. For more than two centuries, mankind has shown a genuine interest to uncover the mystery of how we came to manage languages.

1.1. From classical models to contemporary approach of the neurobiology of language.

In the 19th century, the pioneering neuropsychological work of Paul Broca and Carl Wernicke on patients with brain lesions evidenced a clear demarcation between speech production and perception. These findings have had major and longstanding consequences on our theoretical approach of language, as relying on two functionally and anatomically segregated brain regions: the so-called "Broca's area" in the inferior frontal gyrus (IFG) defined as the seat of speech production and "Wernicke's area" encompassing the superior temporal gyrus (STG) underlying speech comprehension. Recently, an increasing movement in the specialized scientist community is attempting to warn their pairs that the resilience of a 150 years-old model of language neurobiology is compromising the advancement of the field (Iacoboni & Wilson, 2006; Cogan et al., 2014; Friedrich et al., 2018). Despite many contemporary scientists of the field agree that the classic model of language neurobiology has served an important heuristic function, they also admit it is now outdated. Still, the terminology inherited by this model remains widely used in papers, classrooms and among clinical experts (Tremblay & Dick, 2016). Referring to brain areas associated to language production and perception as "Broca and Wernicke's areas", respectively, is problematic for multiple reasons. First, this terminology unnoticeably sustains the dogma of localizationism and modularity of cognitive and brain functions. Also, the most relevant problem for our purpose is that those two "language epicenters" (Papathanassiou et al., 2000) have dangerously led to a functional reductionism. Finally, this perception/production dichotomy is particularly misleading when it comes to integrate the growing body of evidences that consistently suggests that speech perception and production both recruit distributed regions of the brain including frontal, temporal but also inferior parietal regions (Peelle, 2019; Fridiksson, 2010, Romanski, 2007).

Following a behavioral-structural correlational approach ("lesions method"), Lichtheim's model (Lichtheim, 1885), later retaken by Geschwind in a model often referred to as the Wernicke-Lichtheim_Geschwind (WLG) model (Hagoort, 2016), introduced the notion that language is represented in the brain in three centers: an auditory language center (O), a semantic center (S) and a motor language center (π). It is noteworthy that interactions among the centers are achieved in a unidirectional fashion, with no possible direct interaction from motor to auditory language centers (see Figure 1). Evidence from patient studies, computational models and cognitive neuroscience has shown this model to be most likely erroneous. Perhaps, the most striking evidence is that an "extensive resection of tumors such as [diffuse low-grade glioma] situated within the pars opercularis and/or triangularis of the left inferior frontal gyrus can be achieved without causing any permanent language impairment" (Duffau, 2018, p.77). On the other hand, it is now widely acknowledged that motor/premotor and auditory cortices interact during language processing.

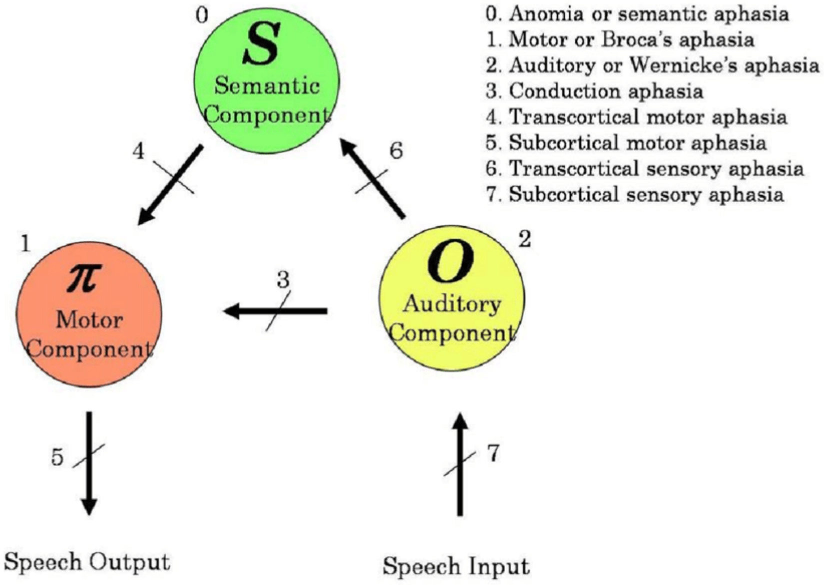
0
S
Semantic Component
1
π
Motor Component
2
O
Auditory Component
3
4
5
6
7
Speech Output
Speech Input
0. Anomia or semantic aphasia
1. Motor or Broca's aphasia
2. Auditory or Wernicke's aphasia
3. Conduction aphasia
4. Transcortical motor aphasia
5. Subcortical motor aphasia
6. Transcortical sensory aphasia
7. Subcortical sensory aphasia

Figure 1.1: Lichtheim's model of language representation in the brain (extracted from Oller, 2010).

The motor theory of speech perception developed by Liberman et al. (1967) and revised by Liberman and Mattingly (1985) sharply contrast with traditional models by proposing that listeners perceive the phonetic features of speech sounds by tracking the intended articulatory gestures of the speakers. In that sense, they were the first to hypothesized intimate links between speech perception and production. Although the authors did not discuss any underlying neural organization, they claimed this process of "translation" from articulatory to phonetic to be achieve by a language-specific module that rather than being acquired via associative learning is innately specified. This theory was vividly criticized for three main reasons: its modular approach, its domain specificity and its innate nature. The existence of such properties suffered intense discussion and the theory was mostly rejected (Massaro & Palmer, 1998; Lane, 1965) because of its incompatibility with contemporary acknowledged properties underlying broader cognitive and brain architecture.

Accounting for the advances in neurobiology of sensory systems, Hickok & Poeppel (2004) proposed a model in which speech perception initiates in the bilateral STG and then dissociate in two main neural pathways. Projecting from the superior temporal auditory cortex into the ventrolateral prefrontal cortex (VLPFC), the ventral stream is postulated to be involved in mapping speech sounds representation onto meaning (analogously to the O-to-S link depicted in Fig.1). The dorsal stream on the other hand, projecting from posterior STG (pSTG) and reaching frontal regions is proposed to allow the association of speech sounds and articulatory representations of speech (the origin and scope of this model is further discussed in Chapter II). Although Hickok and Poeppel's model has provided valuable insights, namely respect to the missing auditory-motor link in former models (suggesting a possible O-to-π link in Fig.1), it still represents a modular approach of language with a sound to meaning network for speech recognition and a sound to action network for speech production. Also, the authors stated that "although the proposed dorsal stream represents a tight connection between processes involved in speech perception and speech production, it does not appear to be a critical component of the speech perception process under normal (ecologically natural) listening conditions, that is when speech input is mapped onto a conceptual representation" (Hickok and Poeppel, 2004, p. 73).

More recently and from a complementary approach, neuroscientists have extended their knowledge about Hebbian learning mechanisms and cell assemblies' characteristics in an intent to develop computational models of language. An influential computational model has been proposed by Friedmann Pulvermüller and collaborators (Pulvermüller & Fadiga, 2010, 2016, Garagnani & Pulvermüller, 2013) whose central claim is that brain language mechanisms are built on the action-perception neural circuits reused for communicative purpose (Pulvermüller, 2018). Importantly, this model defends the idea that motor system is necessarily involved during speech perception. The later assumption is consistent with the several contemporary theories. First, the discovery of populations of neuron in primate's brains that fire during both the observation and the execution of an action (Rizzolatti & Arbib, 1998; Rizzolatti & Craighero, 2004) supports the possibility that speech production circuits activate during speech perception. This is also in line with studies conducted in the framework of embodied theories of language which have demonstrated that semantic processing is associated to somatotopic sensorimotor activity, including during metaphoric and abstract language processing (Vigliocco, Perniss & Vinson, 2014). Finally, the intimate link between speech perception and production proposed by computational models requires anatomical and functional interactions between temporal-auditory and frontal-motor areas as well as subcortical areas. These interactions are assured by a complex set of fasciculi which have been intensively investigated in both humans and other primates and have inspired fascinating evolutionary theories of language (Aboitiz, 2018a).

Consistently with neurobiological (Skipper, Nusbaum & Small, 2005) and computational (Schomers & Pulvermüller, 2016) models of language, the field of psycholinguistic also reject the longstanding and misleading perception/production dichotomy introduced by the classical model. Pickering and Garrot (2013) proposed an integrated theory of language production and comprehension in which speakers and listeners use covert imitation and forward modeling to make predictions and monitor the upcoming linguistic input (i.e., at phonological, semantical and/or syntactical levels). This theory is illustrated in Figure 2 where individual B's production command feeds into B's production implementer resulting in B's utterance $p[sem,syn,phon]_B(t)$. Individual A covertly imitates B's utterance and uses his forward production model to predict B's forthcoming utterance at time (t+1). B simultaneously constructs the next production command and uses his forward production model to predict his forthcoming utterance $p[sem,syn,phon]_B(t+1)$.

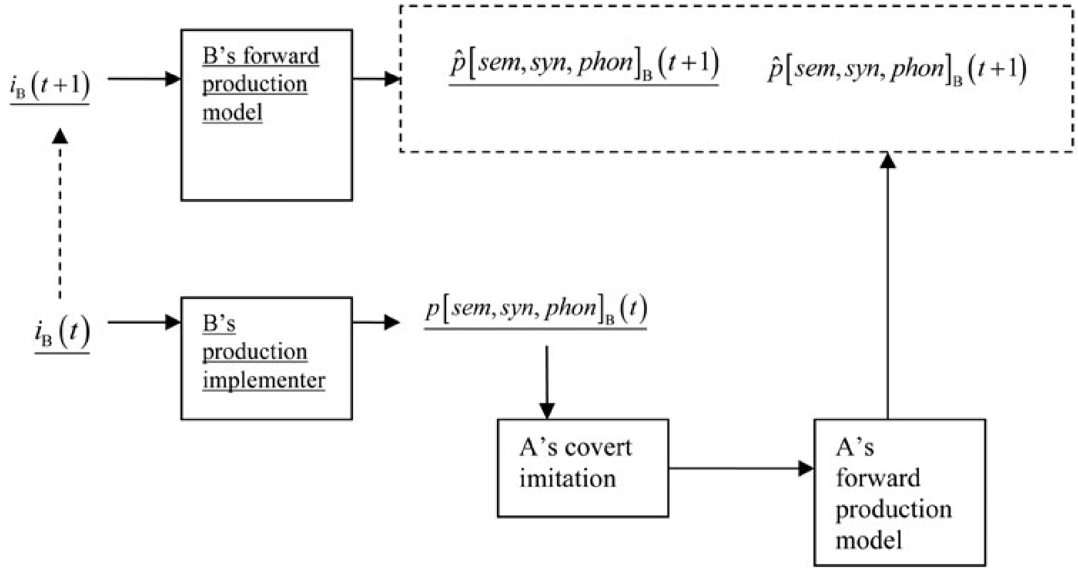
$i_B(t+1)$
B's forward production model
$\hat{p}[sem, syn, phon]_B(t+1)$
$\hat{p}[sem, syn, phon]_B(t+1)$
$i_B(t)$
B's production implementer
$p[sem, syn, phon]_B(t)$
A's covert imitation
A's forward production model

Figure 1.2: Integrated theory of language production and comprehension (extracted from Pickering & Garrot, 2013)

To summarize, far from a Fodorian conception of domain-specific and encapsulated modules for language (Fodor, 1975), it is now increasingly consensual that language involves a distributed anatomical architecture which includes cortical and subcortical structures and heavily relies on domain-general neural and cognitive mechanisms, such as imitative behaviors and prediction.

1.2. On the relevance of visual speech cues.

It is noteworthy, after the revision of the last century's most influential models of language (section 1.1.), that very little attention has been allocated to the processing of visual speech cues. All along the present dissertation, a special emphasis will be put on the importance of orofacial gestures for language development, speech processing and language evolution. As it will be exposed in forthcoming chapters, the visual access to speaker's mouth movements and the availability of orofacial effectors of the perceiver have an important role in several communicative circumstances.

1.2.1. Visual cues for speech acquisition in infancy.

During the first years of life, a critical period for language development, the principal mean of communication between the toddler and his parents are face to face interactions. When we watch and listen to someone speak, the brain associates the visual information of the movements of the speaker's mouth with the speech sounds that are produced by these movements (Saito et al., 2005; Cogan, 2016). As depicted in Figure 3, by the age of 6-months infants start to adopt visual strategies for face exploration focusing more on the mouth than on the eyes of speakers (Lewkowicz & Hansen-Tift, 2012). This reorientation of visual attention towards the lower part of the face occurs when infants enter in the canonical phase of babbling (i.e. repetition of syllables) where they explore their phonatory apparatus and finally achieve the production of their first words by the end of the first year of life (see Chapter II and III). From a developmental approach, these explorative strategies are thought to contribute to a multimodal mapping between vocal gestures, speech sounds and their visual counterparts (see Chapter II, section 2.2.4. for Trimodal repertoire: phoneme, viseme, *articuleme*).

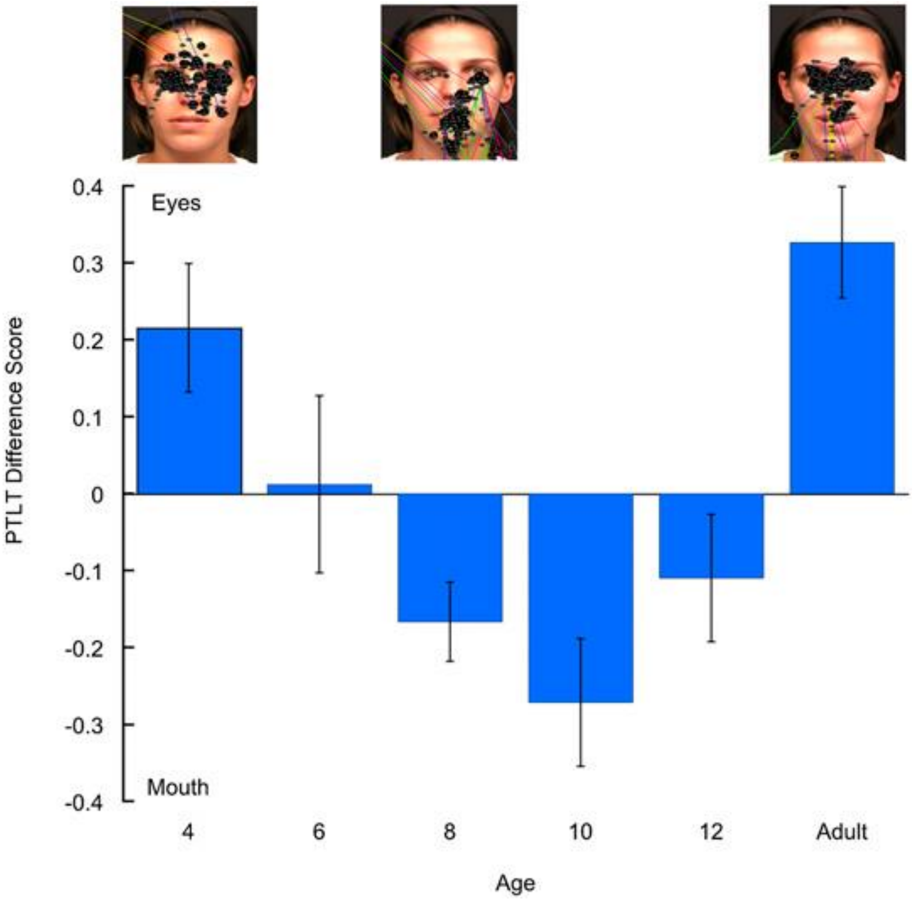

Figure 1.3: Proportion of total looking time (PTLT) between eyes and mouth regions in 4 to 12 months-old infants (extracted from Lewkowicz & Hansen-Tift, 2012).

1.2.2. In challenging hearing situations.

Another circumstance under which the access to visual speech cues reaches major relevance is when the acoustic channel is somehow disturbed. Perhaps the most illustrative example is the reorganization of sensorial cortices and the development of compensatory visual strategies for speech processing in individuals suffering deafness or hearing impairments (Dole, Méary & Pascalis, 2017; Mitchell, Letourneau & Maslin, 2013; also see Chapter III). Similarly, recent neurophysiological studies have demonstrated the importance of visual speech input for resolving perceptual ambiguity in the context of noisy environment. Since visual input precedes the associated auditory signal during face to face interactions, it has the potential to serve a predictive function in facilitating auditory processing of speech, “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; ten Oever et al., 2014). For example, in a crowded bar where the acoustic channel is overloaded by surrounding conversations, music and laughter, articulatory movements inform us about when the speaker is initiating a conversation. In addition to provide temporal information respect to the onset of speech, visual speech contains cues about the speaker’s articulators that are particularly informative when acoustic stream is degraded because the shape of the lips and/or the position of the tongue restrains the subsequent auditory input to a possible set of phonemes (e.g., bilabial /b/ vs. alveolar /d/ vs. velar /g/) (van Wassenhove et al., 2005). The use of visual speech cues has also been intensively studied in individuals with autism spectrum disorder (Williams et al., 2004; Irwin et al., 2011; Irwin & Brancazio, 2014; Irwin et al., 2015; Schelinski, Riedel & von Kriegstein, 2014).

1.2.3. For foreign language learning

Learning a foreign language in adulthood can be somewhat difficult. Because of a lifelong exposition to a native language, our ability to discriminate non-native phonemes diminish (i.e., a phenomenon called perceptual narrowing). Recent studies have shown that adults achieve phoneme discrimination in a foreign language when stimuli are presented audiovisually but fail to detect the differences when presented auditorily only (Navarra and Soto-Faraco, 2007; Hirata and Kelly, 2010). Interestingly, the pattern of visual attention orientation documented in infants (see section 1.2.1) seems to be adopted by adult foreign language learners. Barenholtz et al. (2016) reported that language familiarity modulates the relative attention to the eyes or mouth, with increasing attention to the mouth for unfamiliar languages. Moreover, the visual processing of known language elicited higher motor excitability than unknown language, “suggesting that motor resonance is enhanced specifically during observation of mouth movements that convey linguistic information (Swaminathan et al., 2013). These findings contribute to a central aspect of the present thesis, namely the interactions between visual and motor cortices during speech perception (i.e., the viseme to articuleme link). It is noteworthy that technological support for foreign language learning (e.g., apps or CDs) are generally based on oral and written language, informing the learners about the acoustic outcome of the new word production (i.e. what does it have to sound like?). But since learners rarely have access to the corresponding orofacial movements of speakers on such devices, they lack information about the motoric sequence needed to pronounce the word adequately (i.e. how do I make it sound like this?). The recently raising multimodal approach of speech processing and the contribution of the trimodal repertoire hypothesized along this dissertation might has the potential to guide technological development in a way that foster learning outcomes.

1.3. Epistemological and methodological assumptions.

In this introductive section, we have reviewed behavioral evidences of the importance of visual input for speech perception as well as neurobiological accounts of language proposing an interactive and multimodal network for audio-visual integration. First, orienting attention to visual speech information helps to construct sensori-motor models that match speech sounds with their respective articulatory movements. Those models seems to be crucial for emergence of speech production by the end of the first year of life, and later in life, for foreign language learning. Second, because visual speech cues are perceived before speech sounds, they play a key role in speech perception. Namely, visual input serves as a potential cue to predict the up-coming speech sound and consequently it can disambiguate auditory processing when acoustic input is deteriorated (e.g., in noisy environment or in persons with hearing impairment) or interfere with auditory processing when there is a mismatch between visual and auditory information (e.g., Mc Gurk effect). On the other hand, neuroimaging studies consistently report activity in the movement sensitive visual areas (V5/MT) and the pSTS, thought to be responsible for auditory-motor integration of speech (see section 2.4.2. and Hickok & Poeppel, 2007, for a review). In addition to visual and auditory cortices, there is increasingly evidences that the pars opercularis of Broca's area as well as motor and premotor cortices are also recruited during visual speech perception since the activity of regions is modulated by the saliency of visemic content. The interplay of these regions strongly suggests the existence of a highly interactive network that include not only typical brain areas involved in language perception but also brain areas involved in language production.

Here, we proposed a trimodal network that allows crossmodal transductions by which visual cues are converted into speech sounds on the basis of the articulatory motor plan necessary to produce them. This crossmodal transduction results, along with exposure to consistent linguistic input, in the development of a trimodal repertoire composed of phonemes, visemes and their motoric counterparts, henceforth, articulemes. Since part of the neural circuitry for specular activity and imitation are also involved during speech perception, we hypothesize that the mechanisms supporting this trimodal mapping for speech involve the mirror neuron system (see section 2.5.). To test this hypothesis, we recorded electrophysiological responses elicited by different orofacial movements displayed in silent videos: syllables, backward played syllables, non-linguistic movements and non-human movements. Importantly, three types of syllables were presented that differed with respect to their visual salience and their place of articulation (PoA: bilabial, alveolar and velar). In order to assess whether speech articulatory effectors are solicited during speech perception, participants were asked to observe the videos in two conditions: normal observation vs observation while they hold an effector depressor horizontally between their teeth in order to restrict articulators' motion. Event-Related Potentials (ERPs) and time-frequency dynamics were analyzed from the EEG signal.

