UNIVERSITY OF GENOVA

PHD PROGRAM IN BIOENGINEERING AND ROBOTICS

Sensorimotor Processes in Speech Listening and Speech-based Interaction

by

Sankar Mukherjee

Thesis submitted for the degree of *Doctor of Philosophy* (31° cycle)

December 2018

Prof. Alessandro D'Ausilio, Prof. Luciano Fadiga, Dr. Leonardo Badino — Supervisor
Prof. Giorgio Cannata — Head of the PhD program

Thesis Jury:
Prof. Matteo Candidi, *University* — Università "La Sapienza" Roma
Prof. Chiara Begliomini, *University* — Università di Padova

Dibris

Department of Informatics, Bioengineering, Robotics and Systems Engineering

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. Part of the text in this thesis comes from either manuscripts that are already published or submitted to journals and conferences. This dissertation contains fewer than 40,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 50 figures.

Sankar Mukherjee
February 2019

Acknowledgements

First and foremost I would like to express my deepest gratitude to my revered supervisor Prof. Alessandro D'Ausilio for his invaluable guidance and encouragement throughout my work. His guidance and support is far beyond duty. His constant motivation, support and infectious enthusiasm have guided me towards the successful completion of my work. My interactions with him have been of immense help in defining my research goals and in identifying ways to achieve them. His encouraging words have often pushed me to put in my best possible efforts. Above all, the complete belief that he has entrusted upon me and has instilled a great sense of confidence and purpose in my mind, which I am sure, will stand me in good stead throughout my career.

It gives me immense pleasure to thank the head of the department Prof. Luciano Fadiga for extending me all the possible facilities to carry out the research work. My sincere thanks to Dr. Leonardo Badino and Prof. Noël Nguyen for their valuable suggestions during my research.

Lastly, during this period of my postgraduate study there are lots of people whose guidance, support, encouragement and sacrifice has made me indebted for my whole life. I wanna thank them all. But not individually. Because that will take one more page. But you all know who you are and I thank you. I thank you all.

Abstract

The thesis deals with two extreme end of speech perception in cognitive neuroscience. On its one end it deals with single isolated person brain responses to acoustic stimulus and missing articulatory cues, and on the other end it explores the neural mechanisms emerging while speech is embedded in a true conversational interaction. Studying these two extremities requires the use of relatively different methodological approaches. In fact, the first approach has seen the consolidation of a wide variety of experimental designs and analytical methods. Otherwise, the investigation of speech brain processes during a conversation is still in its early infancy and several technical and methodological challenges still needs to be solved. In the present thesis, I will present one EEG study using a classical attentive speech listening task, analyzed by using recent methodological advancement explicitly looking at the neural entrainment to speech oscillatory properties. Then, I will report on the work I did to design a robust speech-based interactive task, to extract acoustic and articulatory indexes of interaction, as well as the neural EEG correlates of its word-level dynamics. All in all, this work suggests that motor processes play a critical role both in attentive speech listening and in guiding mutual speech accomodation. In fact, the motor system, on one hand reconstruct information that are missing in the sensory domain and on the other hand drives our implicit tendency to adapt our speech production to the conversational partner and the interactive dynamics.

Table of contents

Chapter 1

Introduction

1.1 Classical Model of Language in the Brain

In 1861, Paul Broca began to study brain lesion patients affected by a disturbance to language. Specifically, his landmark patient - Leborgne - maintained near normal comprehension abilities while production was virtually absent. Broca later collected more cases, leading him to believe that specific linguistic difficulties were associated with a damage in specific parts of the cortex and hence these regions were crucial for speech production. The region identified in this manner, located in the posterior part of the inferior frontal gyrus, was later called Broca's area (Broca (1861)). Inspired by these early studies, a German doctor and anatomist named Carl Wernicke suggested that another region of the brain was linked to linguistic understanding Figure 1.1). This region too, located in the posterior part of the superior temporal gyrus, was later named after him as Wernicke's area (Wernicke (1974)).