In the following chapter (Chapter II) a theoretical proposal resulting from an integration of two complementary approaches is presented. The understanding of the evolutive constrains and the neuroanatomical circumstances in which speech appeared across phylogenetic history of human lineage is discussed alongside to the main idea of this dissertation, namely the existence of a trimodal network supporting the ontogenetic development of speech and lifelong language processing and learning. In Chapter III, the results of ERPs analysis are presented and discussed in the frame of crossmodal prediction theories. The results of time-frequency analysis are exposed in Chapter IV where the contribution of motor regions for speech perception, and the role of articulemes are discussed. Finally, Chapter V aims to integrate and contrast the thesis of a trimodal network for speech perception with the empirical outcomes of the current experimental manipulation. Conclusions are drawn, not only contributing to the theoretical framework of an enriched model for the neurobiology of language but also contributing to understanding of the origin and evolution of human language.

Chapter II:

Origin and Evolution of Human Speech:

Emergence from a trimodal auditory, visual and vocal network

This chapter is based on:

Michon, M., López, V., Aboitiz, F. (2019). Origin and evolution of human speech: emergence from a trimodal auditory, visual and vocal network. In Evolution of the human brain: from mater to mind. *Progress in Brain Research*, chapter 14, vol.251.

Abstract

In recent years, there have been important additions to the classical model of speech processing as originally depicted by the Broca-Wernicke model consisting of an anterior, productive region and a posterior, perceptive region, both connected via the arcuate fasciculus. The modern view implies a separation into a dorsal and a ventral pathway conveying different kinds of linguistic information, which parallels the organization of the visual system. Furthermore, this organization is highly conserved in evolution and can be seen as the neural scaffolding from which the speech networks originated. In this paper we emphasize that the speech networks are embedded in a multimodal system encompassing audio-vocal and visuo-vocal connections, which can be referred to an ancestral audio-visuo-motor pathway present in non-human primates. Likewise, we propose a trimodal repertoire for speech processing and acquisition involving auditory, visual and motor representations of the basic elements of speech: phoneme, observation of mouth movements, and articulatory processes. Finally, we discuss this proposal in the context of a scenario for early speech acquisition in infants and in human evolution.

Keywords: human speech, dorsal and ventral streams, vocalizations, orofacial movements, trimodal repertoire, motor system

2. Origin and evolution of human brain: Emergence from a trimodal auditory, visual and vocal network.

2.1.Introduction

Language is arguably the hallmark of humanity, marking a qualitative difference between our species and other primates. However, the process by which our direct ancestors acquired language remains one of the most challenging problems for modern evolutionary theory. While traditional linguistics has been resistant to the notion that a system as complex as human language could have emerged through the process of biological evolution, more modern views based on comparative psychology and neuroscience have pointed to the existence, in non-human primates (NHPs), of auditory-vocal neural networks similar but simpler to those involved in human language processing (Aboitiz, 2018). The organization of these networks follows a general pattern that is conserved in different sensory modalities, and in which there is space for cross-modal interactions where one network establishes associations with neighboring networks. This evidence fits the emerging notion that human communication is a multimodal process, where different sensory modalities contribute to transmit and perceive meaningful signals.

In this article, we will address the problem of speech origins in human evolution, discussing recent evidence from comparative neuroanatomy and developmental psychology. We claim that this evidence supports the concept that speech originated as a result of the co-evolution of several sensory and motor domains that conflated in the generation of the auditory-visual-vocal network that participates during modern speech processing in infants and adults. Our focus will be on speech instead of language because speech is a behavior more dependent on well-defined sensorimotor systems and therefore more tractable in evolutionary analyses. Likewise, we will not address other forms of communication like hand gestures and pantomimes, which may have been important channels for communication in our early ancestors. There is now a fashionable hypothesis that language evolved first by gestures and speech is a late development, to which we will make some brief reference in the article.

More precisely, we will first review the evidence of the parallel organization of the visual and auditory systems into dorsal and ventral components, each processing different aspects of the perceived stimulus. We will also point out to the high conservation of these pathways in humans and in non-humans primates, and discuss how the networks involved in speech processing could have emerged from this ancestral scaffolding. Particularly, in this paper we focus on the convergence of auditory and visual afferents into the homolog of anterior Broca's area in the NHP, proposing this audio-visual region as a node for the evolution of speech in our lineage. An important point is that we distinguish a trimodal repertoire that is relevant for speech acquisition, that includes an auditory representation of sounds (phonemes), a visual representation of mouth movements (visemes), and a motoric representation of the articulation process during speech execution (articulemes). In this line, we discuss the role of visual processing in vocal imitation and speech acquisition in infants, where predictive coding mechanisms and mirror neurons may participate both in the visual and auditory modalities.

2.2. Parallelism between the neuroanatomy of auditory and visual systems.

2.2.1. The dual pathway organization of the visual system.

In the early nineties, based on previous neuropsychological, electrophysiological and behavioral evidence, Goodale and Milner (1992) proposed one of the most influential models for understanding the visual pathways for perception and action in the human brain. This framework postulates an anatomical and functional segregation of visual processing in two independent streams (Goodale & Milner, 1992). The ventral stream, anatomically running along the inferior temporal lobe, is also known as the 'what' pathway, and is widely thought to support the processing of visual information about the identity of objects. The dorsal stream, well known as the 'where' pathway, runs along the parietal lobe reaching the prefrontal cortex and is hypothesized to underlie the processing of visual information about objects' spatial location and contribute to execute movements under visual control. For this reason, some authors have preferred to term this as the 'how' pathway (Arbib, 2012). Well before Goodale and Milner proposed their model, experiments with split-brain monkeys already led to the idea of a functional segregation of visual processing and suggested that anatomically distinct brain mechanisms are involved of the vision of space and vision of object identity (Trevarthen, 1968).

2.2.2. Ventral and dorsal streams for speech.

Interestingly, the auditory system has revealed a similar dual pathway organization with a ventral stream projecting from the superior temporal lobe (auditory cortex) into the ventrolateral prefrontal cortex (corresponding to Broca's region in the human), which is crucial for auditory object identification and recognition, and a dorsal pathway from the posterior auditory cortex, that runs along the inferior parietal and frontal lobes and is involved in spatio-temporal processing of sounds and auditory-motor transformations (Kaas & Hackett, 1999; Romanski, 2007; Romanski et al., 1999).

In a series of articles summarized in Petrides (2014), Petrides and collaborators described the neuroanatomy of the ventral and dorsal pathways where auditory speech input coming from the ventral pathway is received in the anterior ventrolateral prefrontal cortex (aVLPFC, area 45 in Fig. 2.1). On the other hand, the dorsal stream maps auditory input from auditory area Tpt from the posterior superior temporal lobe into vocal motor representations in areas 44 and 45 of the VLPFC, via the arcuate fasciculus and the superior longitudinal fasciculus (Fig. 2.1). As most cortico-cortical projections, both streams carry not only bottom-up (from sensory to motor regions) but also top-down (from motor to sensory areas) information. Accordingly, the dorsal pathway has been suggested to provide backward efference copies of ongoing motor activity from prefrontal and motor cortices towards auditory areas via the dorsal pathway (Rauschecker, 2012).

Progress in our understanding of the functional anatomy of the human visual and auditory systems, alongside with the increasing use of new brain imaging techniques such as diffusion tensor imaging (DTI), have inspired new theoretical approaches for the neurobiology of language. These studies confirmed that auditory and visual projections are similarly configured in humans and in non-human primates, both subdivided in dorsal and ventral streams (Petrides, 2014, see Fig. 2.1). In line with these findings, Hickok and Poeppel (2004) proposed that from early processing stages, speech perception initiates in the bilateral superior temporal gyrus and then dissociates in two main processing streams. The ventral stream is postulated to be involved in mapping sound-based representations of speech onto meaning, while the dorsal stream projects from the posterior superior temporal lobe and reaches frontal regions, allowing the association of speech sounds and articulatory representations of speech (Hickok & Poeppel, 2004). Of special interest in this context, an area called the Sylvian-parieto-temporal region (Spt) has been proposed as an interface between auditory and vocal representations in the dorsal pathway (Hickok & Poeppel, 2007; Rauschecker & Scott, 2009). This region probably overlaps with cytoarchitectonic area Tpt (see Fig. 2.1). The authors remark, however, that the characterization of the dorsal stream as the "where" area may not be sufficient to reflect its critical role in visuo-motor and auditory-motor integration when visually or auditory guided orienting actions are required. Importantly, they argue that those sensorimotor interactions are domain general, serving both linguistic and non-linguistic processes.

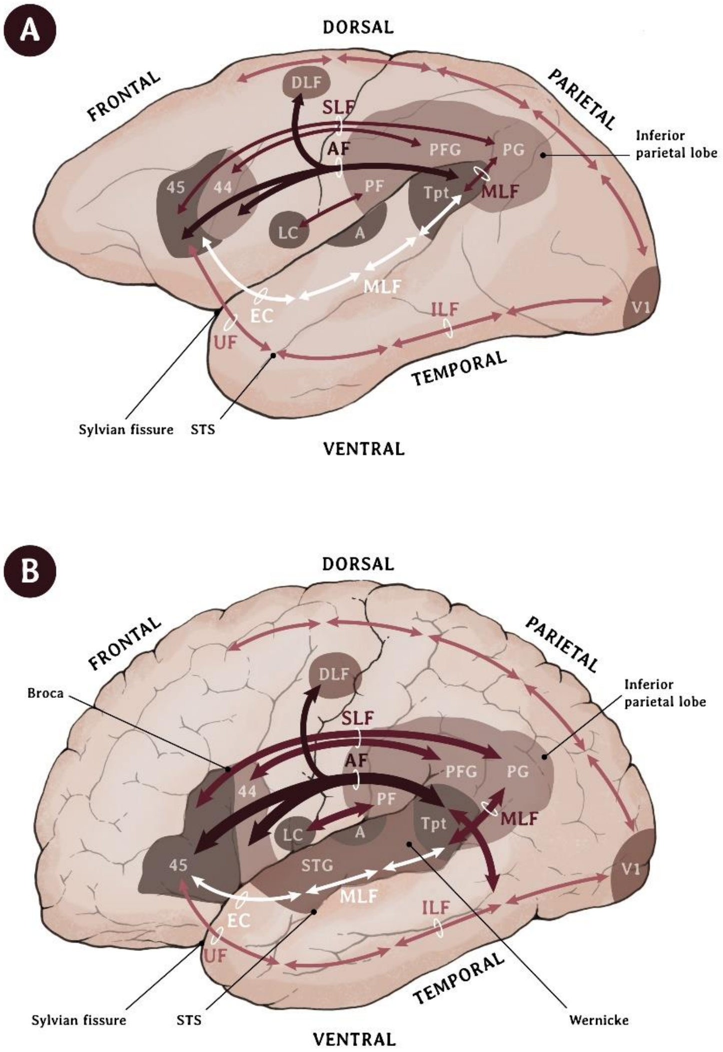

Figure 2.1: Auditory-vocal cortical connectivity in non-human primates (A) and humans (B). White arrows depict the ventral pathway and solid dark arrows represent the dorsal pathway, running via the arcuate fasciculus (AF) and the superior longitudinal fasciculus (SLF). For reference, the dorsal and ventral visual pathways are schematized as a series of arrows emerging from V1 (primary visual area). A, primary auditory area; AF, arcuate fasciculus; DLF, dorsolateral frontal cortex; EC, extreme capsule; ILF, inferior longitudinal fasciculus; LC, laryngeal and orofacial cortex; MLF, medial longitudinal; PF, PFG, PG, inferior parietal areas; SLF, ventral superior longitudinal fasciculus; STG, superior temporal gyrus; Tpt, cytoarchitectonic area Tpt; UF, uncinate fasciculus; V1, primary visual area.

2.2.3. The dual pathway model revisited: What, where and when?

Although a highly influential model, the anatomico-functional segregation of these two sensorimotor pathways has been reconsidered in recent studies. Accumulating evidence from both human and non-human primate (NHP) studies show that the “where” dorsal pathway is also implicated in “what” aspects of visual processing and has a functional role in object recognition, independently of object-directed action planning or execution. Similarly, the ventral stream is not solely involved in object vision but also contributes to the processing of several other visual features (Freud et al., 2016). Recent studies about the retinotopic organization of the visual system in primates and humans showing that visual space is sampled differently in the ventral and dorsal pathways have inspired a new dichotomy consisting of high versus low visual resolution processing (Arcaro & Livingstone, 2017; Janssens et al., 2014; Silson et al., 2018). Sheth and Young (2016) propose that the receptive fields of the ventral stream cells are focal compared with those of the dorsal stream, generally including the fovea and parafovea, consequently providing more detailed information about objects features for their categorization. Cells of the dorsal stream, in contrast, have larger receptive fields providing information about the presence and the location of salient object in peripheral vision and thus, contributing mostly in visuospatial processing (Sheth & Young, 2016).

In the auditory system, evidence coming from very different fields robustly suggests that the dorsal stream has higher temporal resolution compared with the ventral stream (Rauschecker, 2018). The importance of the auditory dorsal stream for the processing of temporal characteristics of sounds becomes particularly relevant for speech perception and production. Speech is a continuous stream of auditory information in which the meaning of sentences can be extracted because word sequences follow syntactic rules. This sequence analysis of auditory input is accomplished largely but not exclusively by the dorsal pathway (Rauschecker, 2018; Wilson et al., 2011), connecting preferentially area 44 with the superior and medial temporal gyri (STG and MTG) (Aboitiz, 2018a;

Friederici, 2016) (see Fig. 2.1). Lexico-semantic processing, on the other hand, preferentially recruits the ventral stream, including anterior Broca’s area (area 45) and anterior temporal lobe (Friederici, 2011). Nonetheless, studies with neurological patients has shown that structural integrity of several ventral pathway areas is also required for correct syntactic processing, especially during meaningful discourse processing (Brennan et al., 2012; Tyler et al., 2011). Using artificial grammars as stimuli, it has been found that processing of simple, linear grammatical forms relies preferentially on the ventral pathway and its connections to the ventral frontal cortex. On the other hand, complex syntactic forms involving recursion activate both the operculum and Broca’s area and its connection to the posterior superior temporal gyrus via the dorsal pathway (Friederici et al., 2006). However, the subtracting technique used to analyze these findings tends to underestimate the participation of widespread networks that may be relevant for natural sentence comprehension. Furthermore, processing grammatical meaningful sentences may require interactions between the dorsal pathway and the ventral pathway, and their interactions with the multimodal neural systems in the superior temporal sulcus, inferior temporal lobe, and temporoparietal and frontal areas supporting semantic representations depicting movement, actions and events (Beauchamp, 2015; Binder et al., 2009; Binder & Desai, 2011). In this line, the dorsal pathway may contribute to manipulate and process structural rearrangements of the sentences, while the ventral pathway and additional regions confer meaningful elements to grammatical processing like the distinction between subject and object, some aspects of verb processing, and more generally the lexical component (Aboitiz, 2017; Langacker, 2013). Thus, the close interaction between the dorsal and the ventral auditory and visual processing pathways, a large part of which takes place in Broca’s area and its vicinities, may be critical for correct grammatical and syntactic processing.

2.3. The ventrolateral prefrontal cortex: a convergence area for multimodal integration.

2.3.1. The dorsal stream for auditory-articulatory transduction.

As mentioned, both the dorsal and the ventral auditory pathways have been shown to have fiber projections into the VLPFC (see Fig. 2.2). The dorsal stream targets preferentially the posterior portion of Broca's area (area 44) which receives phonological information from the posterior auditory cortex and the anterior inferior parietal lobe, as well as sensorimotor inputs from premotor and motor cortices. This network has been proposed to be particularly important for the transduction of phonemes into motor-articulatory gestures (Papoutsi et al., 2009). Recent studies using intraoperative recordings challenge the traditional role of Broca's area as the "seat of speech area" and postulate that area 44 rather works as a functional gate that authorizes the phonetic translation preceding the articulation of speech, thus acting at a pre-articulatory stage (Duffau, 2018; Ferpozzi et al., 2018). During phonetic encoding, the syllables to be produced are organized into temporal motoric sequences, which are later executed by pre-motor and motor areas that directly control phono-articulatory musculature (Long et al., 2016). These predictions are supported by a study of Flinker and collaborators (2015) that investigated the timing in the recruitment of Broca's area respect to other language regions during cued speech production using electrocorticography (ECoG). The fine temporal and spatial resolution of ECoG allowed them to detect activity in Broca's area as soon as 240 ms after cue onset, but this activity surprinsingly shuts down when the production of words commenced, strongly suggesting that Broca's area is recruited for pre-articulatory phonetic processing rather than for the on-line coordination of speech articulators (Flinker et al., 2015).

2.3.2. Overlap of auditory and visual ventral streams for faces-voices associations.

The ventral auditory stream, on the other hand, projects into the anterior portion of Broca's area (area 45) and neighboring regions where afferences from the superior temporal sulcus

(STS) and anterior/inferior temporal regions are also received, providing multimodal sensory and lexico-semantic information. This network is thought to have a key role in the transformation of speech sounds to meaning (Saur et al., 2008). Noteworthy, the projections of the ventral visual pathway and those of the ventral auditory pathway overlap in areas 45 and 47, making these areas especially suited to link visual aspects of the speaker's face and mouth with auditory aspects of her voice (Romanski, 2007). In rhesus monkeys, Romanski and collaborators reported the existence of a neuronal population located in the VLPFC that is responsive to both visual processing of conspecific faces and to auditory processing of vocalizations (Diehl & Romanski, 2012; Romanski, 2012; Sugihara et al., 2006), strongly suggesting that the ventral auditory and visual streams converge in this region and integrate orofacial gestures and vocalizations. Interestingly, a few years later Hage and Nieder (2015) reported that the discharge rate of single neurons located in the VLPFC was modulated both when monkeys produced vocalizations or passively listened to other monkey calls. These results indicate that, in addition to the integration of audio-visual (i.e., voice-face) inputs, the VLPFC is well positioned for the integration of audio-motoric input (i.e. voice-articulation), turning this region into a critical region for the evolution of a complex audio-visual-motoric integration that may have contributed to the emergence of human speech (Hage & Nieder, 2015).

2.3.3. Homologies and differences between human and non-human primates.

2.4. Trimodal repertoire: phoneme, viseme, *articuleme*.

2.4.1. A visual counterpart of vocal articulations.

An aspect that has long been understudied is the visual processing of speech cues. Historically, the question of human language perception has been prevalently focused on auditory processing of speech. Recent investigations, however, postulate that language perception rather relies on an interactive multi-modal system, including not only auditory but also visual (Bernstein & Liebenthal, 2014) and motor systems (Glenberg & Gallese, 2012; Pulvermüller & Fadiga, 2010). The ability to associate motor articulation of speech and its auditory counterpart is important because it provides a model for the target outcome of the vocalization (e.g. What does it have to sound like?), but associating the motor sequence of an articulation with its visual counterpart is also relevant because it informs about the orofacial movements needed to produce the vocalization (e.g. How to make it sound like this?).

An illustration of the importance of the visual counterpart of articulations can be found in speech ontogenesis. As early as 4 months of age, infants have been shown to detect a switch from native to non-native language (and vice-versa) in silently displayed videos, suggesting that visual input alone is sufficient for native language discrimination at these early ages (Sebastián-Gallés et al, 2012; Weikum et al., 2007). At about 6 months of age, when infants begin to produce repeated syllables, a developmental stage called canonical babbling, they begin to shift the orientation of their gaze from the eyes to the mouth of their interlocutors. This reorientation of visual attention to the lower part of the face has been interpreted as a strategy used by prelingual infants to collect complementary audiovisual information about speech cues (Lewkowicz & Hansen-Tift, 2012; Tenenbaum et al., 2013). At 12 months, when infants start to produce their first words, their gaze shifts back to the talker’s eyes, possibly looking for other kind of socio-communicative intentions like adults do. Noteworthy, when audiovisual speech is desynchronized, losing its informative status, 10 months old infants no longer show the typical gaze pattern of preference for the mouth (Hillairet de Boisferon et al., 2016). Importantly, it has been demonstrated that orienting attention to the talker’s mouth is more strongly associated with expressive language skills than chronological age between 6 and 12 months in both monolingual and bilingual infants (Tsang et al., 2018). Altogether, this evidence ascribes a more important role than previously thought to visual speech processing in infancy. That is, visual information alone is sufficient for native tongue discrimination and the orientation of visual attention to articulatory movements of the mouth very likely help to construct a sensory-motor model for the emerging speech production by the end of the first year of life. Noteworthy, although the visual system provides an additional source of information supporting language acquisition, the access to visual speech cues is not strictly necessary for language development. For instance, after profound reorganization of visual areas for auditory functions, syntax and spoken language develop within normal range in congenitally blind subjects (Bedny et al., 2015; Lane et al., 2015; Perez-Pereira, 2006, Röder et al., 2002; Watkins et al., 2013).

The relevance of visual speech processing for language perception is not limited to infancy. Nowadays, the literature consensually supports that vision provides a complementary source of information that improves the perception of speech in adulthood as well. Having access to the visual speech cues afforded by the interlocutor's face can be especially advantageous in a noisy environment (Ross et al., 2007; Sumby & Pollack, 1954) and when hearing acuity is impaired (Auer & Bernstein, 2007; Bernstein et al., 2000). Pimperton, Ralph-Lewis and MacSweeney (2017) have recently corroborated that patients with early deafness adopt an adaptive strategy known as perceptual compensation (i.e., tendency to rely more on other modalities when one modality is impaired). They also reported a positive correlation between lip reading abilities and the age of cochlear implantation; the older you're implanted the better you lip read. Eye-tracking studies revealed that early deaf adults proportionally orient their visual attention more toward the mouth than the eyes whereas hearing adults typically and almost exclusively look to the top half of a face (Dole et al., 2017; Mitchell et al., 2013). This evidence indicates that early audition deprivation leads to a prolonged dependence on vision during speech perception in deaf people, which in turn results in a reorientation of visual attention in order to improve the perception of visual speech cues provided by orofacial movements. Finally, evidence from investigations in non-native speech perception show that language familiarity modulates the relative attention to the eyes and mouth, the attention to the mouth increasing in response to an unfamiliar language (Barenholtz et al., 2016). Moreover, adults who often fail to hear the difference between certain nonnative phonemic contrasts when presented only auditorily can successfully distinguish these contrasts when presented audiovisually (Hirata & Kelly, 2010; Navarra & Soto-Faraco, 2007). The latter suggests that, as infants do, adult learners of a foreign language look at the mouth taking advantage of orofacial movements in order to facilitate phoneme discrimination.

Intriguingly, it has been widely documented that an incongruence between auditory and visual speech information can lead to a decreased auditory perception. A convincing illustration of this phenomena is the so-called Mc Gurk effect, where the syllable |ba| was auditorily presented in a videoclip where a face simultaneously articulated the syllable |ga|. As a result, the auditory processing is biased by the visual input and subjects consistently reported to hear the syllable |ga| or |da| (Mc Gurk & Mac Donald, 1976). Because orofacial speech movements (i.e., especially lips) are typically perceived before the auditory signal, visual information precedes auditory information. Paris, Kim & Davis (2013), reported that when speech is presented audiovisually, the prior access to visual speech form speeds up the processing of auditory speech compared to when speech is presented only auditorily. They argued that the temporal priority of visual speech may serve as a potential cue to predict aspects of up-coming auditory signal. The more the articulatory movements were salient and predictive of a possible speech sound, the speediest was the processing of auditory signal. 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 et al., 2005).

2.4.2. Neuronal correlates of visual speech perception.

2.4.2.1. The involvement of visual and auditory cortices.

The use of functional magnetic resonance imaging (fMRI) and magnetoencephalographic (MEG) recordings has refined the localization of brain regions involved in visual processing of articulatory movements. Calvert and collaborators (1997) reported that silent lip-reading activated areas of the temporal auditory cortex that considerably overlap with those activated by auditory speech perception. Furthermore, the activity in the angular gyrus in the inferior parietal lobe shows significant differences in response to a static face compared to a face silently articulating words. This area is known to be involved in the mapping of a visual form (including words and numbers) to its linguistic representation. The activation of angular gyrus in response to orofacial movements suggests that this region may also be involved in mapping visual speech cues to their corresponding verbal representations. The same group published a fMRI study where the pattern of cortical activity between still images of a talking face and video-clips of a face naturally speaking were compared. The neural circuit activated was highly similar for both stimuli with an involvement of traditional language regions of the left hemisphere, including the auditory cortex (areas 41/42, lateral Heschl's gyrus), the STS and the VLPFC or areas 44/45. Different responses were observed however in the visual cortex, where still faces predominantly activated the primary visual cortex (V1/V2) whereas moving faces elicited more activity in the visual movement areas (V5/MT) in the occipito-temporal lobe (Calvert & Campbell, 2003).

More recent evidence has consistently reported that the posterior part of the left STS is involved in visual speech recognition and appears to be activated to a greater extent when auditory input mismatches visual speech input. For instance, Blank and von Kriegsten (2013) demonstrated a greater functional connectivity between left pSTS and auditory-speech areas when a visual cue mismatches an upcoming auditory cue. The pSTS is thought to have a crucial role in predicting up-coming auditory speech based on visual information that typically precedes the acoustic signal in a natural face to face conversation, this predictive role being even more crucial in a noisy environment or in hearing-impaired persons. In line with this statement, a MEG study investigating the neural activity elicited by viewing mouth producing non-linguistic movements reported a clear activation of occipito-temporal areas (V5/MT) but no activation of the STS (Miki et al., 2004). This result suggests that the involvement of the STS is specific to mouth movement associated to speech production (but see Puce et al., 1998). Moreover, when the pSTS is stimulated by transcranial direct current the behavioral performance of both visual-only and auditory-only speech recognition is altered compared to a control group who receive no current stimulation (Riedel et al., 2015). Using intracortical ERPs in temporal cortices, a study of Besle and collaborators (2008) provided critical insights into the cortical dynamics of visual speech processing both in its spatial and temporal dimension. They first replicated the finding that lip-reading activates the temporo-occipital junction and the posterior middle temporal gyrus, areas which correspond to the movement-sensitive visual cortex (V5/MT) as well as the secondary auditory cortex. Importantly, the activation of these visual areas occurred around 140 ms after stimulus onset and within the 10 subsequent milliseconds superior temporal regions were activated as well, suggesting that V5/MT directly feedforwards visual information to the secondary auditory cortex (Besle et al., 2008). Altogether, the evidence indicates that the neural basis of visual speech processing includes the temporo-occipital junction (V5/MT area) involved in movement processing as well as the posterior part of the auditory cortex (pSTS) with a possible role of angular gyrus in mapping the visual form of the mouth to its corresponding phoneme.

2.4.2.2. The involvement of motor cortices.

Although movement-sensitive visual areas and auditory cortex have been consistently reported to be involved in visual speech processing, it seems that other areas are also recruited. A considerable number of studies demonstrate the recruitment of motor cortices during speech processing. Dubois et al. (2012) investigated visemic processing that is the visual counterpart of phonemic processing based on *visemes*, the distinctive visual units of speech. In this study, participants were asked to discriminate syllables and non-phonological stimuli (i.e., the same syllables played backward) presented audiovisually either with images of a speaking facial configuration (i.e., still condition) or with videos of dynamic mouth movements associated to the articulation of speech (i.e., dynamic condition). They reported an increase in discrimination performance in the dynamic compared to the still condition that was associated with hemodynamic activity in the bilateral occipito-temporal visual areas V5/MT and the left premotor cortex. Interestingly, the former was responsive to facial movements independently of its linguistic content (phonological = non-phonological) whereas the activation of the latter was specific to the phonological discrimination. However, because stimuli were presented in audiovisual modality, it is difficult to determine if the activation of premotor cortex is attributable to the sensori-motor representations associated with the phonological form of the syllables or to the visemic processing per se (Dubois et al., 2012).

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 show that audiovisual speech perception activated not only regions that are typically associated with language perception but also regions associated with language production. First, the activity of posterior superior temporal gyrus and sulcus, already known to be a hub of multimodal integration for speech perception, was modulated by the saliency of articulatory movements, becoming more active as visemic content increases. Second, the activation of Broca's area and particularly of pars opercularis (i.e. area 44) was greater in the visual-only condition compared to the audiovisual condition. Based on their shared functional properties and connectivity, the authors suggest that pSTS and area 44 work together to associate the sensory patterns of phonemes and/or visemes with the motor commands needed to produce them. Finally, similarly to the pSTS, the activity of the dorsal precentral gyrus and sulcus (i.e., premotor and motor cortices) was 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) based on the sensori-motor representation generated by the pSTS and pars opercularis (Skipper et al., 2005). Even though Broca's area, premotor and motor cortices have traditionally been associated with language production, it seems that they also are part of a highly interactive network that "translates" visual information of mouth movements into phonetic information based on the motor commands required to generate those movements. As it will be exposed in the next section, the existence of such a network implies that, over a lifetime of speech production, humans develop a kind of multimodal repertoire associated with the speech forms of their language(s).

2.4.3. Trimodality, the missing link?

The evidence reviewed above (see sections 2.4.2.1. and 2.4.2.2.) emphasizes the importance of the visual counterpart of speech sounds and the involvement of motor cortices during visual speech perception, leading the scientific community of the field to reconsider the weight accorded to visual and motor processes in the models of speech perception. Neural circuits for language have long been separated into perceptive and productive components. In this section, the multi-modal nature of speech will be addressed in an attempt to bridge the gap between perception and production. An insightful beginning would be to define to the fundamental units of speech in the different modalities involved in language processing. The smallest meaningful unit of sound that can distinguish two words from each other ("light" vs "right") is known as a phoneme. Phonemic contrasts have been extensively investigated by linguists and psycholinguists both in childhood and adulthood. In contrast, the visual equivalents of phonemes, known as visemes, have received relatively little attention. In fact, visemes are not precisely defined. No one-to-one correspondence exists between a phoneme and a viseme, instead an identical configuration of the lips can be mapped with several phonemes (Bear & Harvey, 2018). Last but not least, the third fundamental unit of speech is the motor sequence of articulatory movements required to produce a given phoneme. We will call this motor unit the *articuleme* in the rest of this chapter. Accounting for the above-mentioned neuroanatomic evidence from both monkeys and humans, we propose the existence a trimodal repertoire where the auditory (i.e., phoneme), the visual (i.e., viseme) and the motor (i.e., *articuleme*) counterparts of speech forms are interactively linked.

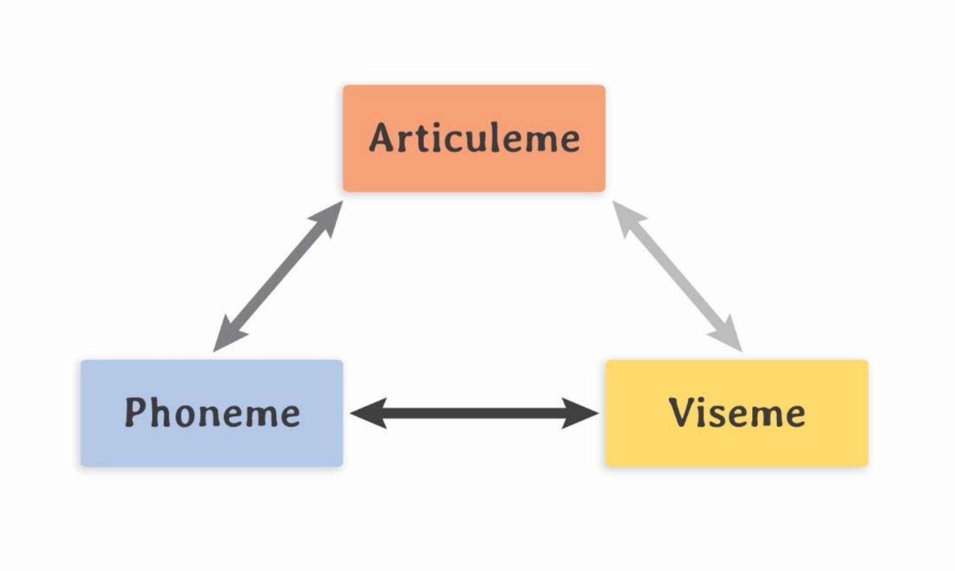

Figure 2.2: Trimodal repertoire for speech.

The overlap between the projections of ventral auditory and visual pathways in VLPFC in primates and the responsiveness of neural populations within the same area to both visual and auditory input provide a first robust argument in favor of the integration of orofacial gestures and vocalizations (Romanski, 2007; see black link in Figure 2.2). Moreover, and despite the lack of one-to-one mapping, the involvement of auditory cortex in lip-reading tasks strongly suggests that phonemes and visemes are somehow associated. Namely, visual input serves as a potential cue to predict the up-coming speech sound and consequently it can disambiguate auditory processing when acoustic processing is difficulted (e.g. in noisy environment, for foreign language learners or for persons with hearing impairment) or interfere with auditory processing when visual and auditory information mismatch (e.g. Mc Gurk effect). Respect to the link between articulemes and phonemes (see dark grey link in Figure 2.2), the pars opercularis (area 44) is preferentially positioned since it receives phonological information from the posterior auditory cortex and the anterior inferior parietal lobe, as well as sensorimotor inputs from premotor and motor cortices. This network in the dorsal auditory stream has been proposed to be particularly important for the transduction of phonemes into motor-articulatory gestures (Papoutsi et al., 2009). Moreover, a recent ECoG study reports that the inferior frontal gyrus represents both gestures and phonemes, while ventral precentral gyrus represents gestures to a greater extent than phonemes (Mugler et al., 2018). On the other hand, neuroimaging studies consistently advocate for the pSTS to be responsible of auditory-motor integration of speech (see Hickok & Poeppel, 2007, for a review). In addition to visual and auditory cortices, there is increasing evidence that the pars opercularis of Broca's area (area 44) as well as motor and premotor cortices are also recruited during visual speech perception since the activity of these regions is modulated by the saliency of visemic content. Finally, and may be the less documented link within the proposed trimodal repertoire, is the *articuleme*-viseme link (see light grey link in Figure 2.2). A recent study proposes a visual cortical network in the dorsal stream for viseme-phoneme mapping where motor-related areas exert top-down control on visual cortex (Hauswald et al., 2018). This is in line with developmental studies documenting that strategical orientation of the attention to the mouth during the first years of life is associated with greater expressive language skills (Tsang et al., 2018). Orienting attention to visual speech information may help to construct a highly interactive network that "translates" visual information of mouth movements into phonetic information based on the motor commands required to generate those sounds. These interactions between viseme, phoneme and *articuleme* must have been as relevant for the appearance of a proto-lexicon in early humans as it is for the emergence of speech production by the end of the first year of life.

2.5. Mirror neuron system, predictive coding and imitative behaviors

This section aims to discuss the possible general domain mechanisms involved in the construction of the above hypothesized trimodal repertoire for speech.

2.5.1. The Mirror Neuron System

In order to better understand the nature of the sensori-motor representations of speech, the literature on the human mirror neuron system (MNS) becomes relevant. Mirror neurons are neurons that fire both when an individual executes an action and when she observes the same action being performed by another individual. Mirror neurons were discovered in the premotor area F5 of monkeys while they were performing or observing a hand grasping action (Rizzolatti et al., 1996). Comparative neuroanatomical studies have demonstrated that area F5 in macaques, where mirror neurons are located, is homologous to the ventral premotor area 6v in humans (Aboitiz, 2012; Petrides et al., 2005). The MNS has been considered by some as a link between interacting individuals that "mirror" the action performed by the other person in one's own system allowing a common understanding of the willed communicative act (Rizzolatti & Arbib, 1998). This motor resonance, also known as specular activity of other's action, is thought to be the mechanism that underlies imitative behaviors. Vocal imitation and mimicry have been proposed to play a crucial role both in the ontogenetic development (Messum & Howard, 2015; Nguyen & Delvaux, 2015) and in the phylogenetic evolution (García et al., 2014) of language. As postulated by the motor theory of language perception developed by Lieberman in the1980s: "In all communication, sender and receiver must be bound by a common understanding about what counts; what counts for the sender must count for the receiver, else communication does not occur. Moreover, the processes of production and perception must somehow be linked" (Lieberman, 1993, see also Lieberman, 2015). However, the link between the observed and the executed may not be straightforward: cetaceans and parrots can imitate the human voice with sound-producing systems very different than the human vocal system, which indicates that the motor programs of the imitator and the imitated do not need to be similar (Aboitiz, 2018a; Abramson et al., 2018).

The human MNS involves an anterior area encompassing the posterior inferior frontal gyrus (IFG), the adjacent ventral premotor cortex (vPMC) and a posterior area including the rostral part of the inferior parietal lobe (IPL) that corresponds to area PF in the macaque. In this line, Iacoboni and Dapretto (2006) propose a neural circuitry for imitation of manual gestures and actions that in addition to the MNS involves the pSTS.

They postulate that visual input from pSTS is sent to the IPL where a motoric description of the action is elaborated and forwarded to the frontal areas of MNS which is more dedicated to the goal of the action. Neurons of the STS that respond selectively to biological motion both in humans and monkeys are increasingly accepted to be part of the MNS (Keysers & Perrett 2004). Similar to the human MNS network proposed by Iacoboni and Dapretto (2006), premotor area F5 in the macaque brain is reciprocally connected to parietal area PF, that is reciprocally connected to STS providing sensory information to the MNS (see Fig. 2.3).

Proponents of the MNS hypothesis of language origins claim that language may have emerged from a MNS originally involved in hand action imitation which was later refined for speech mimicry (Iacoboni, 2005; Iacoboni et al., 2001). They argue that the propensity for gestural mimetic capacity has been a fundamental aspect in language evolution, emphasizing the role of imitation of manual gestures as precursors of vocal imitation and the subsequent origin of speech (Arbib, 2012). Although we agree with a general role of imitation and with an involvement of mirror neuron activity in the origin of speech (see Fig. 2.3), we also consider the possibility that vocal imitation leading to early speech is an ancient character in the human lineage, and may have developed together, rather than deriving from, manual imitation. Vocal imitation is observed very early in human infancy (12 to 20 weeks) and can be observed in other highly social species that lack grasping abilities, indicating that the latter is not a requisite for the former (Abramson et al., 2018; Kuhl & Meltzoff, 1996). In this sense, a vocal mirror circuit may have developed in our ancestors and other vocal learning species, in parallel or independently to a grasping mirror neuron circuit (Aboitiz, 2017, 2018a).

2.5.2. Predictive coding and imitative behaviors.

Some years ago, Kilner et al. (2007) proposed a feedforward recognition model in which the observation of an action results in “the firing of neurons in the STS, which drives activity in area PF, which in turn drives activity in area F5”, suggesting that low-level visual and kinematic information are transformed in high-level representations of intentions subjacent to the action performed. The authors suggest that MNS function can be understood within a predictive coding framework appealing to a statistical approach called empirical Bayesian statistics. Predictive coding relies on the bidirectional interactions between the frontal and motor areas mentioned above (see Figure 2.3), that exert strong top-down influences on sensory cortices. These projections serve to anticipate the perception of the executed action, minimizing prediction errors and refining the subsequent motor sequences (Aboitiz, 2017; Kilner et al., 2007; Rauschecker, 2012). In the domain of action perception, prior contextual knowledge makes it possible to construct a representation of the goals and intentions of the person performing the action and therefore predict her motor commands and kinematics on the basis of their resonance in the observer’s own motor system (Kilner et al., 2007). The comparison between the observed and the predicted kinematics permit to elaborate a prediction error that is used to update our representation of other’s actions.

Figure 2.3. depicts a proposed scheme for a predictive coding network anchored in the speech circuit rather than in the visuo-manual gestural circuit. The blue arrows are hypothesized to be re-afferent connections that send copies of motor commands back to the pSTS and sensory regions, that are matched with the sensory representation of the observed behavior which may be one’s own or others’ (i.e., “Am I actually doing what I have seen?”). In audio-visual speech interactions, there is no visual information about one’s own mouth movements, and the observation of other’s gestures during social interactions becomes particularly relevant (Ray & Heyes, 2011). During face-to-face interaction, humans spontaneously and unconsciously mimic a variety of behaviors, a phenomenon called automatic mimicry. These imitative behaviors emerge early in ontogenetic development and are thought to be crucial for predictive coding and error minimizing during speech perception (Cook et al., 2014; Pickering & Garrod 2013; Schomers & Pulvermüller 2016; Skeide & Friederici 2016; Skipper et al. 2017). In this context, we previously mentioned the findings of Kumar and collaborators (Kumar et al., 2016) showing a much more robust connectivity between the laryngeal motor cortex and the inferior parietal lobe (particularly area PF) in the human than in the macaque. This may be anatomical evidence of a strengthened predictive coding circuit for learned vocalizations (Hickok, 2016), that probably conveys both auditory and visual information, and includes mirror neurons as well.