About 100 years later the American behavioral neurologist Norman Geschwind revised and updated the neurological model of language processing known as Wernicke–Geschwind model or Classical model of language in the brain (Geschwind (1970)). This model, by revising all available neuropsychological data, crystallized the idea that the language faculty consists of two basic functions: comprehension and production. For comprehension, the speech sound travels through the primary auditory cortex to reach Wernicke's area where word and sentence meaning is extracted. For production, linguistic concepts are generated at the interface between Wernicke's area and the inferior parietal lobe (often called the Geshwind territory; Catani and ffytche (2005)) and then sent to Broca's area which holds representations required for articulating words. Finally, in order to articulate speech sounds, speech instructions are sent form Broca's area to the facial area of the motor cortex where it relays information to facial muscles. What critically changes from previous conceptions is

that here we have a fundamentally connectionist model. In fact, linguistic functions may also reside in the computations allowed by specific anatomical connection between brain areas. For instance, the arcuate fasciculus, by connecting Wernicke's and Broca's areas allow the interface between the brain centers for production and comprehension. Lesion of the arcuate fasciculus produces the very specific pattern of symptoms, called conduction aphasia.

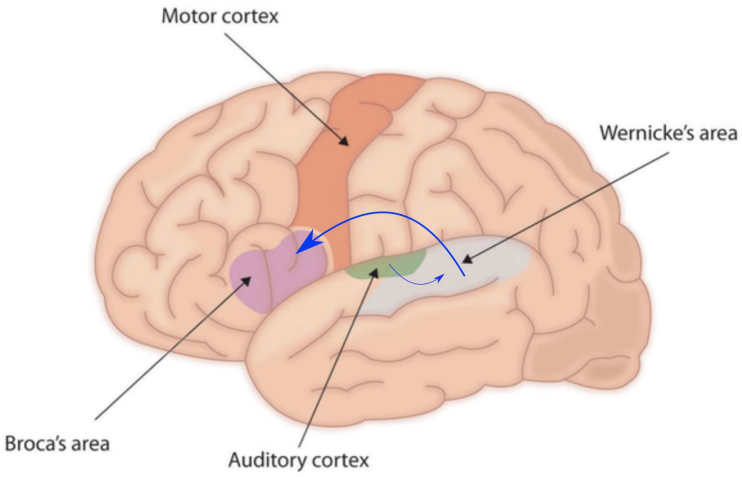

Figure 1.1 *Regions of the brain involved in speech and language processing. The blue arrow line represents interaction between auditory cortex with Wernicke's area during language comprehension and Wernicke's area with Broca's areas during language production.*

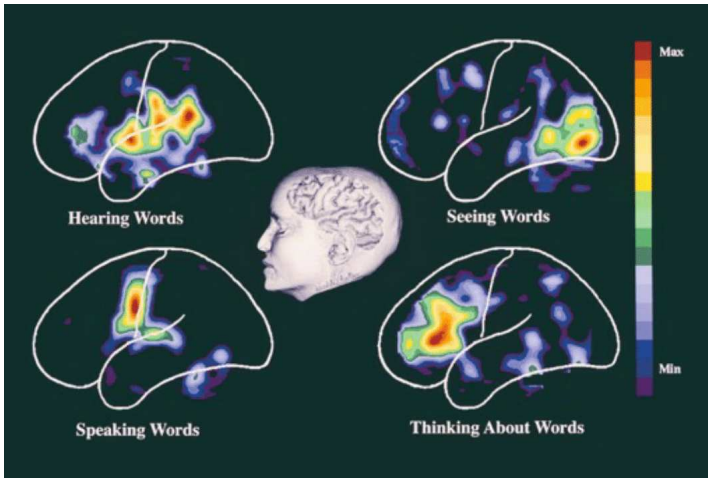

Figure 1.2 *The figure shows PET imaging of brain activities during linguistic tasks. Different brain areas light up when we hear, see, speak and think about words. This reveals different regions of the brain involved during speech production and comprehension. Image adapted from: Posner and Raichle (1994)*