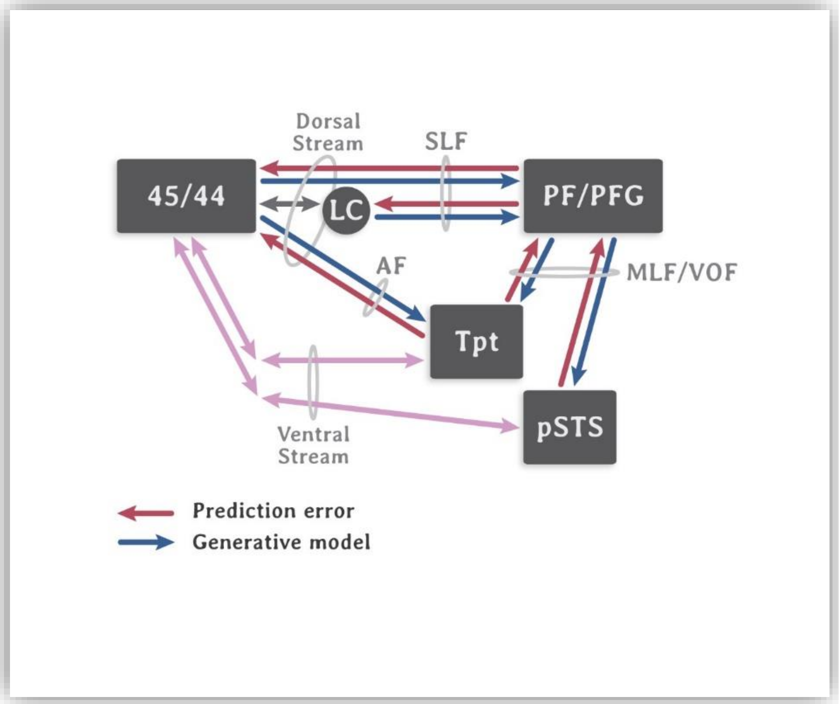

Figure 2.3: Proposed network for speech processing, distinguishing backward projections (Generative model) providing ongoing motor information into the sensory systems, which is compared with actual sensory input and sent forward as a prediction error. This network partly overlaps with regions of the Mirror Neuron System, especially in the posterior superior temporal sulcus (pSTS) and in inferior parietal areas (PF/PFG).

We have mentioned that proponents of the MNS hypothesis claim that manual imitative gestures triggered the development of vocal mimicry in early humans. In contrast, in the trimodal repertoire model proposed in this chapter, the emphasis is rather put on the auditory-visual-vocal circuits and the capacity to imitate vocalizations supported by the observation of others' orofacial gestures. Noteworthy, mirror neurons have been demonstrated to fire not only in response to executed or seen actions but also in response to heard actions (Keysers et al., 2003). As mentioned above, Hage & Nieder (2015) reported that single neurons located in the VLPFC were modulated both when monkeys produced or listened to conspecific vocalizations, a feature strongly reminiscent of mirror neuron properties. These results indicate that, in addition to the integration of audio-visual (i.e., phoneme-viseme) inputs, the VLPFC seems well positioned for the integration of audio-motoric input (i.e., phoneme-*articuleme*) turning it into a critical region for the evolution of a complex audio-visual-motoric integration that may have contributed to the emergence of speech in humans (Hage & Nieder, 2015). Analogously, from an ontogenetic perspective the existence of those imitative behaviors seems to be relevant for speech perception and the construction of the phonologic repertoire during the first year of life. A behavioral study conducted on 6 months-old infants who used teething toys in order to interfere with their imitative orofacial movements, showed that the auditory discrimination between nonnative phonemes was impaired in these subjects compared to infants not using teething toys. The outcomes of this experimental manipulation on infant's performance provides evidence of how the sensorimotor information from the articulators influences speech perception (Bruderer et al., 2015).

2.6. Discussion

Structural, functional and behavioral evidence comparing the evolution of the vocal system of humans and NPHs and their neural circuits have been reviewed along this chapter. We first mentioned the anatomical parallelism and overlap between auditory and visual systems which are organized in a dorsal and a ventral pathway and are present in both species. The dorsal stream projects into area 44 conveying phonological information from the posterior auditory cortex and the anterior inferior parietal lobe, as well as sensorimotor inputs from premotor and motor cortices. In this sense, the dorsal pathway has an important role in the transduction of phoneme into *articuleme.* On the other hand, the projections of auditory and visual ventral streams overlap in the VLPFC where voices and faces are associated. Homologies and differences between human and NHP neuroanatomy have also been addressed to foster our understanding of the evolutionary trajectory that may have led to the emergence of speech. First, we saw that the structural connectivity of the dorsal stream via the AF and the SLF seems to have strengthened from monkeys to humans since it is more robust and direct in humans compared to NHPs. Likewise, behavioral and functional connectivity analyses suggest that NHPs are more likely to rely on the ventral auditory stream to associate vocalizations and orofacial gestures whereas the human lineage took advantages from the growth of the dorsal pathway to increase vocal working memory capacity and amplify the vocal repertoire (Aboitiz, 2017, 2018b). Finally, the auditory-vocal network is claimed to have evolved in step with the control of the larynx via direct and robust projections to the nucleus ambiguus.

Special attention has been given to the visual processing of speech in this chapter. Evidence consistently reports that not only visual but also auditory and motor cortices are engaged during the perception of orofacial gestures and that the access to congruent or incongruent visual speech cues has a large influence on auditory processing. The latter leads us to propose the hypothesis of a trimodal repertoire where auditory, visual and motor aspects of speech are transduced; the smallest units of these modalities being the phonemes, the visemes and the *articulemes*, respectively. Observation and imitation are proposed as tentative general mechanisms fostering the construction of the trimodal repertoire. The presence of neural populations displaying mirror properties within the traditional speech circuitry in the left hemisphere may support the associative learning of observed orofacial gestures and the motoric sequence required to produce that gesture. Moreover, the reciprocity of the projections permits to generate predictions and minimize errors during speech perception.

Alongside with the evolution of the brain matter, the human mind has also evolved. The increasing complexity of ancestral societies urged adaptive and selective behaviors. The cortical and sub-cortical adaptations for language are necessary but may not be sufficient to fully explain the origin of speech. The need to organize communication within highly social communities is likely to have contributed to the emergence of increasingly richer vocalizations. Consistent with this idea, the complexity of vocalizations in birds is associated with more elaborate social behavior such as collaborative breeding (Leighton, 2017) and highly social mammals like cetaceans, elephants or orangutans are able to imitate the human voice, which points to an increasing relevance of vocal communication as social interactions become more complex (Aboitiz, 2018).

Finally, we would like to stress that human communication is, and has always been, a multimodal process in which gestures, vision, voice and audition play a role. The gestural theory of language origins views learned gestures like pantomimes and so-called "protosigns" as the initial channel for symbolic human communication, which through some unknown process triggered the acquisition of vocal mimicry and the development of speech (Arbib, 2012). However, there is no compelling reason or evidence indicating that vocal learning had to wait for symbolic gestural communication to develop first. As we have said, other species have developed vocal learning capacity without need of a hand gestural repertoire. In other words, our point is not whether the first symbolic messages were first carried by gestures or by voice (a purely speculative issue for which that we may never have a definite answer), but rather to track the evolution of speech mechanisms from ancestral multimodal networks in non-human primates. Considering this, there is another possible scenario where vocal imitation appeared early in the genus *Homo* (or even in australopithecines), and coexisted with gestural communication for a long time. According to this view, in these animals vocal mimicry was more related to social bonding, particularly between mother and child, rather than transmitting referential contents. This kind of communication was supported by the trimodal visual-auditory-motor repertoire we have proposed in Fig. 2.2. At some undetermined point in our history, early humans started making reference to events using gestural pantomimes, vocal imitations, vocal alarms like those of vervet monkeys and any other possible way to communicate simple meanings (Aboitiz, 2018). However, the progressive development of the dorsal auditory-vocal pathway enabled them to displace the gestural modality and to establish early speech as the main communication channel. This perspective has the advantage of providing a preexisting visual-auditory-vocal scaffolding from which the speech circuitry could later emerge to transmit symbolic content, an aspect that is in our view neglected by the gestural theory of language origins.

Chapter III:

Do you see what I am saying?

Electrophysiological dynamics of visual speech processing and the role of orofacial effectors for cross-modal prediction.

This chapter is based on:

Michon, M., Boncompte, G., Lagos, R. & López, V. (2019). Do you see what I am saying? Electrophysiological dynamics of visual speech processing and the role of orofacial effectors for cross-modal prediction. Under Review in *Neurobiology of Language*

Abstract

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 showed 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.

Keywords visual speech; place of articulation; orofacial movements; motor system; crossmodal prediction; ERPs

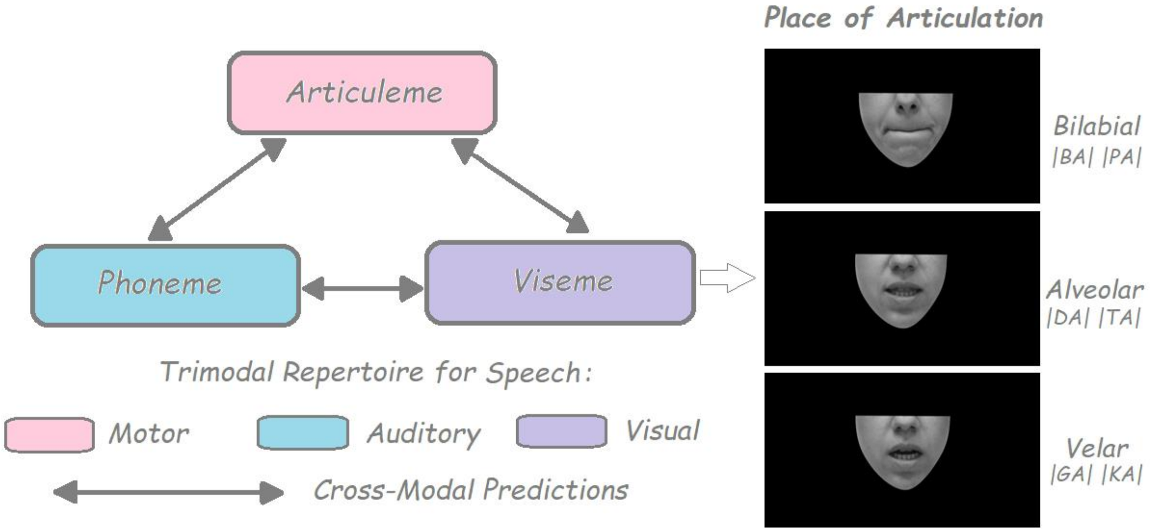

Figure 3.1: Graphical Abstract

Highlights:

- *The amplitude of N270 and N400 was modulated by the place of articulation of the syllables.*
- *No modulation was observed for backward syllables or other types of orofacial movements.*
- *The amplitude of N400 was significantly reduced when orofacial effectors were not available.*
- *The language production system is recruited for pre-lexical perception and prediction.*

3. Do you see what I am saying? Electrophysiological dynamics of visual speech perception and the role of orofacial effectors for cross-modal prediction

3.1. 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 in silent videos a switch from native to non-native language (and vice versa), 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 seem 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 to consider the activity 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 (Miller et al., 2018; 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. The left posterior superior temporal sulcus (pSTS), considered as a central hub for multimodal integration, 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 (Arnal & Giraud, 2012; Peelle & Sommers, 2015). For instance, greater functional connectivity has been found between left pSTS and auditory-speech areas when visual cue mismatch the 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). Even though 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 (Arnal et al., 2009; Paris, Kim, & Davis, 2013). 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 and silent 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.

In contrast to the growing body of studies evidencing the involvement of motor and premotor cortices during language perception, the electrophysiological data available about the visual processing of speech is still scarce. If language perception implies the participation of the same motor systems necessary to articulate what is being perceived, then the visual processing of orofacial movements should be differentially affected when the movement is language related or not. This information should be integrated early enough to activate the trimodal repertoire. And this process could be affected by interrupting the effectors of the perceived action. In the frame of the hypothesized trimodal repertoire for speech, we expect that 1) the absence of auditory input generally perceived subsequently to the orofacial movements and 2) the restrained mobility of lips and tongue will differentially affect the electrophysiological responses evoked by linguistic versus non-linguistic stimuli.

3.2. Methods

3.2.1. 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 types 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 kinds 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 fourth condition, non-linguistic orofacial movements (e.g., tongue protrusion, 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, 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 (Pierce, 2007). Importantly, all the videos were silently displayed (i.e., audio removed) and they 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.

3.2.2. 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 ranges 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.

3.2.3. Procedure

Participants sat at a distance of 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 (Peirce, 2007). 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. The imitation condition was included as a sham task to the participants. 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.

3.2.4. Electroencephalographic recording parameters

Electrophysiological activity was registered with a 64-channels 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-referencing. Data pre-processing was performed using MatLab (The Mathworks, Inc.), EEGLAB (Delorme & Makeig, 2004) and ERPLAB toolbox (López-Calderón & Luck, 2014). The signal was first down-sampled at a rate of 512 Hz, re-referenced to mastoids 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 Independent Component Analysis (ICA) decomposition and the rejection of components typically associated with eye-blinking, epochs exceeding maximum amplitude of ±100 μV were removed.

3.2.5. 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 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 experiment (1 and 2), with conditions as fixed factor and subjects as aleatory factor. Statistical analysis was performed using the *lme4* (Bates et al., 2015) package of R. The pairwise comparison within conditions was obtained using *emmeans* (Lenth, 2016) package of R, that permit to compare slopes in a mix model.

3.3. Results

3.3.1. Experiment 1

The evoked responses over FCz electrode for all conditions can be observed in Figure 1a. The ERP components P2, N270, N400 and P1000 can be identified. Upon closer inspection of these waveforms 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 non-linguistic orofacial gestures in electrode FCz (t=3.186, p=0.0279, ETA= 0,398, d= 0,87 and t=3.528, p=0.0095, ETA=0,46, d= 0,96 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 was more located in the posterior-occipital regions (Fig.3.2b).

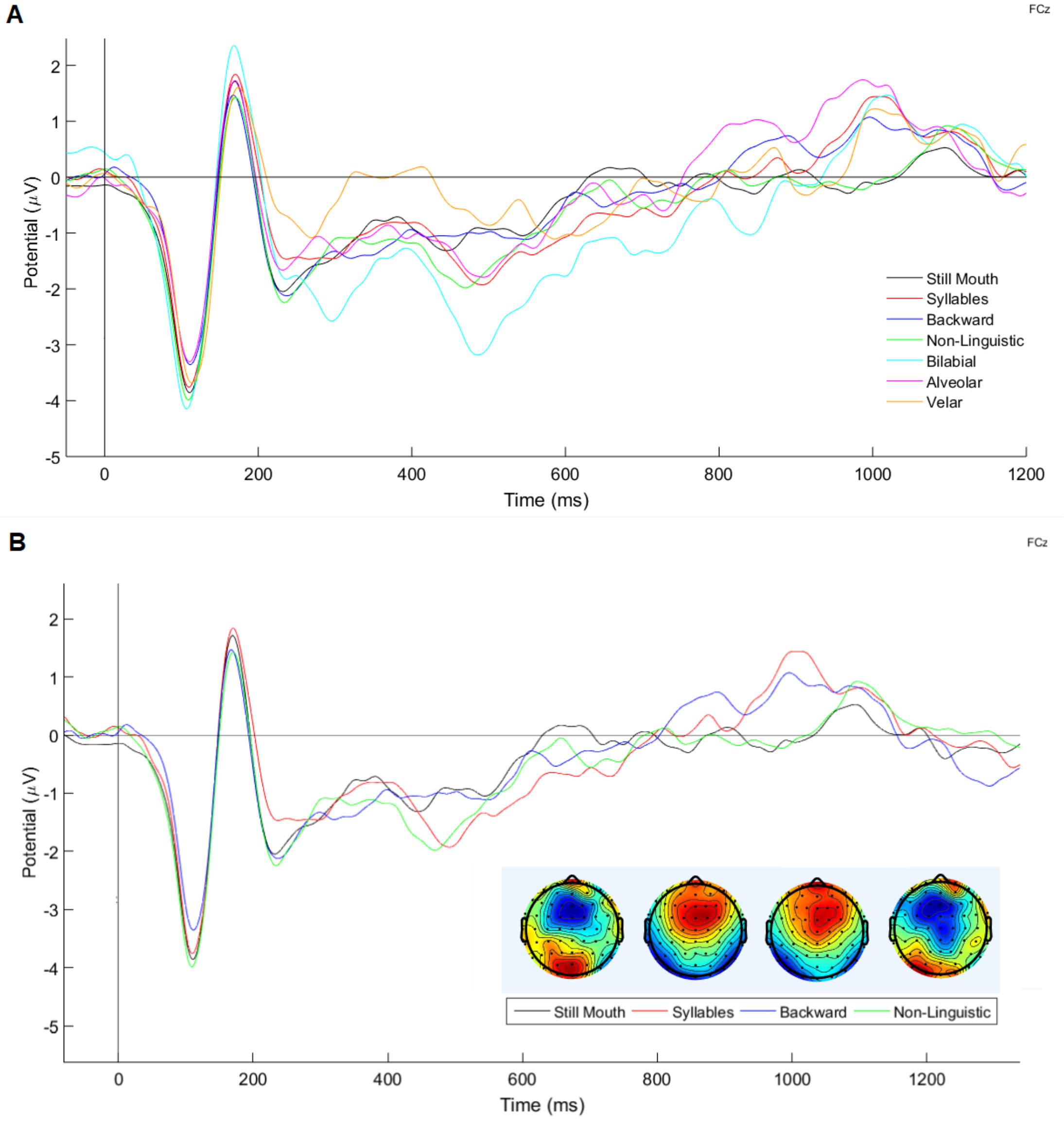

Figure 3.2: ERPs across all conditions (electrode FCz).

Since ERPs responses were strongly modulated 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= -3.757, p=0.0043, ETA= 0,45, d= 1,02) and in N400 compared with velar syllables (t= -4,609, p = 0.0002, ETA= 0,53, d= 1,25) and alveolar syllables (t= -3,099, p=0.0361, ETA= 0,39, d= 0,84). No significant differences were found for components P2 and P1000 (Fig.3.3).

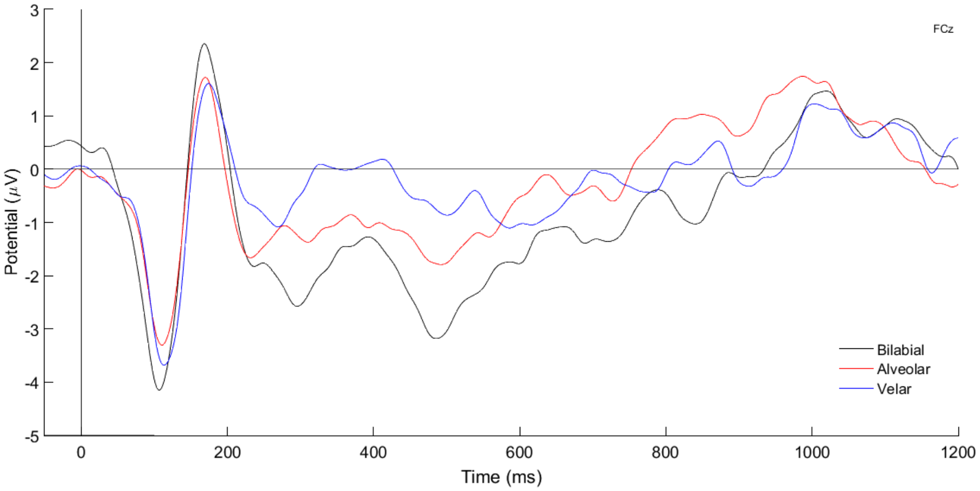

Figure 3.3: ERPs of the syllables as a function of their PoA (electrode FCz).

3.3.2. Experiment 2

The amplitude of the N400 component was significantly greater for the bilabial syllables in experiment 1 compared to experiment 2, in which participants held an effector depressor between their teeth (t= 3.423, p= 0.043, ETA= 0,34, d=0,74). In contrast, the amplitude of N270 didn't differ between the two experiments (Fig. 3.4). No significant differences were found between experiment 1 and 2 for the other type of syllables nor for non-linguistic movements.