Several years later, the pioneering work of Stephen E. Petersen and colleagues opened the era of brain imaging with positron emission tomography(PET). Starting from 1988 they showed that also in the healthy brain the same regions were recruited during speech production and comprehension (Petersen and Fiez (1993)). This was essentially confirming and reinforcing all indirect claims made in decades of neuropsychological studies on patients. As clearly visible in a landmark image presented as Figure 1.2, different brain areas were activated when subjects heard, read, spoke or thought about words.

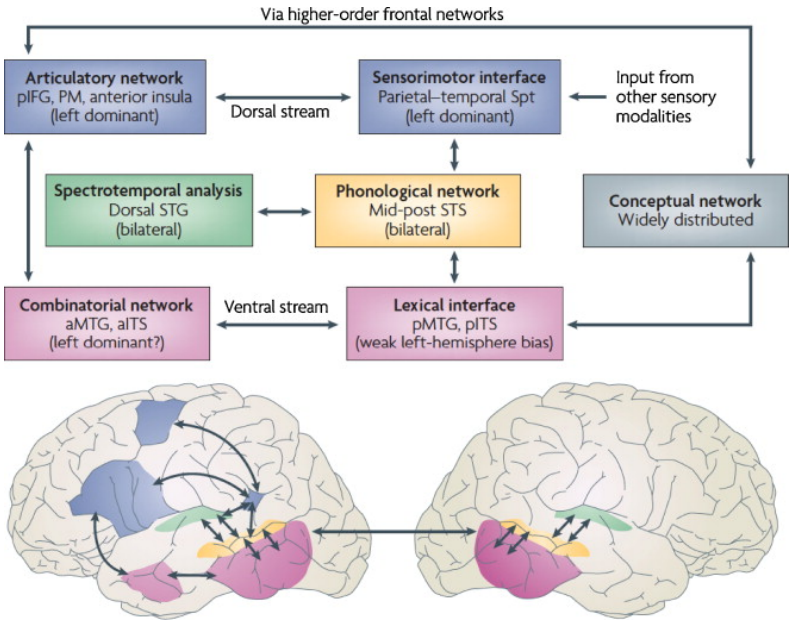

Figure 1.3 *The dual stream model of speech processing (Hickok and Poeppel (2007)). The dual-stream model states that there exists two functionally distinct neural networks for language and speech processing. The first is recruited in sensory/phonological and conceptual-semantic systems processes while the second in sensory/phonological and motor-articulatory ones. Image adapted from: Hickok and Poeppel (2007)*

Much closer to our times, the accumulation of an important number of neuroimaging studies allowed a further revision of the Classical model. In fact, in 2007, the dual stream model of speech processing was introduced (Hickok and Poeppel (2007); Figure 1.3). This model suggests that the early stages of speech processing occur bilaterally in the temporal lobe. In this regards, two temporal areas play a central role: the superior temporal gyrus (STG) (green) and superior temporal sulcus (STS) (yellow). STG is thought to be involved in the spectrotemporal analysis, whereas STS is implicated in phonological-level processes. Then, speech processing diverges into two broad streams: a ventral stream and a dorsal stream. The first, projecting to the middle temporal gyrus and later in the anterior temporal

cortex. The second, to the parieto-temporal junction which in turn would send projections to the anterior language node, in the frontal cortex. The first, is believed to be essential in speech comprehension whereas the second, in speech repetition. Specifically, lexical access and combinatorial processes (pink) is supported by the ventral stream. Whereas sensory–motor integration (i.e. the transformation from sensory-based auditory representations to motor-based production representations) is supported by the left dominant dorsal stream. The conceptual network (gray box) is distributed across several regions of the cortex. These include the superior temporal sulcus (STS), inferior temporal sulcus (ITS), Sylvian parietal–temporal junction (spt), superior temporal gyrus (STG), premotor (PM), inferior frontal gyrus (IFG) and middle temporal gyrus (MTG). The dual stream model treats speech processing in the brain as two separate functions, one involved with the acoustic-phonetic structure of the speech signal and the other involved with its articulatory transformation.