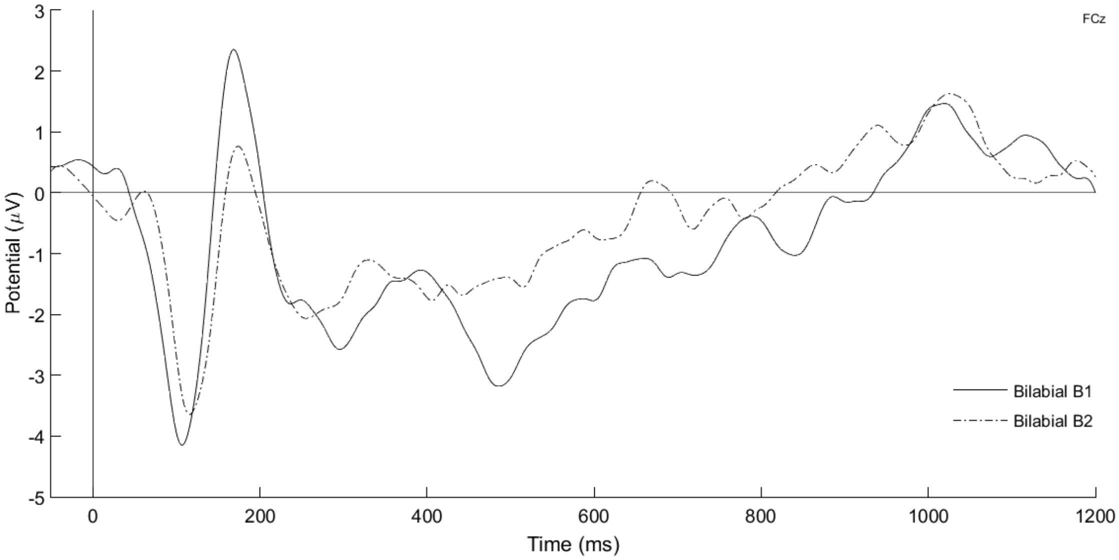

Figure 3.4: The effect of effector depression on bilabial syllables processing (electrode FCz).

3.4. 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 on the amplitude of N400 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" (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 strengthens 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 counterpart.

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 the auditory and motoric counterparts. An alternative interpretation can be formulated based on 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 of 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 agrees with results suggesting that the availability of 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, therefore the subsequent auditory cues should be more predictable. So, when the sound 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 these 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 forward 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 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 involvement of motor systems. Further studies analyzing time-frequency domain or electromyographic responses of orofacial muscles could be more conclusive in that respect.

In summary, it was hypothesized that the omission of the auditory input as well as the restriction of lips and tongue movements would differentially affect the ERPs elicited by linguistic versus non-linguistic stimuli. We reported two main electrophysiological findings. First, the amplitude of N270 and N400 were clearly modulated by the visual salience of the syllables, the first component suggesting the detection of a perceptual incongruency and the latter indicating cross-modal prediction error. Second, the effect of the PoA was significantly reduced by the effector depression, suggesting the involvement of motor system during prelexical speech perception and prediction.

Chapter IV:

Visemic salience modulates μ suppression during silent speech perception: preliminary evidence for the articuleme.

This chapter is based on:

Michon, M., Boncompte, G., López, V. Visemic salience modulates μ suppression during silent speech perception : preliminary evidence for the articuleme. To be submitted.

Abstract:

The involvement of motor cortices during auditory speech perception has been increasingly documented. Particularly, mirroring activity has been reported in EEG studies to be evidenced by the suppression of µ rhythms in central electrodes. In the frame of sensorimotor integration of speech, a fascinating question remains regarding the potential motor involvement in visual speech perception. The current study aims to determinate if the processing of visual counterpart of phonemes (i.e., the visemes) activates motor regions. Thirty participants observe or imitate silent videos displaying a talking face either producing different syllables (stop consonant + vowel) or backward syllables (i.e., nonlinguistic orofacial movements). The syllables presented differed in their place of articulation (bilabial vs alveolar vs velar) and consequently in their visemic salience. The time-frequency dynamics of the EEG signal were analyzed and evidenced a significant suppression of µ rhythms during the perception of syllables compared to backward syllables. Interestingly, the suppression of µ rhythms was modulated by the visemic salience, the bilabial syllables eliciting more suppression than velar syllables. Moreover, when the orofacial effectors of the participants were restricted, the µ suppression found for bilabial syllables significantly diminished. Altogether, these results suggest that visual speech cues perceived without any auditory counterpart are represented in motor cortices and that the amount of this motoric representation depends on visual salience and on the functional availability of the perceiver's orofacial effectors. The results of the current study are discussed in the frame of a hypothesized trimodal model network where phonemes, visemes and *articulemes* (i.e., the vocal-motor sequence required to pronounce a phoneme) are mapped in a cross-modal repertoire.

Keywords: µ-suppression; visual speech; place of articulation; articuleme

4. Visual salience modulates μ suppression during silent speech perception: preliminary evidence for the articuleme.

4.1.Introduction

4.1.1. Audiovisual speech integration.

During most face-to-face interactions, the perception of the speakers' orofacial movements provides listeners complementary information for speech processing. Visual speech cues have been demonstrated to be particularly helpful to disambiguate acoustic input during challenging listening conditions, such as noisy environment, competing talkers, foreign accent and hearing impairment (Peelle, 2019). Since they typically precede their auditory counterparts when speech is perceived, visual speech cues facilitate the auditory processing of speech. In a behavioral study, Paris, Kim and Davis (2013) reported that the access to orofacial movements speeds up the processing of speech sounds and that this facilitation effect was modulated by the salience of visual speech. These results are in line with the suppression of the amplitude and latency of auditory component N1 and P2 during audiovisual compared to auditory only speech perception, an effect documented by several EEG studies (van Wassenhove, Grant & Poeppel, 2005, 2007; see Baart, 2016 for a meta-analysis). Using fMRI, Calvert and al. (1997) showed that visual the perception of a speaker's lip movements activates auditory cortex. A posterior study based on intracranial recordings in human brain demonstrated that secondary auditory areas are activated very shortly (~10ms) after the activation of movement visual area MT/V5 (Besle et al., 2008). In addition to this temporal facilitation, the visual cues of the speaker's orofacial movements offer a restricted set of possible subsequent auditory input, allowing the listener to generate prediction about the forthcoming speech (Peelle & Sommers, 2015). Accordingly, performing Grainger causality analysis of MEG signal, Hauswald and al. (2018) observed that, when presented to silent lip movements, participants' visual cortex exhibits stronger entrainment to the absent acoustic envelope of intelligible (i.e., forward) vs. unintelligible (i.e., backward) speech. Interestingly, the authors also reported that the respective occipital region received more top-down input from motor cortices and proposed that motor-related areas of the dorsal stream mediate the viseme-phoneme mapping by exerting top-down control of the visual cortex (Hauswald et al., 2018).

4.1.2. Auditory-motor mapping.

The association between speech sounds and articulatory representations of speech are thought to be supported by a cortical circuit extending from temporo-parietal vicinities and projecting into frontal regions. In the dual-route model of neurobiology of language (Hickok & Poeppel, 2007; Rauscheker, 2012) this circuit is known as the dorsal stream and postulated to be essential for speech production with a minor, marginal role during speech perception. Recently however, this view has been challenged by an increasing number of studies that consistently reports the presence of activity among motor-related areas during auditory processing of speech (Wilson et al., 2004; Wilson & Iacoboni, 2006; Cheung et al., 2016). Using event-related fMRI, Pulvermüller et al. (2006) demonstrated that the perception of speech sounds requiring lip or tongue movements (e.g., ‘p’ vs ‘t’) to be produced elicited a somatotopic activation in the precentral gyrus, suggesting that specific motor circuits are involved when perceiving distinctive phonetic features as a function of the articulatory effectors recruited to produce them. This somatotopic mapping of heard speech sounds into articulatory codes according to the place of articulation was recently documented using EEG (Bartoli et al., 2016). In this study, peripheral electrical stimulation of the participants’ lips was used while they performed an auditory discrimination task. The results showed that the lips stimulation elicited a decrease in the beta-rebound, a spectral modulation known as an index of the “return to baseline stage of somatosensory processing” (Bartoli et al., 2016, p.2). The authors proposed that return to baseline could not happened due to the involvement of the same neural circuit during speech perception. Interestingly, this reduction in the beta-rebound occurred when listening to bilabial syllables (i.e., produced with the lips) but did not occurred for dental syllables (i.e. produced with the tongue).

An emerging field of cognitive neuroscience, known as neural entrainment, is providing promising insights about the neural mechanisms underlying the binding between auditory and motor cortices. Assaneo & Poeppel (2018) examined the synchronization between auditory and motor-speech regions in response to different speech rates and reported that oscillations in the respective cortices synchronize around a restricted range of frequency, reaching its higher synchronization at 4,5Hz. The theta band (4-7 Hz) has been associated with the mean rate of syllabic production across languages (Arnal & Giraud, 2012; Ding et al., 2017). Relevantly to the purpose of the current study, it has been shown that auditory cortex not only tracks speech sounds dynamics, but it also entrains to visual speech rhythm at low-frequency neural oscillations (Chandrasekaran et al., 2009; Luo, Liu & Poeppel, 2010; Crosse, Butler & Lalor, 2015).

4.1.3. Visuomotor binding: the missing link?

Six months-old infants have been shown to discriminate their native language from a non-native language by simply watching silent videos of a talking face (Weikum et al. 2007). By orienting their gaze and attention to orofacial movements (Lewkowicz et al., 2012), they take advantage of redundant visual speech information that, in turn, improve their expressive language skills (Young et al., 2009; Tsang et al., 2018). This visual orientation of attention deployed by infants in the second half of their first year of life have also been observed in adults, who increase their attention to the mouth relative to the eyes when exposed to unfamiliar languages (Barenholtz, Mavica & Lewkowicz, 2016). Together these results provide evidence that, when learning native tongue or, later in life, for understanding a foreign language, listeners spontaneously reorient the focus of visual attention to the lower part of the interlocutor face searching for additional cues that very likely help speech acquisition. Interestingly, in a TMS study conducted with adults, Swaminathan and collaborators (2013) reported that visual perception of speech elicited activity in motor areas and that this motor excitability was greater for known language than for unknown language or nonspeech mouth movements. The authors interpret this finding to suggest that motor resonance is especially robust during the observation of orofacial movements conveying meaningful linguistic information. Motor resonance between the speaker and the listener, as a result of the activity of mirror neuron system (MNS), represent a core argument in theories of speech evolution (Michon, López & Aboitiz, 2019).

In order to investigate the involvement of motor cortices during speech perception, an increasing number of researchers are analyzing the suppression of μ-rhythm (Cuellar et al., 2012; Saltuklaroglu et al., 2018; Thornton et al., 2018; Bowers et al., 2019). The human sensorimotor cortex displays oscillatory waves called μ-rhythm which are observable in the 10-to-12/13 Hz range depending on authors and whose activity is greater when the body is physically at rest. The suppression of this rhythm is thought to be linked with the activity of MNS since μ-power is reduced during both action execution and observation (see Fox et al., 2016 for a meta-analysis, but also see Hobson & Bishop, 2016). This assumption has been experimentally addressed in studies combining fMRI and EEG recordings in order to clarify if the brain areas of MNS identified by fMRI studies are effectively the origin of μ-suppression in the EEG signal. Arnstein et al. (2011) reported a covariation of μ-power around 10 Hz with blood oxygen-level dependent (BOLD) activity in the traditional MNS regions (i.e., IPL, premotor and primary sensorimotor cortices) during both observing and executing a hand-action. More recently, a study combining EEG and local field potential recordings in the ventral premotor cortex in monkey also advocates for the contribution of MNS to μ suppression (Bimbi et al., 2018).

In line with the later, and accounting for the evidence reported above about audiovisual, audio-motor and visuo-motor integration during speech perception, the current represent an in attempt to inquire the existence of a trimodal repertoire for speech in which the auditory (phonemes), visual (visemes) and motor (articulemes) units of speech are mapped and reciprocally predictable when one modality is missing or interfered. The current study aims to address an important but understudied aspect of speech processing, namely the binding between visual and motor cortices during silent speech processing. More specifically, through two experiments, we investigated 1) whether the perception of visual speech cues compared to nonlinguistic orofacial movements elicited a power suppression in the μ-frequency band and 2) whether the pattern of activity in motor cortices indexed by μ-suppression is affected by orofacial effectors depression. Congruently with the proposed trimodal network, we expect to observe a stronger suppression of μ-rhythms power in response to visual speech movements compared to other kind of movements. On the other hand, we hypothesize that the effector depression will interfered with the motor resonance processes, especially those elicited by visual speech cues.

4.2. Methods

4.2.1. Participants

30 right-handed subjects (16 females) with normal or corrected-to-normal vision and hearing and without any history of psychiatric or neurological disorders performed the experiments. Participants' ages 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 Pontificia Universidad Católica de Chile. Before the experiment started each participant was explained the procedure and signed an informed consent form. Six participants were removed from final analysis because of poor signal-to noise ratio in the EEG/ERP.

4.2.2. 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 types 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, meaning that they are articulated by closing the airway to impede the flow of air, thus generating a slight 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 kinds of syllables have been reported to have different levels of visual salience; bilabial syllables being more salient then velar syllables (Jesse & Massaro, 2010; Paris, Kim, & Davis, 2013; van Wassenhove, Grant, & Poeppel, 2007). In the third condition, the same syllables were played backwards. Because of their particular motoric sequence, stop syllables cannot be pronounced backwards. 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 (see section 1.3.). In the fourth condition, non-linguistic orofacial movements (e.g., tongue protrusion, lip-movements) were produced. This condition was introduced in order to control the activity associated to the processing of orofacial movements with no linguistic content. Finally, 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 (Pierce, 2007). Importantly, all the videos were silently displayed (i.e., with audio removed) and they only showed 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.

4.2.3. Procedure

Participants sat at a distance of 70 cm from the computer screen in a quiet and dark room 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 indicating the instruction, either “Observe” or “Imitate”. After 100 to 150 ms, a fixation cross appeared for 250 ms. In the observation condition, the video was displayed one time, 1000 to 1500 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 seconds (see Figure 4.1.). The imitation condition was included as a sham task to the participants in order to give them a meaningful task and the electrophysiological responses elicited in this condition were not analyzed or presented here. 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 perturbate orofacial effectors motion. 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] consisted 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.

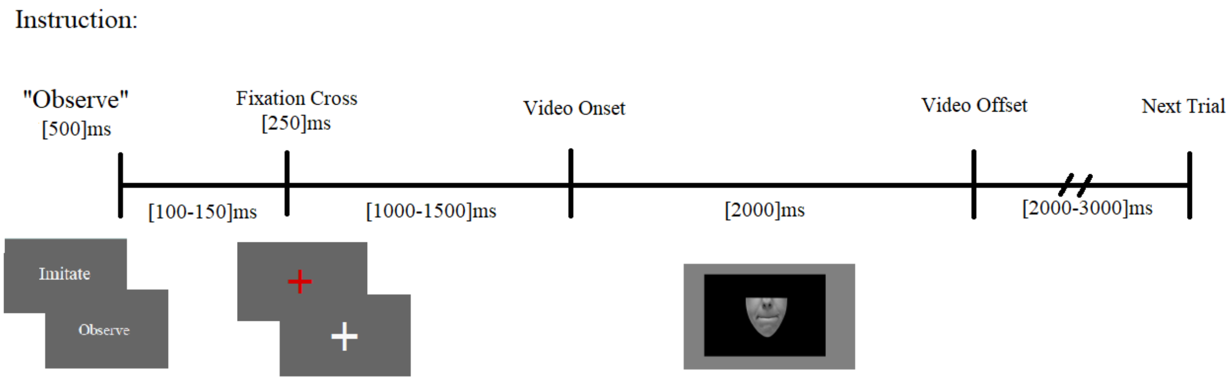

Figure 4.1: Description of the temporal sequence of the task.

4.2.4. Electroencephalographic recording parameters and statistical analysis.

Brain electrophysiological activity was registered with a 64 active channels (AgCl) EEG system (Biosemi ® ActiveTwo) with electrodes positioned according to the extended 10-20 international system. The sampling rate was of 2048 Hz. Four external electrodes were used to monitor eye movements. Two were placed in the outer canthi of the eyes in order to record horizontal electrooculogram (EOG), and the other two above and below the right eye, which recorded vertical EOG. Two additional external electrodes were placed on bilateral mastoids for re-referencing purposes. Data pre-processing was performed using MatLab (The Mathworks, Inc.) and EEGLAB (Delorme & Makeig, 2004). During data acquisition, the signal was recorded using active references (CMS-DRL). Offline, data was first down-sampled at a rate of 1024 Hz, re-referenced to the average of mastoids electrodes and band-pass filtered between 0.1 and 100 Hz. Then, EEG signal was segmented into epochs from -500 ms to 2800 ms respect to stimulus onset. Artifact rejection was performed using a mixture of automatized and manual procedures. Epochs were visually inspected in order to reject large artefacts due to head movements, eye movements or muscular contractions, in addition to an automatized epoch rejection algorithm flagging epochs in which voltage exceeded a maximum of ±100 µV after demeaning. Subsequently, Independent Component Analysis (ICA) decomposition was calculated to reject finer components typically associated with eye-blinking and others.

We calculated the spectral behavior of the signal by means of fast fourier transforms (FFT) applied to a window sliding in time for each artifact-free epoch in order to capture induced oscillatory behaviors. To this aim we used the following procedure. From each epoch (3300 ms) we calculated the FFT of 301 windows of the voltage signal in time (200ms; hamming windowing; zero-padding to 512 points), which overlapped by a 95%. This was done to grasp the time dynamics of the oscillatory activity. To normalize the spectral powers obtained we converted the signal to z-scores respective to a baseline period preceding target presentation (-400 to -100 ms) for each frequency independently. This data was then graphed in time-frequency (TF) plots.

To analyze difference between conditions related to µ power, we selected an area of TF plots corresponding to 10 to 14Hz and from 500 to 1500ms based on the clear behavior observed in TFs of all conditions. The values of these areas of each TF and each condition were averaged and then compared, across subjects, for each pair of conditions by means of the non-parametric Wilcoxson signed rank test for matched pairs. Effect size was computed a posteriori on G-power (Faul et al., 2007) and is reported as Cohen's *d* for significant results.

4.3.Results

As expected, observed movements induced a suppression of power in the µ-frequency band. All orofacial movements induced a significant suppression compared with their respective baseline. None of the control conditions, namely still faces, non-linguistic and non-human movements, elicited significant differences when comparing B1 vs B2 (see Figure 4.2.). However, syllables and backward played syllables, two critical conditions involved in our hypotheses, yielded significant effects. The timing and scalp topography of these physiological effects are depicted in Figures 4.3 and 4.4.

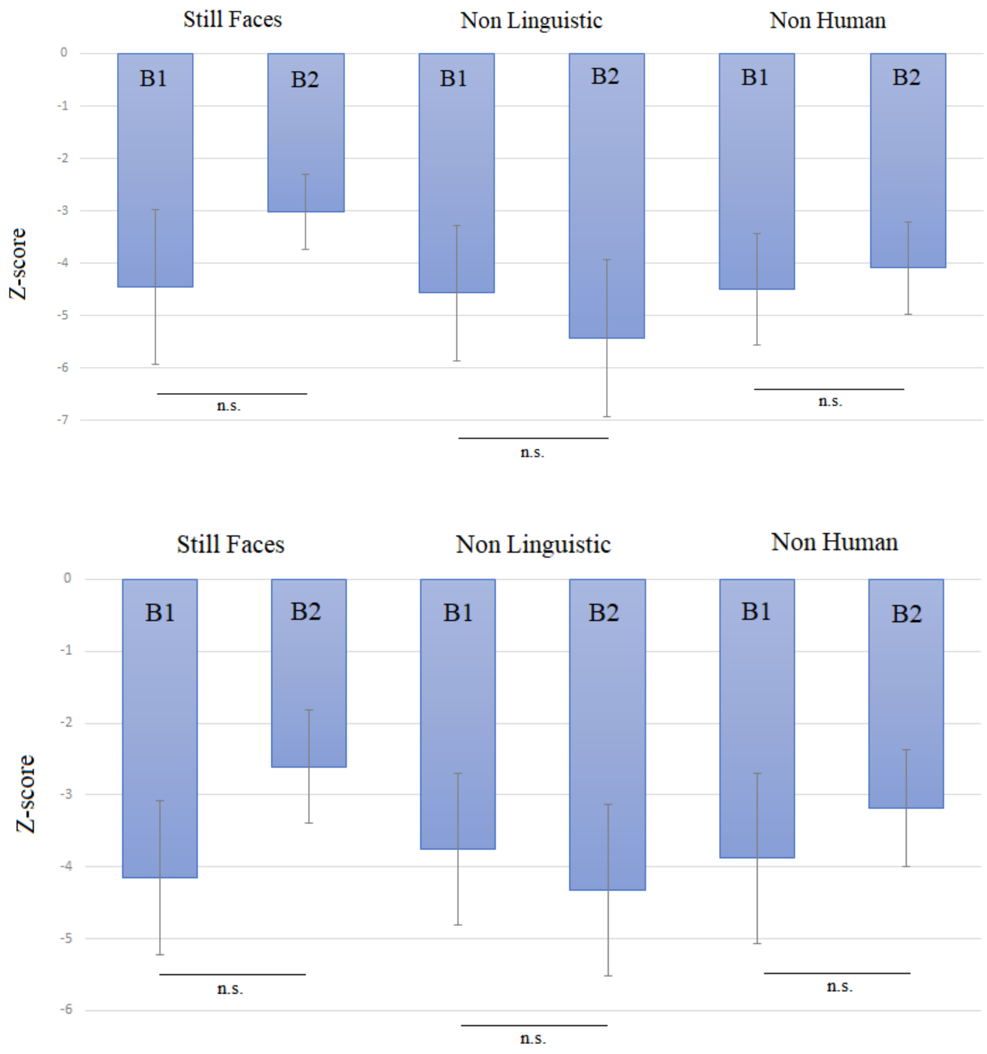

Figure 4.32: μ-suppression expressed in z-scores for control conditions in B1 vs B2. Top panel: Z-scores corresponding to the μ-suppression of control conditions in electrode C3. Bottom panel: Z-scores corresponding to the μ-suppression of control conditions in electrode C4.