This kind of modeling has very important heuristic value in systematizing knowledge accumulated in 160 years of investigations upon the neurobiology of language. However, they are fundamentally inspired by brain lesion studies and reinforced by the now classic subtraction method in neuroimaging research. At the same time, we have to acknowledge the fact that the low temporal resolution of functional magnetic resonance imaging (fMRI) and the inherent properties of the hemodynamic signal do not allow us to explain the necessarily fast network dynamics that needs to be instantiated by this network.

Another techniques such as EEG (electroencephalography) and MEG (Magnetoelectroencephalography) directly record the electrical or magnetic activity of the brain. This are noninvasive technique with the electrodes placed along the scalp, although invasive electrodes are sometimes used such as in electrocorticography (ECoG). M/EEG measures voltage fluctuations resulting from ionic current within the neurons of the brain. In clinical contexts, M/EEG refers to the recording of the brain's spontaneous electrical activity over a period of time, as recorded from multiple electrodes placed on the scalp. In experimental context,generally focus either on event-related potentials or on the spectral content of EEG. The former investigates potential fluctuations time locked to an event like stimulus onset or button press. The latter analyses the type of neural oscillations (popularly called "brain waves") that can be observed in EEG signals in the frequency domain. But as these methods have high temporal resolution they suffer from lower spacial resolution.

A recent study (Gow Jr and Segawa (2009)) combined the high temporal resolution of EEG and MEG with the spatial resolution of fMRI, to tackle this problem. They first reconstruct the neural sources of the M/EEG signals and then by adopting the Granger causality algorithm they examined the functional connectivity between lexical and articulatory side of

speech processing in a phonetic context assimilation task. They found different patterns of causal interaction between the same brain regions during the task. This provides evidence that a distributed dynamic pattern of connections emerge during speech and language processing and one region can act as hub of connections. Most critically, these hubs can provide multiple functional contribution to speech and language processing (Figure 1.4).

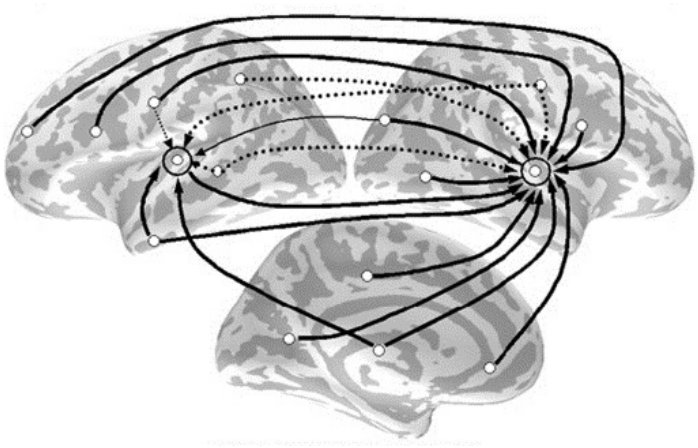

Figure 1.4 *In this figure the white circles represent different regions involved in speech and language processing. Using Granger causality (Gow Jr and Segawa (2009)) able to find different patterns of causal interaction emerged during phonetic context assimilation task. This provides the evidence that a distributed dynamic connections forms during speech and language processing. Image adapted from: Gow Jr and Segawa (2009)*

Another study (Cogan et al. (2014)), used Electrocorticography (ECoG) to record a group of patients' brain activity while they performed a series of tasks such as word listening, repetition and mimicking (movement planning with no actual acoustic feedback). The goal was to evaluate where in the brain it was possible to differentiate purely sensory responses (acoustic), purely motor and mixed sensory-motor. Quite surprisingly, sensory-motor representations that link speech perception and production occurs bilaterally. Furthermore, the authors also found that electrodes over bilateral inferior frontal, inferior parietal, superior temporal, pre-motor, and somatosensory cortices showed sensory-motor neural responses (Figure 1.5). Interestingly, both results do not match more traditional views suggesting a left hemispheric dominance and the presence of sensory-motor activities only in the tempo-parietal junction (Hickok and Poeppel (2007)). As a consequence, this very short review of the available literature, suggests the fact that, despite of the large amount of research using a quite diverse set of methods, (Binder et al. (2009), Price (2010), Vigneau et al. (2006)) the neural basis of language and speech perception still remains unclear.

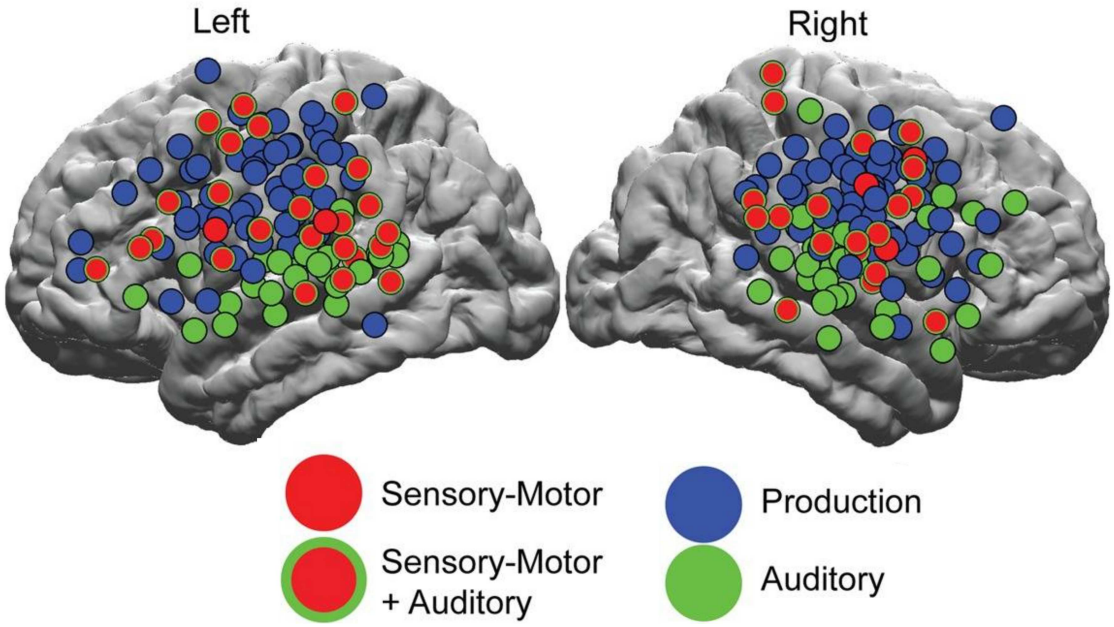

Figure 1.5 *ECoG recording sites of in all subjects from Cogan et al. (2014). The neural responses were typically sensitive to auditory, motor, sensory-motor and combination of sensory-motor and auditory response. This result deviates from the classical left dominant view of speech processing and proves that sensory-motor responses are spread around perisylvian cortices. Image adapted from: Cogan et al. (2014)*

Interestingly, though, in order to move beyond the Classical model and its many contradictions, a new theoretical proposal has emerged in the last 10 years. This proposal do not cancel previous acquisitions but moves from an orthogonal way of investigating the neurobiology of language. Specifically this relatively new trend considers that neural oscillations, more that a specific brain localization, might be essential in making speech segmentation, discrimination and identification possible (Buzsáki and Draguhn (2004), Giraud and Poeppel (2012), Meyer (2018)).