4.3.1. Lateralization of μ-suppression elicited by speech vs speech-like orofacial movements.

Time-frequency analysis for B1 revealed that μ-suppression was significantly greater for the syllable condition compared to the backward syllable condition (Z= -2.09, p=.037; d=0.745). Surprisingly, this effect was observed in electrode C4, located in the right central region of the scalp (see Figure 4.3.), but no significant difference was found in the homologue electrode C3 in the left hemisphere (Z= -0.5143, p=.6071). This lateralization of the syllables vs backward syllables effect in the right hemisphere was confirmed by topographical maps depicted in Figure 4.4.

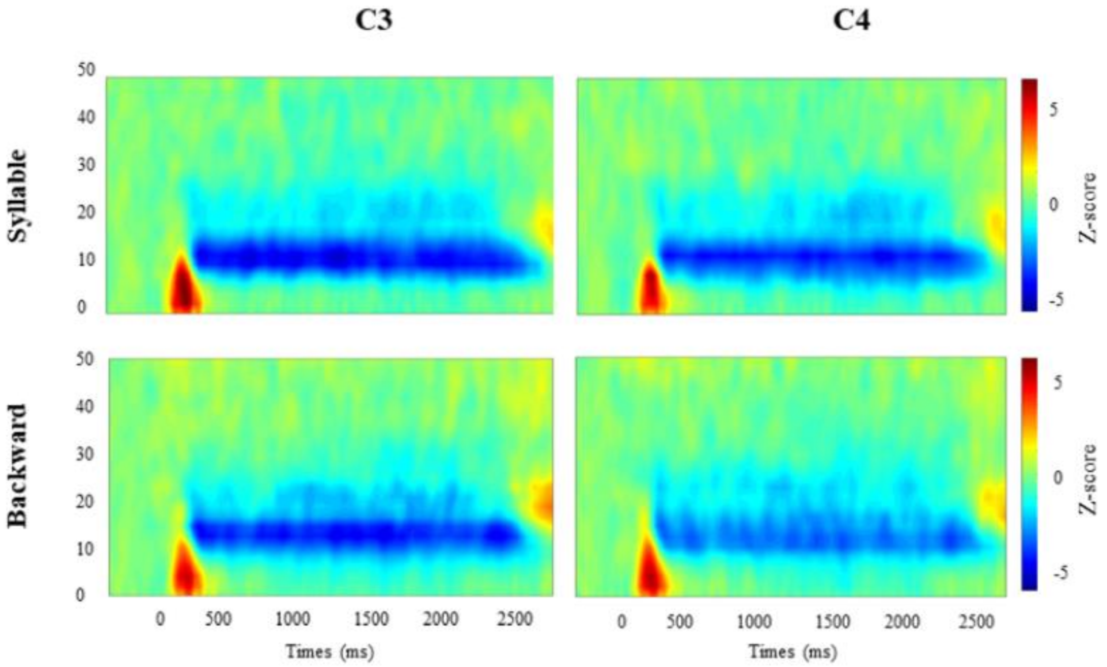

Figure 4.3.1: TF analysis for μ-rhythms for syllable (top panels) vs backward syllable (bottom panels) conditions. Color bar represents z-score values of power respect to the baseline period. Left: TF map illustrating the suppression of μ-rhythms in electrode C3. Right: TF map illustrating the suppression of μ-rhythms in electrode C4.

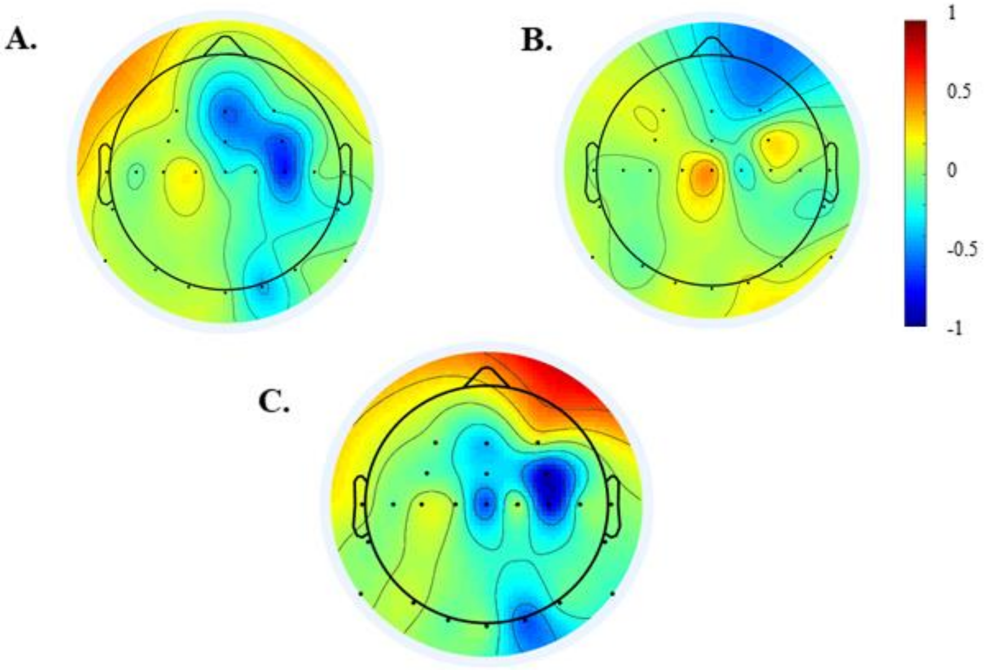

Figure 4.3.14: Topographical maps illustrating the lateralization of the µ-suppression elicited by speech vs speech-like orofacial movements in B1 and B2. A: Syllables minus Backward B1 (p=.0455). B: Syllables minus Backward B2 (n.s.). C: (Syllables minus Backward Syllables B1) minus (Syllables minus Backward Syllables B2).

4.3.2. The effect of effectors' depression on µ-suppression

Interestingly, the difference between syllables and backward syllables was not significant when participants observed the videos with the effector depressor between their teeth (Z= -1.1556, p=.2478; see Figure 4.5.).

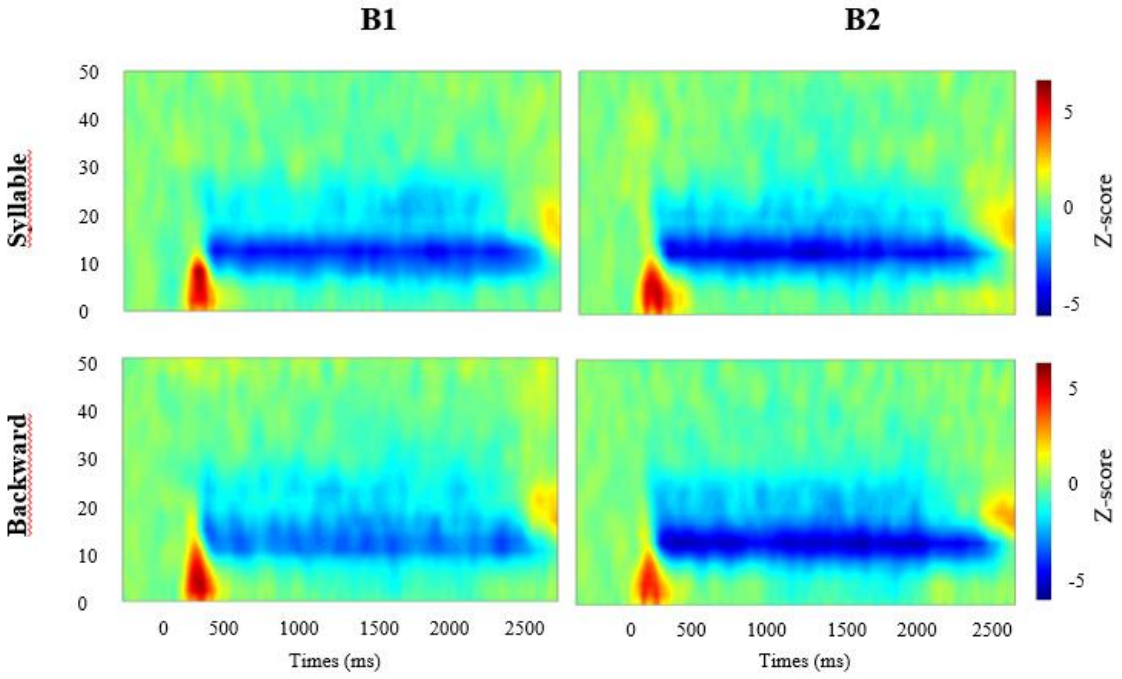

Figure 4.3.25: TF maps for syllables (top panels) vs backward syllables (bottom panels) conditions (electrode C4) with and without effector depression. Color bar represents z-score values of power respect to the baseline period. Left: TF map illustrating the suppression of µ-rhythms in B1. Right: TF map illustrating the suppression of µ-rhythms in B2.

4.3.3. The effect of visual salience and effector depression on µ-suppression.

Time-frequency analysis for B1 revealed that µ-suppression was greater for bilabial syllables compared to velar syllables in electrode C4 (Z= -2.54, p= .011; d= 0.96; Figure 4.6 top panel), but not in C3 (Z= -1.37, p=.170). Interestingly, this effect was not present in the second block when the effectors of the participants were interfered (Z= -0.65, p= .517). To further investigate these differences, we compared the amplitude of µ-suppression evoked by syllables that strongly differ in their visual saliency, namely bilabial and velar, in order to extract the effect of saliency on µ-suppression. This comparison was performed by subtracting µ-suppression amplitude of velar from bilabial syllables of B2 from B1. This subtraction was significantly greater than zero (mean= -1.72, Z=-2.543, p=.011, d=0.961) for Bl but was not significantly different from zero when the effectors were impeded in B2 (mean -0.221, Z=-0.648, p=.517). These differences were also significantly different from each other (Z=2.114, p=.0345, d= 0.766; Figure 4.6 bottom panel).

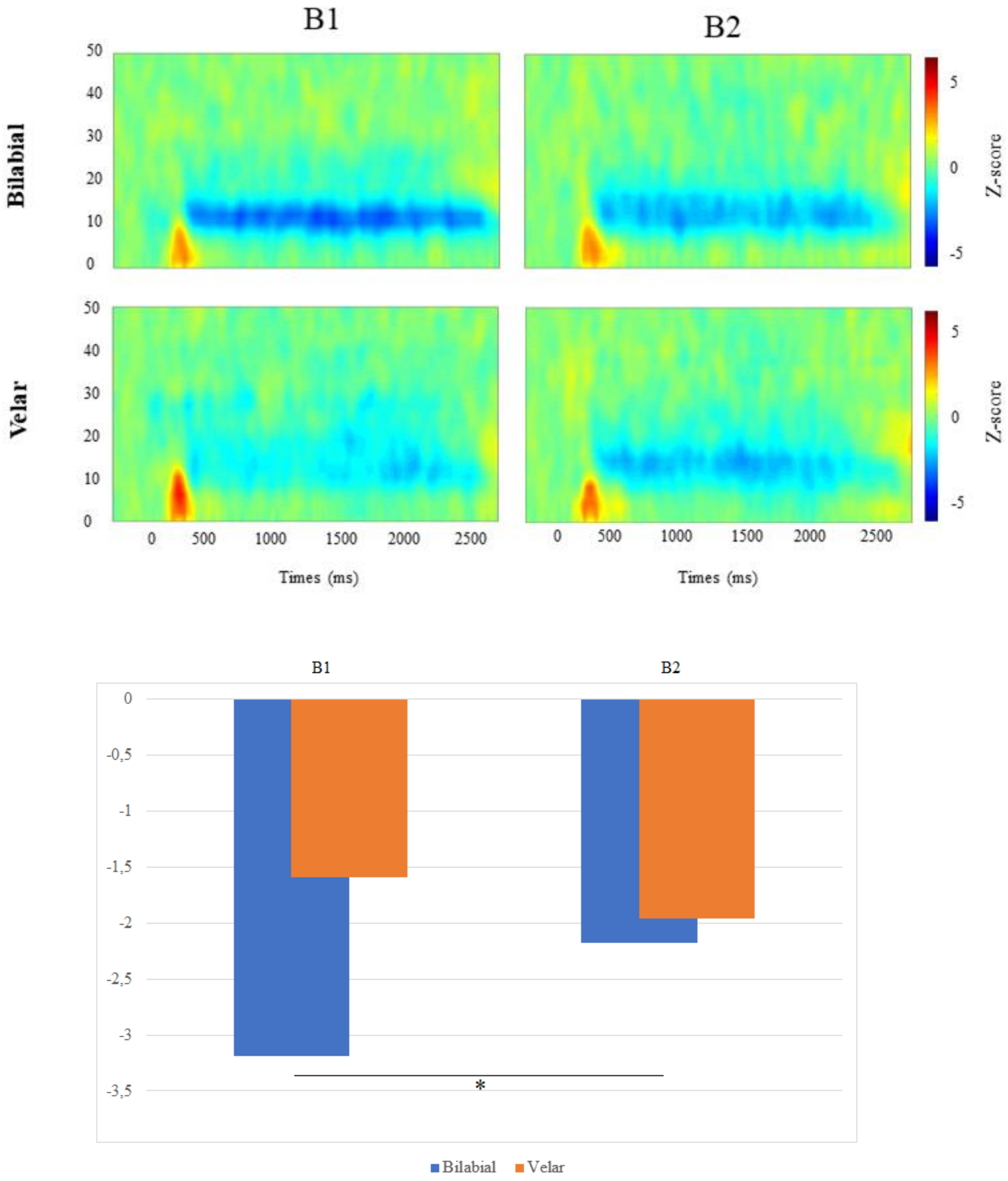

Figure 4.3.36: TF analysis of μ-rhythms as a function of the place of articulation of the syllables in electrode C4 for B1 (left) and B2 (right). Top: TF maps illustrating the suppression of μ-rhythms for bilabial and velar syllables in electrode C4. Bottom: μ-suppression expressed in z-scores for bilabial vs velar syllables in B1 and B2.

4.4. Discussion

The first objective of the current study was to address whether sensorimotor regions activated in response to visual perception of silent syllables versus nonlinguistic orofacial movements. To do so, we compared the power suppression of the µ-band frequency, known to be an index of mirror neurons activity, in response to linguistic and non-linguistic orofacial movements. Since stop syllables are not pronounceable backward, we suspected that they lack multimodal representation in the hypothesized trimodal repertoire. As a result, we expected them to elicit a smaller modulation of the µ-rhythms. In accordance with our hypothesis, the µ-suppression effect was significantly greater for the syllable condition compared to the backward played syllables. This result suggest that the observation of visual speech movements produced a specular activity in the observer's own motor system whereas biomechanically impossible movements did, but to a significantly lesser extent. Surprisingly however, this difference between speech and speech-like stimuli seems to be lateralized in the right hemisphere. More specifically, time-frequency analysis and topographical maps showed that this effect achieve its maximal power in the right frontocentral electrode C4, with no significant effect for its homologue in the left hemisphere (i.e., electrode C3). The interpretation of this result is challenging because it contrasts from the longstanding idea that linguistic processes are left-lateralized. Since the trimodal repertoire is hypothesized to develop in infancy through imitative behaviors (Kuhl & Meltzoff, 1982, 1996; Legerstee, 1990), a possible explanation for the right lateralization of this effect is that the activation of motor areas by visual speech perception is supported by the same imitation network involved during the emergence of crossmodal mapping in infancy. Nevertheless, in order to confirm that the generator of this frequency power difference is located in the right hemisphere it would be necessary to perform source analyses. Noteworthy, there is increasing evidence coming from ECoG and TMS studies that sensory-motor transformations for speech occur bilaterally (Cogan et al., 2014; Nuttall et al., 2018).

The second objective of the current study was to address whether the pattern of activity in motor cortices indexed by μ-suppression was affected by orofacial effectors depression. In order to test this assumption, we asked participants to observe the same visual speech and non-speech mouth movements while their orofacial effectors were blocked. This experimental manipulation aimed to interfere the automatic imitation processes known to occur during action observation. In line with our expectations, the difference in μ-power elicited by syllables versus backward syllables in B1 were not significant in B2. The pronounceable syllables are supposedly represented in the trimodal repertoire whereas non-pronounceable backward syllables do not belong to the trimodal repertoire. The fact that specular motor activity did not differentiated between biomechanically possible and impossible movements provides further evidence for the involvement of motor cortices during visual speech perception. Several lines of evidence support this view. For instance, Meister et al. (2007) reported that when transcranial magnetic stimulation is applied to the premotor cortex the ability to perform a phonetic discrimination task is disrupted, suggesting that motor cortices play an essential role in speech perception. Interestingly, blocking speech effectors, using similar strategies to the one used in the current study, has been demonstrated to affect not only visual speech recognition (Turner, McIntosh & Moody, 2014) but also the ability to generate lexical predictions during sentence comprehension (Martin, Branzi & Bar, 2018).

Unexpectedly, we also found an effect of the visual salience associated with the place of articulation of the different syllables. More specifically, μ-power in response to bilabial syllables was significantly greater compared to velar syllables. As previous studies reported, bilabial syllables are visually more salient compared to velar (Jesse & Massaro, 2010; Paris, Kim, & Davis, 2013; van Wassenhove, Grant, & Poeppel, 2007). Interestingly, the effect in μ-suppression for bilabial vs velar syllables was significant when participants freely observed the videos but was not significant when participants were asked to hold the effector depressor between their teeth, thus interfering with automatic imitation. This absence of effect in B2 is consistent with the proposed notion of articuleme. In effect, the movements the most affected by the effector disruption are the movements produced with the lips, particularly by bilabial occlusion. Since speech sounds perception has been demonstrated to be somatotopically represented in the motor cortex (Pulvermüller et al., 2006, Bartoli et al., 2016), it is very likely that lip-articulated syllables (|ba|, |pa|) were more affected by effector suppression than tongue-articulated syllables (|da|, |ta|). The later could explain why specular motor activity decreased in bilabial but not velar syllables.

In conclusion, the results of the current study provide preliminary evidence supporting the trimodal repertoire hypothesis. However, since no source reconstruction analysis were performed it is difficult to understand the results from an anatomical view. Replicating similar experimental procedure using neuroimaging techniques with greater spatial resolution could offer critical insights respect to the trimodal brain network underlying the binding of phonemes, visemes and articulemes for a seamless speech perception.

Chapter V:

General conclusions

« Lorsque Broca fit voir d'une manière précise les conséquences de la lésion de la troisième circonvolution frontale, il crut, et tout le monde avec lui, que le siège de la faculté du langage articulé était découvert. C'était une erreur : on avait simplement trouvé le groupe de cellules volitives qui, dans le but de l'articulation des mots, actionnent les muscles du pharynx, de la langue et des parois de la bouche. »

Fauvelle, 1886, p. 636.

« When Broca made a clear view of the consequences of the injury of the third frontal convolution, he believed, and everyone with him, that the seat of the faculty of articulated language was discovered. It was a mistake: we had simply found the group of volitive cells which, for the purpose of articulating words, activate the muscles of the pharynx, the tongue and the walls of the mouth. »

5. General conclusions

This chapter aims to integrate and contrast the thesis of a trimodal network for speech perception with the empirical outcomes of the current experimental manipulation. Conclusions are drawn, not only contributing to the theoretical framework of an enriched model for the neurobiology of language but also contributing to the understanding of the origin and evolution of human language.

As discussed along this dissertation, increasingly robust and consistent empirical data are challenging classical models of neurobiology of language. To account for this evidence that emphasizes the interplay between language perception and language production systems, a new framework is needed to understand brain mechanisms and cognitive processes underlying ecological communication. In an attempt to contribute to that necessary enterprise, we proposed the existence of a trimodal network model for speech perception, which highlights the importance of visual processing of speech related orofacial movements and its representation in motor cortices. In that sense, we hypothesized that auditory (phonemes), visual (visemes) and motor (articulemes) aspects of speech are bonded in a trimodal repertoire. This proposal arose from the crossroad between both phylogenetic evolution and ontological development of human language. We addressed the evolutive constrains and the neuroanatomical circumstances under which speech appeared across phylogenetic history of human lineage. Also, we revised evidence that infants adopt visual strategies during face processing to form their phonological and articulatory repertoire.