1.2 A new approach to study language and speech processing in the brain

In physics terms a sound contains a set of longitudinal, sinusoidal waves that represent displacement of pressures in a medium (i.e. air). These changes of air pressure constitute a sound waveform which has periodic qualities that can be decomposed in its spectral constituents. This periodicity contains cues which are repeated in predictable intervals.

Intuitively, speech processing requires the segmentation and identification of these features in the acoustic spectrum which convey very specific properties of the sound source (i.e. speech articulation patterns). These features are phonemes or syllables. A series of phonemes constitutes a syllable and a series of syllables constitutes an intonation phrase.

Neural oscillations might be essential in the segmentation and identification of these speech units into three different timescales by aligning their phase and amplitude with the onsets and offsets of syllables, phonemes and intonation phrases (for review, see Giraud and Poeppel (2012); Kösem and Van Wassenhove (2017)). The delta-band oscillations synchronize with intonational phrase boundaries (Bourguignon et al. (2013)), the theta-band oscillations synchronize with syllables (Peelle et al. (2012)) and the gamma-band oscillations synchronize with the phonemes (Di Liberto et al. (2015)). In addition to bottom-up neural responses to stimuli, an independent top-down hierarchical structure plays key role in speech processing. This hierarchical structure, via low frequency neural oscillations, top-down align high frequency neural oscillations through phase–amplitude coupling (Giraud and Poeppel (2012); Gross et al. (2013); Kayser et al. (2015); Riecke et al. (2015)). This amplifies bottom-up information extraction. Therefore, the temporal alignment between different oscillations might be the result of the brain extracting meaningful linguistically distinctive spectral information from speech. Top-down modulation are indexed by delta-theta (Lakatos et al. (2005)) and theta-gamma (Morillon et al. (2012)) frequency coupling. Delta-theta coupling (phase of delta and amplitude of theta) has been proposed in grouping syllables into intonational phrases (Giraud and Poeppel (2012)). Whereas theta-gamma coupling (phase of theta and amplitude of gamma) binds phonemes to syllables (Canolty et al. (2006); Morillon et al. (2012)).

After successful segmentation of speech segments and identification of corresponding phonological percepts, the encoded meaning must be decoded in order to facilitate language comprehension. In this context, there are two psycholinguistics and neurolinguistics accounts for language comprehension. These are two parallel processes, the syntactic processing stream and the predictive processing stream. The syntactic processing stream is thought to be involved in composing complex meaning through chunking of multiple words into syntactic phrases, storing them in verbal working memory, retrieving information from working memory as well as long-term memory and finally imposing syntactic knowledge, thereby creating meaningful relationships within which the sentence is used. Recent work (Ding et al. (2015); Ding and He (2016); Meyer et al. (2017); Bonhage et al. (2017)) suggests that delta-band cycles might involved in grouping words into syntactic phrases. Increased alpha-band power modulations has been found to be associated with storage of syntactic phrases in

verbal working memory (Haarmann and Cameron (2005); Weiss et al. (2005); Meyer et al. (2013); Bonhage et al. (2017)). Coherence and power in theta-band oscillations was found to be increased during language comprehension (Bastiaansen et al. (2002); Bastiaansen et al. (2008); Bastiaansen and Hagoort (2006)). The predictive processing stream deals with the lexical-semantic prediction of upcoming words based on the prior frequency of occurrences stored in long-term memory in a probabilistic fashion (Marslen-Wilson (1973); Kutas and Hillyard (1980); Ehrlich and Rayner (1981); Petten (1993); Hagoort et al. (2004); Kutas and Federmeier (2011)). This cannot be achieved from acoustic cues alone but from the lexical-semantic knowledge. Recent works (Lewis and Bastiaansen (2015); Lewis et al. (2016)) suggests that the beta-band and gamma-band may be involved in the top-down semantic-driven prediction of upcoming words which correctly reflects the cumulative build-up of meaning derived from prior word sequence. Finally the beta–gamma interplay during language comprehension is thought to be the all-purpose theory of cognition in the predictive coding framework (e.g. Friston (2005)). Figure 1.6 summarize the different ways neural oscillation impacts speech and language processing.

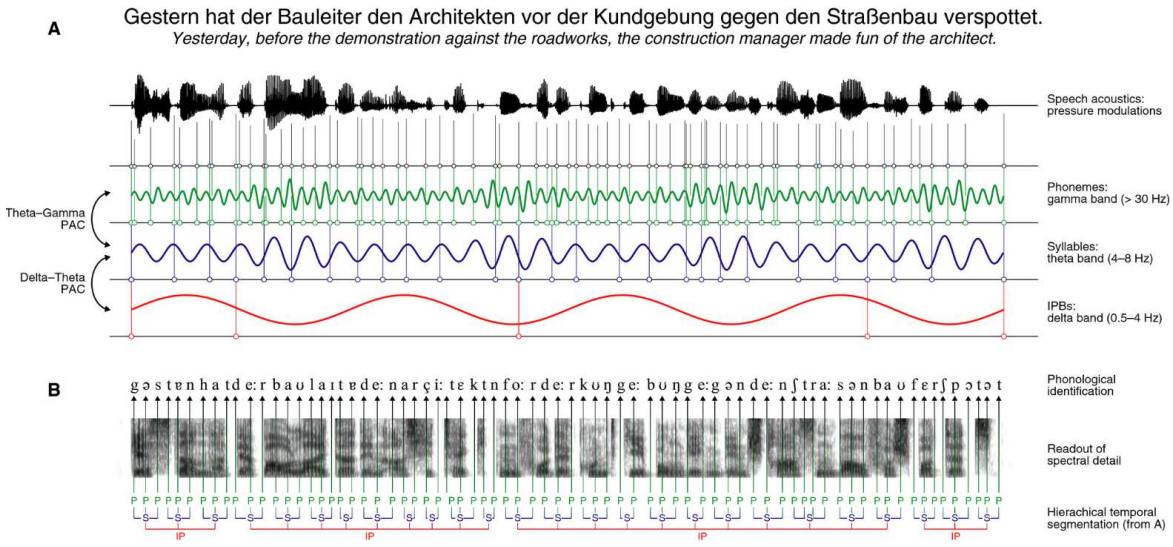

Figure 1.6 *(A) Phonemes (green), syllables (blue) and intonational phrase boundaries (IPB; red) resonates maximally with gamma- (green), theta- (blue) and delta-band oscillations (red). Delta–theta and theta– gamma phase–amplitude coupling (PAC) could serve as binding mechanisms of lower level phonological units into higher level features. For example in (B) phonemes groups into syllables and syllables groups into intonation phrases. This binding mechanisms helps in identification of critical linguistic features which in turn results in speech comprehension. Image adapted from: Meyer (2018)*

1.2.1 Neural Entrainment to Speech Sounds

Verbal interaction is a remarkable part of our behavioural repertoire and our brain is tuned to the decoding of speech from the senses. Neural entrainment is the synchronization of cortical oscillations with different patterns of temporal information in the speech signal. Temporal amplitude envelope of speech is of high importance as it has been shown, in the single neuron studies that neurons in the primary auditory cortex encode speech envelop via phase locked neural firing (Wang et al. (2003)). In the last years cortical entrainment to the temporal envelope of speech has been demonstrated in humans using magnetoencephalography (MEG; Ahissar et al. (2001); Luo and Poeppel (2007)), electroencephalography (EEG; Aiken and Picton (2008)), and electrocorticography (ECoG; Nourski et al. (2009)).