The aim of this dissertation was to investigate the hypothesized mechanisms underlying this trimodal network, namely mirror activity, automatic mimicry and cross-modal predictions. In order to test our hypothesis, we recorded EEG signal while participants were attentively observing different type of linguistic and non-linguistic orofacial movements in two conditions: under normal observation and observation holding a speech effector depressor horizontally between their teeth. ERPs analyses of the signal provide evidence of cross-modal predictions indexed by the N270 and the N400-like components. The amplitude of these components was specifically modulated by the visual salience of visual speech cues; the more salient the more predictable. Interestingly, when orofacial effectors were restricted, the amplitude of N400 was significantly reduced, suggesting that language production system is recruited for predictions. The time-frequency analysis, on the other hand, demonstrated the involvement of motor cortices for visual speech perception. More specifically, a significant difference in the µ-suppression was observed between linguistic and non-linguistic orofacial movements. The power of the µ-suppression was modulated by visual salience but diminished for the more salient visual speech cues when the participants orofacial effectors were blocked.

Predictive coding relies on the bidirectional interactions between frontal and motor cortices, that in turn exert top-down influences on sensory cortices (auditory and visual; see section 2.5.2). As mentioned before, in the absence of a modality the remaining sensory cortex tracks the missing input. The results of experiment 1 suggest that speech-sound omission produces greater prediction error, as indexed by the N400, when the syllables were particularly salient. The later strongly suggests that visual cortex tracks the missing auditory signal (Park et al., 2016; Hauswald et al., 2018) and advocates for the existence of phoneme-viseme binding. Moreover, since this effect did not occur when articulatory system was unavailable, it also indicates a functional role of the articuleme for generating cross-modal prediction errors. Vocal imitation and mimicry, on the other hand, have been proposed to play a crucial role both in the ontological development and in the phylogenetical evolution of speech. These imitative behaviors have been extensively studied in both humans and NHPs and represent a core argument in several evolutive theories of human language origin. Moreover, they emerge very early in infancy (Kuhl and Meltzoff, 1996) and are thought to be crucial for predictive coding and error minimizing during speech perception (Ray & Heyes, 2011). As mentioned in Chapter 2, one of the most influential model of neural circuitry for imitation (Iacoboni & Dapretto, 2006) suggests a strong overlap with MNS. The model includes the IFG, the vPMC, the

IPL and pSTS in the right hemisphere. Strikingly, the left-hemisphere homologues of the very same areas are largely known to be part of language circuits (see section 2.5.2.). Our results surprisingly showed that the µ-suppression differences in response to forward versus backward syllables were significant only in right central electrodes. At the first glance, this result seems contradictory to the mainstream knowledge that most language processes are left-lateralized. Despite this longstanding idea, there is now "increasing evidence of right hemisphere involvement in language processing" (Hagoort, 2017, p. 200; Gajardo-Vidal et al., 2018; Vilasboas, Herbet & Duffau, 2017; Rolland, Herbet & Duffau, 2018). Since we have no visual feedback about our own orofacial movements, the observation and imitation of our conspecifics' or care-givers' visemic production gains all its relevance for the acquisition of our articulatory repertoire (Venezia et al., 2016). Because "imitation requires the imitator to solve the correspondence problem – to translate visual information from modelled action into matching motor output" (Ray & Heyes, 2011, p. 92), it turns the right hemisphere imitation network into a possible candidate that supports the development of viseme-articuleme binding.

Some limitations of the current dissertation have to be addressed. Perhaps the more important aspect that limits our interpretation of the results is that the shame task introduced in order for the participants to be attentively involved in the task. Participants were asked to carefully observe the presented orofacial movements in 90% of the trials and to imitate them in 10% of the trials. The expectancy of the imitation trials may have predisposed the participants to covertly mimic the gestures, including in observation trials. If this is the case, it is likely that this volitional covert mimicry exacerbated the suppression of µ-rhythms. In that sense, the effects observed in our results could reflect an intentionally drive imitation rather than an automatic, unconscious mimicry as we hypothesized. This is unlikely, however, because no high expectancy-related electrophysiological or behavioral indicators were found regarding imitation trials. Besides, such expectancy effect should be evenly distributed among all conditions, rendering difficult to explain between-condition differences (i.e: syllable vs backward) from these more general factor. The other main limitation concerns the possibility to account for the neurobiological underpinnings of the hypothesized trimodal repertoire. Since EEG recordings have limited spatial resolution and given that no brain source reconstruction analyses were performed, the interpretation of our results in terms of neural circuits is restricted.

As a whole, the thesis developed along this dissertation is an attempt to demonstrate that speech processing is a multimodal phenomenon which is achieved by the mean of domain general, social and cognitive abilities such as imitative behaviors and predictive coding, among others. In chapter 2, we discussed the evolution of a dual neural pathway for speech along with a possible network for cross-modal prediction generation that involve the pars opercularis and triangularis of the IFG, the IPL, the pSTS, the laryngeal cortex and area Tpt. However, in order to effectively test the neural correlates of the trimodal repertoire, further investigations using neuroimaging techniques with higher spatial resolution are crucially needed. Particularly, a finest understanding of how the visuomotor binding of speech is supported by brain activity has the potential to shade lights on a number of related fields of research. For instance, respect to the right hemisphere involvement in speech processing, Correia, Jansma & Bonte (2015) used fMRI to study the neural decoding of the articulatory features of different syllables and reported that place of articulation was mainly decoded in the right temporo-parietal and frontal regions (Correia, Jansma & Bonte, 2015; Archila-Meléndez et al., 2018). Interestingly, µ-rhythms dynamics have been reported to differentiate between people who stutter and fluent speakers during both overt and covert production tasks but also during passive listening tasks and auditory discrimination tasks (Saltuklaroglu et al., 2017; Jenson et al., 2018), reflecting auditory-motor integration deficits. Research on the potential role of the visuomotor network in stuttering people is still scarce. In our opinion, this is a question that might deserves more dedicated research, since the increased reliance on visual speech cues could be beneficial for people who stutter. As mentioned earlier, the use of visual speech cues has also been intensively studied in individuals with autism spectrum disorder. The processing of visible articulatory information have been proved to improve the perception of spoken words in autistic children (Schelinski, Riedel & von

Kriegstein, 2014), leading a group of researchers to develop an application called Listening to Faces (L2F). In this application, children are presented with videos of a speaker producing monosyllabic words asked to associate these words with the corresponding image. When they fail to correctly associate the word with the picture, children receive a feedback consisting in a red arrow pointing to the mouth of the speaker and an auditory message saying “Look at the mouth”. The authors claimed that 8 to 10 years-old autistic children improved their performance for untrained words (Irwin et al., 2015). Similarly, the development of an application that integrate the speaker’s orofacial gestures could have greater efficacy in the frame of foreign language learning and particularly improve pronunciation.

In summary, we propose that understanding speech processing in the frame of a trimodal repertoire may have profound implications to the field of language research and its practical applications.

6. References

Aboitiz, F., 2012. Gestures, vocalizations, and memory in language origins. Frontiers in evolutionary neuroscience, 4, 2. *https://doi.org/10.3389/fnevo.2012.00002*

Aboitiz, F., 2017. A Brain for Speech. Palgrave Macmillan UK.

Aboitiz, F., 2018a. A Brain for Speech. Evolutionary Continuity in Primate and Human Auditory-Vocal Processing. Frontiers in neuroscience, 12, 174. *https://doi.org/10.3389/fnins.2018.00174*

Aboitiz, F., 2018b. Voice gesture and working memory in the emergence of speech. Interaction Studies, 19(1-2), 70–85.

Aboitiz, F., García, R., 1997. The evolutionary origin of the language areas in the human brain. A neuroanatomical perspective. Brain Research Reviews, 25(3), 381-396.

Abramson, J. Z., Hernández-Lloreda, M. V., García, L., Colmenares, F., Aboitiz, F., Call, J., 2018. Imitation of novel conspecific and human speech sounds in the killer whale (Orcinus orca). Proc. R. Soc. B, 285(1871), 20172171.

Arbib, M. A., 2005. From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics. Behavioral and brain sciences, 28(2), 105-124.

Arbib, M. A., 2012. How the brain got language: The mirror system hypothesis (Vol. 16). Oxford University Press.

Arcaro, M. J., Livingstone, M. S., 2017. Retinotopic Organization of Scene Areas in Macaque Inferior Temporal Cortex. The Journal of Neuroscience, 37(31), 7373–7389.

Archila-Meléndez, M. E., Valente, G., Correia, J. M., Rouhl, R. P., van Kranen-Mastenbroek, V. H., & 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*(2).

Arnal, L. H., & Giraud, A. L. (2012). Cortical oscillations and sensory predictions. *Trends in cognitive sciences*, *16*(7), 390-398. https://doi.org/10.1016/j.tics.2012.05.003

Arnal, L. H., Morillon, B., Kell, C. A., & Giraud, A. L. (2009). Dual neural routing of visual facilitation in speech processing. *Journal of Neuroscience*, *29*(43), 13445-13453. *https://doi.org/10.1093/cercor/bhu103*

Arnstein, D., Cui, F., Keysers, C., Maurits, N. M., & Gazzola, V. (2011). μ-suppression during action observation and execution correlates with BOLD in dorsal premotor, inferior parietal, and SI cortices. *Journal of Neuroscience*, *31*(40), 14243-14249.

Assaneo, M. F., & Poeppel, D. (2018). The coupling between auditory and motor cortices is rate-restricted: Evidence for an intrinsic speech-motor rhythm. *Science advances*, *4*(2), 3842.

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)

Baart, M. (2016). Quantifying lip-read-induced suppression and facilitation of the auditory N1 and P2 reveals peak enhancements and delays. *Psychophysiology*, *53*(9), 1295-1306.

Barenholtz, E., Mavica, L., Lewkowicz, D. J., 2016. Language familiarity modulates relative attention to the eyes and mouth of a talker. Cognition, 147, 100-105.

Bartoli, E., Maffongelli, L., Campus, C., & D'Ausilio, A. (2016). Beta rhythm modulation by speech sounds: somatotopic mapping in somatosensory cortex. *Scientific reports*, *6*, 31182.

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

Bear, H. L., Harvey, R., 2017. Phoneme-to-viseme mappings: the good, the bad, and the ugly. Speech Communication, 95, 40-67.

Beauchamp, M. S., 2015. The social mysteries of the superior temporal sulcus. Trends in cognitive sciences, 19(9), 489-490.

Bedny, M., Richardson, H., Saxe, R., 2015. "Visual" cortex responds to spoken language in blind children. Journal of Neuroscience, 35(33), 11674-11681.

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

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

Besle, J., Fischer, C., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., Giard, M. H., 2008. Visual activation and audiovisual interactions in the auditory cortex during speech perception: intracranial recordings in humans. Journal of Neuroscience, 28(52), 14301-14310.

Binder, J. R., Desai, R. H., 2011. The neurobiology of semantic memory. Trends in cognitive sciences, 15(11), 527-536.

Binder, J. R., Desai, R. H., Graves, W. W., Conant, L. L., 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cerebral Cortex, 19(12), 2767-2796.

Bimbi, M., Festante, F., Coudé, G., Vanderwert, R. E., Fox, N. A., & Ferrari, P. F. (2018). Simultaneous scalp recorded EEG and local field potentials from monkey ventral premotor cortex during action observation and execution reveals the contribution of mirror and motor neurons to the mu-rhythm. *NeuroImage*, *175*, 22-31.

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*

Blank, H., von Kriegstein, K., 2013. Mechanisms of enhancing visual–speech recognition by prior auditory information. NeuroImage, 65, 109-118.

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*

Bowers, A., Saltuklaroglu, T., Jenson, D., Harkrider, A., & Thornton, D. (2019). Power and phase coherence in sensorimotor mu and temporal lobe alpha components during covert and overt syllable production. *Experimental brain research*, *237*(3), 705-721.

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*

Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., Pylkkänen, L., 2012. Syntactic structure building in the anterior temporal lobe during natural story listening. Brain and Language, 120(2), 163–173. *http://doi.org/10.1016/j.bandl.2010.04.002*

Bruderer, A. G., Danielson, D. K., Kandhadai, P., Werker, J. F., 2015. Sensorimotor influences on speech perception in infancy. Proceedings of the National Academy of Sciences, 112(44), 13531-13536.

Calvert, G. A., Campbell, R., 2003. Reading speech from still and moving faces: the neural substrates of visible speech. Journal of cognitive neuroscience, 15(1), 57-70.

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*

Catani, M., Bambini, V., 2014. A model for Social Communication And Language Evolution and Development (SCALED). Current Opinion in Neurobiology, 28, 165–171. *http://doi.org/10.1016/j.conb.2014.07.018*

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS computational biology*, *5*(7), e1000436.

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 Weren't. *Journal of Neuroscience*, *36*(32), 8305–8316. *http://doi.org/10.1523/jneurosci.1125-16.2016*

Cheung, C., Hamilton, L. S., Johnson, K., & Chang, E. F. (2016). The auditory representation of speech sounds in human motor cortex. *Elife*, *5*, e12577.

Cogan, G. B. (2016). I see what your're saying: The motor cortex in the brain tracks lip movements to help with speech perception. eLife 5:e17693.

Cogan, G. B., Thesen, T., Carlson, C., Doyle, W., Devinsky, O., & Pesaran, B. (2014). Sensory–motor transformations for speech occur bilaterally. *Nature*, *507*(7490), 94.

Cook, R., Bird, G., Catmur, C., Press, C., Heyes, C., 2014. Mirror neurons: from origin to function. Behavioral and Brain Sciences, 37(2), 177-192.

Correia, J. M., Jansma, B. M., & Bonte, M. (2015). Decoding articulatory features from fMRI responses in dorsal speech regions. *Journal of Neuroscience*, *35*(45), 15015-15025.

Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *Journal of Neuroscience*, *35*(42), 14195-14204.

Cuellar, M., Bowers, A., Harkrider, A. W., Wilson, M., & Saltuklaroglu, T. (2012). Mu suppression as an index of sensorimotor contributions to speech processing: evidence from continuous EEG signals. *International Journal of Psychophysiology*, *85*(2), 242-248.

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*

Dichter, B. K., Breshears, J. D., Leonard, M. K., Chang, E. F., 2018. The Control of Vocal Pitch in Human Laryngeal Motor Cortex. Cell, 174(1), 21–31.e9. *http://doi.org/10.1016/j.cell.2018.05.016*

Diehl, M. M., Romanski, L. M., 2012. Representation and Integration of Faces and Vocalizations in the Primate Ventral Prefrontal Cortex. In Integrating Face and Voice in Person Perception (pp. 45–69). Springer New York. *http://doi.org/10.1007/978-1-4614-3585-3_3*

Ding, N., Melloni, L., Yang, A., Wang, Y., Zhang, W., & Poeppel, D. (2017). Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). *Frontiers in human neuroscience*, *11*, 481.

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*

Dubois, C., Otzenberger, H., Gounot, D., Sock, R., Metz-Lutz, M. N., 2012. Visemic processing in audiovisual discrimination of natural speech: a simultaneous fMRI–EEG study. Neuropsychologia, 50(7), 1316-1326.

Duffau, H., 2018. The error of Broca: From the traditional localizationist concept to a connectomal anatomy of human brain. Journal of Chemical Neuroanatomy, 89, 73–81. *http://doi.org/10.1016/j.jchemneu.2017.04.003*

Dunn, J. C., Smaers, J. B., 2018. Neural Correlates of Vocal Repertoire in Primates. Frontiers in Neuroscience, 12, 534.

Eichert, N., Verhagen, L., Folloni, D., Jbabdi, S., Khrapitchev, A. A., Sibson, N. R., … Mars, R. B., 2018. What is special about the human arcuate fasciculus? Lateralization projections, and expansion. Cortex. *http://doi.org/10.1016/j.cortex.2018.05.005*

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*(2), 175-191.

Fauvelle C. (1886). Du langage articulé. In: Bulletins de la Société d'anthropologie de Paris, III° Série. Tome 9,. pp. 636-653

Ferpozzi, V., Fornia, L., Montagna, M., Siodambro, C., Castellano, A., Borroni, P., … Cerri, G., 2018. Broca's Area as a Pre-articulatory Phonetic Encoder: Gating the Motor Program. Frontiers in Human Neuroscience, 12. *http://doi.org/10.3389/fnhum.2018.00064*

Fleagle, J. G., 2013. Primate adaptation and evolution. Academic Press.

Flinker, A., Korzeniewska, A., Shestyuk, A. Y., Franaszczuk, P. J., Dronkers, N. F., Knight, R. T. & Crone, N. E. (2015). Redefining the role of Broca's area in speech. Proceedings of the National Academy of Sciences, 112(9), 2871–2875. *http://doi.org/10.1073/pnas.1414491112*

Fodor, J. A. (1975). *The language of thought* (Vol. 5). Harvard university press.

Fox, N. A., Bakermans-Kranenburg, M. J., Yoo, K. H., Bowman, L. C., Cannon, E. N., Vanderwert, R. E., ... & van IJzendoorn, M. H. (2016). Assessing human mirror activity with EEG mu rhythm: A meta-analysis. *Psychological Bulletin*, *142*(3), 291.

Freud, E., Plaut, D. C., Behrmann, M., 2016. 'What' Is Happening in the Dorsal Visual Pathway. Trends in Cognitive Sciences, 20(10), 773–784. *http://doi.org/10.1016/j.tics.2016.08.003*

Fridriksson, J., Kjartansson, O., Morgan, P. S., Hjaltason, H., Magnusdottir, S., Bonilha, L., & Rorden, C. (2010). Impaired speech repetition and left parietal lobe damage. *Journal of Neuroscience*, *30*(33), 11057-11061.

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*

Friederici, A. D., 2011. The Brain Basis of Language Processing: From Structure to Function. Physiological Reviews, 91(4), 1357–1392. *http://doi.org/10.1152/physrev.00006.2011*

Friederici, A. D., 2016. Evolution of the neural language network. Psychonomic Bulletin & Review, 24(1), 41–47. *http://doi.org/10.3758/s13423-016-1090-x*

Friederici, A. D., Bahlmann, J., Heim, S., Schubotz, R. I., Anwander, A., 2006. The brain differentiates human and non-human grammars: functional localization and structural connectivity. Proceedings of the National Academy of Sciences, 103(7), 2458-2463.

Friedrich, P., Anderson, C., Schmitz, J., Schlüter, C., Lor, S., Stacho, M., ... & Ocklenburg, S. (2019). Fundamental or forgotten? Is Pierre Paul Broca still relevant in modern neuroscience?. *Laterality: Asymmetries of Body, Brain and Cognition*, *24*(2), 125-138.

Gajardo-Vidal, A., Lorca-Puls, D. L., Hope, T. M., Parker Jones, O., Seghier, M. L., Prejawa, S., ... & Price, C. J. (2018). How right hemisphere damage after stroke can impair speech comprehension. *Brain*, *141*(12), 3389-3404.

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*

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

Garagnani, M., & Pulvermüller, F. (2013). Neuronal correlates of decisions to speak and act: spontaneous emergence and dynamic topographies in a computational model of frontal and temporal areas. *Brain and language*, *127*(1), 75-85.

García, R. R., Zamorano,. F., Aboitiz, F., 2014. From imitation to meaning: circuit plasticity and the acquisition of a conventionalized semantics. Front. Hum. Neurosci. 8:605. doi: 10.3389/fnhum.2014.00605

Ghazanfar, A. A., Takahashi, D. Y., Mathur, N., Fitch, W. T., 2012. Cineradiography of monkey lip-smacking reveals putative precursors of speech dynamics. Current Biology, 22(13), 1176-1182.

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*

Golumbic, E. Z., Cogan, G. B., Schroeder, C. E., & Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". *Journal of Neuroscience*, *33*(4), 1417-1426.

Goodale, M. A., Milner, A. D., 1992. Separate visual pathways for perception and action. Trends in Neurosciences, 15(1), 20–25. *http://doi.org/10.1016/0166-2236(92)90344-8*

Hage, S. R., Nieder, A., 2015. Audio-Vocal Interaction in Single Neurons of the Monkey Ventrolateral Prefrontal Cortex. Journal of Neuroscience, 35(18), 7030–7040. *http://doi.org/10.1523/jneurosci.2371-14.2015*

Hagoort, P. (2016). MUC (Memory, Unification, Control): A model on the neurobiology of language beyond single word processing. In *Neurobiology of language* (pp. 339-347). Academic Press.

Hagoort, P. (2017). The core and beyond in the language-ready brain. *Neuroscience & Biobehavioral Reviews*, *81*, 194-204.

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*

Hickok, G., 2016. A cortical circuit for voluntary laryngeal control: Implications for the evolution language. Psychonomic Bulletin & Review, 24(1), 56–63. *http://doi.org/10.3758/s13423-016-1100-z*

Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition, 92(1-2), 67–99. *http://doi.org/10.1016/j.cognition.2003.10.011*

Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nature Reviews Neuroscience, 8(5), 393–402. *http://doi.org/10.1038/nrn2113*

Hillairet de Boisferon, A., Tift, A. H., Minar, N. J., Lewkowicz, D. J., 2017. Selective attention to a talker's mouth in infancy: role of audiovisual temporal synchrony and linguistic experience. Developmental science, 20(3), e12381.

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)

Hobson, H. M., & Bishop, D. V. (2016). Mu suppression–a good measure of the human mirror neuron system?. *cortex*, *82*, 290-310.

Iacoboni, M. (2005). Neural mechanisms of imitation. *Current opinion in neurobiology*, *15*(6), 632-637.

Iacoboni, M. & Dapretto, M. (2006). The mirror neuron system and the consequences of its dysfunction. *Nature Reviews Neuroscience*, *7*(12), 942.

Iacoboni, M., Hurley, S. & Chater, N. (2005). Perspectives on imitation: From neuroscience to Social Science. Perspectives on imitation: From neuroscience to Social Science, 1.

Iacoboni, M., Koski, L. M., Brass, M., Bekkering, H., Woods, R. P., Dubeau, M. C., ... Rizzolatti, G. (2001). Reafferent copies of imitated actions in the right superior temporal cortex. *Proceedings of the national academy of sciences*, *98*(24), 13995-13999.

Irwin, J. R., & Brancazio, L. (2014). Seeing to hear? Patterns of gaze to speaking faces in children with autism spectrum disorders. *Frontiers in psychology*, *5*, 397.

Irwin, J., Preston, J., Brancazio, L., D’angelo, M., & Turcios, J. (2015). Development of an audiovisual speech perception app for children with autism spectrum disorders. *Clinical linguistics & phonetics*, *29*(1), 76-83.

Irwin, J. R., Tornatore, L. A., Brancazio, L., & Whalen, D. H. (2011). Can children with autism spectrum disorders “hear” a speaking face?. *Child development*, *82*(5), 1397-1403.

Janssens, T., Zhu, Q., Popivanov, I. D. & Vanduffel, W. (2014). Probabilistic and Single-Subject Retinotopic Maps Reveal the Topographic Organization of Face Patches in the Macaque Cortex. *Journal of Neuroscience*, *34*(31), 10156–10167. http://doi.org/10.1523/jneurosci.2914-13.2013

Jenson, D., Reilly, K. J., Harkrider, A. W., Thornton, D., & Saltuklaroglu, T. (2018). Trait related sensorimotor deficits in people who stutter: An EEG investigation of µ rhythm dynamics during spontaneous fluency. *NeuroImage: Clinical*, *19*, 690-702.

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

Jürgens, U. (2009). The Neural Control of Vocalization in Mammals: A Review. *Journal of Voice*, *23*(1), 1–10. http://doi.org/10.1016/j.jvoice.2007.07.005

Kaas, J. H. & Hackett, T. A. (1999). What and where processing in auditory cortex. *Nature Neuroscience*, *2*(12), 1045–1047. http://doi.org/10.1038/15967

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

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

Keysers, C., Kohler, E., Umiltà, M. A., Nanetti, L., Fogassi, L., Gallese, V. (2003). Audiovisual mirror neurons and action recognition. *Experimental brain research*, *153*(4), 628-636.

Keysers, C., Perrett, D. I. (2004). Demystifying social cognition: a Hebbian perspective. *Trends in cognitive sciences*, *8*(11), 501-507.

Kilner, J. M., Friston, K. J., Frith, C. D. (2007). Predictive coding: an account of the mirror neuron system. *Cognitive processing*, *8*(3), 159-166.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science*, *218*(4577), 1138-1141.

Kuhl, P. K., & Meltzoff, A. N. (1996). Infant vocalizations in response to speech: vocal imitation and developmental change. Journal of the Acoustic Society of America, 100(401):2425-2438.

Kumar, V., Croxson, P. L., Simonyan, K. (2016). Structural Organization of the Laryngeal Motor Cortical Network and Its Implication for Evolution of Speech Production. *Journal of Neuroscience*, *36*(15), 4170–4181. http://doi.org/10.1523/jneurosci.3914-15.2016

Lane, H. (1965). The motor theory of speech perception: A critical review. *Psychological Review*, *72*(4), 275.

Lane, C., Kanjlia, S., Omaki, A., Bedny, M. (2015). “Visual” cortex of congenitally blind adults responds to syntactic movement. *Journal of Neuroscience*, *35*(37), 12859-12868.

Langacker, R. W. (2012). Essentials of cognitive grammar. Oxford University Press.

Legerstee, M. (1990). Infants use multimodal information to imitate speech sounds. *Infant behavior and development*, *13*(3), 343-354.

Leighton, G. M. (2017). Cooperative breeding influences the number and type of vocalizations in avian lineages. *Proc. R. Soc. B*, *284*(1868), 20171508.

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

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

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

Lichtheim, L. (1885). On aphasia. *Brain*, *7*, 433-484.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review*, *74*(6), 431.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1-36.

Lieberman, P. (1993). Uniquely human: The evolution of speech, thought, and selfless behavior. Harvard University Press.

Lieberman, P. (2015). The evolution of language. In: Goldstein, S., Princiotta, D., Naglieri, J.A. (Eds.), *Handbook of Intelligence: Evolutionary Theory, Historical Perspective, and Current Concepts*. Springer, New York, pp. 47-64.

Long, M. A., Katlowitz, K. A., Svirsky, M. A., Clary, R. C., Byun, T. M. A., Majaj, N., … Greenlee, J. D. W. (2016). Functional Segregation of Cortical Regions Underlying Speech Timing and Articulation. *Neuron*, *89*(6), 1187–1193. http://doi.org/10.1016/j.neuron.2016.01.032

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

Luo, H., Liu, Z., & Poeppel, D. (2010). Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. *PLoS biology*, *8*(8), e1000445.

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

Massaro, D. W., & Palmer, S. E. (1998). *Perceiving talking faces: From speech perception to a behavioral principle* (Vol. 1). MIT Press.

McGurk, H., MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746.

Messum, P., Howard, I. S., 2015. Creating the cognitive form of phonological units: The speech sound correspondence problem in infancy could be solved by mirrored vocal interactions rather than by imitation. Journal of Phonetics, 53, 125-140.

Michon, M., López, V., Aboitiz, F. (2019). Origin and evolution of human language. In Evolution of the human brain: from mater to mind. *Progress in Brain Research*, chapter 14, vol.251.

Miki, K., Watanabe, S., Kakigi, R., Puce, A., 2004. Magnetoencephalographic study of occipitotemporal activity elicited by viewing mouth movements. Clinical neurophysiology, 115(7), 1559-1574.

Miller, J., Brookie, K., Wales, S., Wallace, S., & Kaup, B. (2018). Embodied cognition: Is activation of the motor cortex essential for understanding action verbs?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(3), 335. http://doi.org/10.1037/xlm0000451.supp

Mitchell, T. V., Letourneau, S. M., Maslin, M. C., 2013. Behavioral and neural evidence of increased attention to the bottom half of the face in deaf signers. Restorative neurology and neuroscience, 31(2), 125-139.

Morrill, G. V., 2012. Type logical grammar: Categorial logic of signs. Springer Science & Business Media.

Mugler, E. M., Tate, M. C., Livescu, K., Templer, J. W., Goldrick, M. A., Slutzky, M. W., 2018. Differential Representation of Articulatory Gestures and Phonemes in Precentral and Inferior Frontal Gyri. Journal of Neuroscience, 38(46), 9803-9813.

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

Neubert, F.-X., Mars, R. B., Thomas, A. G., Sallet, J., Rushworth, M. F. S., 2014. Comparison of Human Ventral Frontal Cortex Areas for Cognitive Control and Language with Areas in Monkey Frontal Cortex. Neuron, 81(3), 700–713. http://doi.org/10.1016/j.neuron.2013.11.012

Nguyen, N., Delvaux, V., 2015. Role of imitation in the emergence of phonological systems. Journal of Phonetics, 53, 46-54.

Nuttall, H. E., Kennedy-Higgins, D., Devlin, J. T., & Adank, P. (2018). Modulation of intra-and inter-hemispheric connectivity between primary and premotor cortex during speech perception. *Brain and language*, *187*, 74-82.

Oller, J. W. (2010). The antithesis of entropy: Biosemiotic communication from genetics to human language with special emphasis on the immune systems. *Entropy*, *12*(4), 631-705.

Ortega, R., López, V., & Aboitiz, F. (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*

Papathanassiou, D., Etard, O., Mellet, E., Zago, L., Mazoyer, B., & Tzourio-Mazoyer, N. (2000). A common language network for comprehension and production: a contribution to the definition of language epicenters with PET. *Neuroimage*, *11*(4), 347-357.

Papoutsi, M., de Zwart, J. A., Jansma, J. M., Pickering, M. J., Bednar, J. A., Horwitz, B., 2009. From Phonemes to Articulatory Codes: An fMRI Study of the Role of Brocas Area in Speech Production. Cerebral Cortex, 19(9), 2156–2165. *http://doi.org/10.1093/cercor/bhn239*

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*

Peelle, J.E. (In press). The neural basis for auditory and audiovisual speech perception. In: *The Routledge Handbook of Phonetics* (Katz and Assmann, eds). Routledge.

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

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of neuroscience methods*, *162*(1-2), 8-13. *https://doi.org/10.1016/j.jneumeth.2006.11.017*

Pérez-Pereira, M., 2006. Language development in blind children. K. Brown (Editorin-Chief), Encyclopedia of Language & Linguistics, 6, 357-361.

Petrides, M., 2005. Lateral prefrontal cortex: architectonic and functional organization. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 360(1456), 781-795.

Petrides, M., 2014. Neuroanatomy of Language Regions of the Human Brain. New York, New York: Academic Press.

Petrides, M., Cadoret, G., Mackey, S., 2005. Orofacial somatomotor responses in the macaque monkey homologue of Broca's area. Nature, 435(7046), 1235.

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*

Pimperton, H., Ralph-Lewis, A., MacSweeney, M., 2017. Speechreading in deaf adults with cochlear implants: Evidence for perceptual compensation. Frontiers in psychology, 8, 106.

Puce, A., Allison, T., Bentin, S., Gore, J. C., McCarthy, G., 1998. Temporal cortex activation in humans viewing eye and mouth movements. Journal of Neuroscience, 18(6), 2188-2199.

Pulvermüller, F., Huss, M., Kherif, F., del Prado Martin, F. M., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences*, *103*(20), 7865-7870.

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*

Pulvermüller, F., & Fadiga, L. (2016). Brain language mechanisms built on action and perception. In *Neurobiology of Language* (pp. 311-324). Academic Press.

Rauschecker, J. P., 2012. Ventral and dorsal streams in the evolution of speech and language. Frontiers in Evolutionary Neuroscience, 4. *http://doi.org/10.3389/fnevo.2012.00007*

Rauschecker, J. P., 2018. Where did language come from? Precursor mechanisms in nonhuman primates. Current Opinion in Behavioral Sciences, 21, 195–204. *http://doi.org/10.1016/j.cobeha.2018.06.003*

Rauschecker, J. P., 2018. Where When, and How: Are they all sensorimotor? Towards a unified view of the dorsal pathway in vision and audition. Cortex, 98, 262–268. *http://doi.org/10.1016/j.cortex.2017.10.020*

Rauschecker, J. P., Scott, S. K., 2009. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nature Neuroscience, 12(6), 718–724. *http://doi.org/10.1038/nn.2331*

Ray, E., & Heyes, C. (2011). Imitation in infancy: the wealth of the stimulus. *Developmental science, 14*(1), 92-105.

Riedel, P., Ragert, P., Schelinski, S., Kiebel, S. J., von Kriegstein, K., 2015. Visual face-movement sensitive cortex is relevant for auditory-only speech recognition. Cortex, 68, 86-99.

Rilling, J. K., Glasser, M. F., Preuss, T. M., Ma, X., Zhao, T., Hu, X., Behrens, T. E. J., 2008. The evolution of the arcuate fasciculus revealed with comparative DTI. Nature Neuroscience, 11(4), 426–428. *http://doi.org/10.1038/nn2072*

Rilling, J., Glasser, M. F., Jbabdi, S., Andersson, J., Preuss, T. M., 2012. Continuity, divergence, and the evolution of brain language pathways. Frontiers in evolutionary neuroscience, 3, 11.

Rizzolatti, G., Arbib, M. A., 1998. Language within our grasp. Trends in neurosciences, 21(5), 188-194.

Rizzolatti, G., Craighero, L., 2004. The mirror-neuron system. Annu. Rev. Neurosci., 27, 169-192.

Rizzolatti, G., Fadiga, L., Gallese, V., Fogassi, L., 1996. Premotor cortex and the recognition of motor actions. Cognitive brain research, 3(2), 131-141.

Röder, B., Stock, O., Bien, S., Neville, H., Rösler, F., 2002. Speech processing activates visual cortex in congenitally blind humans. European Journal of Neuroscience, 16(5), 930-936.

Rolland, A., Herbet, G., & Duffau, H. (2018). Awake surgery for gliomas within the right inferior parietal lobule: new insights into the functional connectivity gained from stimulation mapping and surgical implications. *World neurosurgery*, *112*, e393-e406.

Romanski, L. M., 2007. Representation and Integration of Auditory and Visual Stimuli in the Primate Ventral Lateral Prefrontal Cortex. Cerebral Cortex, 17(suppl 1), i61–i69. *http://doi.org/10.1093/cercor/bhm099*

Romanski, L. M., 2012. Integration of faces and vocalizations in ventral prefrontal cortex: Implications for the evolution of audiovisual speech. Proceedings of the National Academy of Sciences, 109(Supplement_1), 10717–10724. *http://doi.org/10.1073/pnas.1204335109*

Romanski, L. M., Bates, J. F., Goldman-Rakic, P. S., 1999. Auditory belt and parabelt projections to the prefrontal cortex in the Rhesus monkey. The Journal of Comparative Neurology, 403(2), 141–157. *http://doi.org/10.1002/(sici)1096-9861(19990111)403:2<141::aid-cne1>3.0.co;2-v*

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*

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Molholm, S., Javitt, D. C., Foxe, J. J., 2007. Impaired multisensory processing in schizophrenia: deficits in the visual enhancement of speech comprehension under noisy environmental conditions. Schizophrenia research, 97(1-3), 173-183.

Saito, D. N., Yoshimura, K., Kochiyama, T., Okada, T., Honda, M., & Sadato, N. (2005). Cross-modal binding and activated attentional networks during audio-visual speech integration: a functional MRI study. *Cerebral Cortex*, *15*(11), 1750-1760.

Saltuklaroglu, T., Bowers, A., Harkrider, A., Casenhiser, D., Reilly, K., Jenson, D., & Thornton, D. (2018). EEG mu rhythms: Rich sources of sensorimotor information in speech processing. *Brain and language*, *187*, 41-61.

Saltuklaroglu, T., Harkrider, A. W., Thornton, D., Jenson, D., & Kittilstved, T. (2017). EEG Mu (µ) rhythm spectra and oscillatory activity differentiate stuttering from non-stuttering adults. *NeuroImage*, *153*, 232-245.

Saussure De, F. (1916). Cours de linguistique générale. *Paris: PUF*.

Saur, D., Kreher, B. W., Schnell, S., Kummerer, D., Kellmeyer, P., Vry, M.-S., … Weiller, C., 2008. Ventral and dorsal pathways for language. Proceedings of the National Academy of Sciences, 105(46), 18035–18040. *http://doi.org/10.1073/pnas.0805234105*

Schelinski, S., Riedel, P., & von Kriegstein, K. (2014). Visual abilities are important for auditory-only speech recognition: evidence from autism spectrum disorder. *Neuropsychologia*, *65*, 1-11.

Schomers, M. R., Pulvermüller, F., 2016. Is the sensorimotor cortex relevant for speech perception and understanding? An integrative review. Frontiers in human neuroscience, 10, 435.

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*

Shepherd, S. V., Freiwald, W. A., 2018. Functional networks for social communication in the Macaque Monkey. Neuron, 99(2), 413-420.

Sheth, B. R., Young, R., 2016. Two Visual Pathways in Primates Based on Sampling of Space: Exploitation and Exploration of Visual Information. Frontiers in Integrative Neuroscience, 10. *http://doi.org/10.3389/fnint.2016.00037*

Silson, E. H., Reynolds, R. C., Kravitz, D. J., Baker, C. I., 2018. Differential Sampling of Visual Space in Ventral and Dorsal Early Visual Cortex. The Journal of Neuroscience, 38(9), 2294–2303. *http://doi.org/10.1523/jneurosci.2717-17.2018*

Skeide, M. A., Friederici, A. D., 2016. The ontogeny of the cortical language network. Nature Reviews Neuroscience, 17(5), 323–332. http://doi.org/10.1038/nrn.2016.23

Skipper, J. I., Devlin, J. T., Lametti, D. R., 2017. The hearing ear is always found close to the speaking tongue: Review of the role of the motor system in speech perception. Brain and language, 164, 77-105.

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

Sugihara, T., Diltz, M. D., Averbeck, B. B., Romanski, L. M., 2006. Integration of Auditory and Visual Communication Information in the Primate Ventrolateral Prefrontal Cortex. Journal of Neuroscience, 26(43), 11138–11147. http://doi.org/10.1523/jneurosci.3550-06.2006

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

Swaminathan, S., MacSweeney, M., Boyles, R., Waters, D., Watkins, K. E., & Möttönen, R. (2013). Motor excitability during visual perception of known and unknown spoken languages. *Brain and language*, *126*(1), 1-7.

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

ten Oever, S., Schroeder, C. E., Poeppel, D., van Atteveldt, N., & Zion-Golumbic, E. (2014). Rhythmicity and cross-modal temporal cues facilitate detection. *Neuropsychologia*, *63*, 43-50.

Thornton, D., Harkrider, A. W., Jenson, D., & Saltuklaroglu, T. (2018). Sensorimotor activity measured via oscillations of EEG mu rhythms in speech and non-speech discrimination tasks with and without segmentation demands. *Brain and language*, *187*, 62-73.

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

Trevarthen, C. B., 1968. Two mechanisms of vision in primates. Psychologische Forschung, 31(4), 299–337. http://doi.org/10.1007/bf00422717

Tsang, T., Atagi, N., Johnson, S. P., 2018. Selective attention to the mouth is associated with expressive language skills in monolingual and bilingual infants. Journal of experimental child psychology, 169, 93-109.

Turner, A. C., McIntosh, D. N., & Moody, E. J. (2015). Don’t listen with your mouth full: the role of facial motor action in visual speech perception. *Language and speech*, *58*(2), 267-278.

Tyler, L. K., Marslen-Wilson, W. D., Randall, B., Wright, P., Devereux, B. J., Zhuang, J., … Stamatakis, E. A., 2011. Left inferior frontal cortex and syntax: function structure and behaviour in patients with left hemisphere damage. Brain, 134(2), 415–431. *http://doi.org/10.1093/brain/awq369*

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*

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*

Venezia, J. H., Fillmore, P., Matchin, W., Isenberg, A. L., Hickok, G., & Fridriksson, J. (2016). Perception drives production across sensory modalities: A network for sensorimotor integration of visual speech. *NeuroImage*, *126*, 196-207.

Vigliocco, G., Perniss, P., & Vinson, D. (2014). Language as a multimodal phenomenon: implications for language learning, processing and evolution.

Vilasboas, T., Herbet, G., & Duffau, H. (2017). Challenging the myth of right nondominant hemisphere: lessons from corticosubcortical stimulation mapping in awake surgery and surgical implications. *World neurosurgery*, *103*, 449-456.

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*

Watkins, K. E., Shakespeare, T. J., O'Donoghue, M. C., Alexander, I., Ragge, N., Cowey, A., Bridge, H., 2013. Early auditory processing in area V5/MT+ of the congenitally blind brain. Journal of Neuroscience, 33(46), 18242-18246.

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*

Williams, J. H., Massaro, D. W., Peel, N. J., Bosseler, A., & Suddendorf, T. (2004). Visual–auditory integration during speech imitation in autism. *Research in developmental disabilities*, *25*(6), 559-575.

Wilson, S. M., & Iacoboni, M. (2006). Neural responses to non-native phonemes varying in producibility: evidence for the sensorimotor nature of speech perception. *Neuroimage*, *33*(1), 316-325.

Wilson, S. M., Galantucci, S., Tartaglia, M. C., Rising, K., Patterson, D. K., Henry, M. L., … Gorno-Tempini, M. L., 2011. Syntactic Processing Depends on Dorsal Language Tracts. Neuron, 72(2), 397–403. *http://doi.org/10.1016/j.neuron.2011.09.014*

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature neuroscience*, *7*(7), 701.

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*

Young, G. S., Merin, N., Rogers, S. J., & Ozonoff, S. (2009). Gaze behavior and affect at 6 months: predicting clinical outcomes and language development in typically developing infants and infants at risk for autism. *Developmental science*, *12*(5), 798-814.