Neural oscillations are thought to play an essential role in shaping speech perception in time (Giraud and Poeppel (2012); Kösem and Van Wassenhove (2017); Meyer (2018); Zoefel et al. (2018)). In fact, it has been shown that neural oscillatory activities in auditory areas are coupled to the rhythmic properties of speech (Ahissar et al. (2001); Luo and Poeppel (2007); Gross et al. (2013); Ding and Simon (2014)). Both the amplitude and the phase of neural oscillations synchronize with external signals and they are defined as amplitude synchronization and phase synchronization respectively. For amplitude synchronization, the phase of theta and delta frequency bands is shown to be synchronized with speech amplitude modulations, whereas only the amplitude of gamma-band is affected by modulations of speech amplitude (for review, see Hyafil et al. (2015)). For phase synchronization, across the gamma, theta and delta frequency bands, the phase of an ongoing oscillation resets for fast modulations of speech amplitude (Gross et al. (2013)). Also, in the rat auditory cortex the phase of local field potentials resets to high amplitude burst of complex acoustic stimuli (Szymanski et al. (2011)).

The strength of oscillatory coupling positively scales with speech intelligibility (Ghitza (2012); Peelle et al. (2012); Ding et al. (2015); Kayser et al. (2015); Riecke et al. (2018)) and is associated with speech comprehension performance (Ahissar et al. (2001); Luo and Poeppel (2007); Peelle et al. (2012)). Investigations on the cocktail party effect (Cherry (1953); Ding and Simon (2012)) show that selective attention translates into increased neural entrainment to the attended acoustic stream (Golumbic et al. (2013a); Kerlin et al. (2010); Golumbic et al. (2012); O'sullivan et al. (2014); Vander Ghinst et al. (2016)), and this is further boosted in the presence of congruent visual cues (Crosse et al. (2015)). Indeed, the oscillatory dynamics within the visual cortex is also coupled to the periodicity of visual cues (Park et al. (2016)). As for speech-brain coupling, cortical visual coupling is affected by top-down modulation (Park et al. (2015), Park et al. (2016); Park et al. (2018a)) and

enhanced when the acoustic signal is degraded (Giordano et al. (2017)). Not surprisingly, speech comprehension mostly benefits from visual cues in suboptimal listening conditions (Sumby and Pollack (1954); Schroeder et al. (2008); Golumbic et al. (2013b)).

As outlined above, the functional relevance of speech-brain entrainment has been associated with the attentional-driven top-down selection of multimodal sensory cues over time (Park et al. (2015)) which eventually aid speech comprehension. Importantly, selective attention exploits knowledge about the periodicity of speech (Poeppel (2003)) and the statistical associations between audio-visual speech signals (Chandrasekaran et al. (2009); Park et al. (2016); But see Schwartz and Savariaux (2014)). It is not clear, however, whether these top-down influences rely on predictive modeling and are therefore capable of reconstructing missing sensory information.

1.2.2 Role of the Motor System in Top-down Modulations

The motor system is not only involved in motor timing of articulatory units , in fact it is also recruited during rhythm perception in passive listening tasks, even when attention is directed away from the auditory stream. In other words, the efferent motor signals that are generated when producing speech sounds are also generated during the passive listening of temporally structured auditory streams. However, it is not clear how far these motor based attention-in-time effects rely on predictive modeling of the articulatory side of the heard speech or if its role is solely limited to extraction of the rhythmic pattern of speech. One key aspect that would help in sorting these two option out is verifying whether these top-down influences are capable of synthesizing missing articulatory information.

I describe in **Chapter 2** an EEG experiment with participants engaged in a sentence listening task followed by a delayed word recognition task. After the listening phase, participants had to indicate with a button press which word between two options (displayed only when the sentence was completed) was spoken in the previous sentences. Sentences (200 for each subject) were derived from a dataset containing both acoustic and lips movements data recorded via electromagnetic articulography (EMA). The EMA data provides a very accurate characterization of lips-opening and it is commonly used in speech technology research (Savariaux et al. (2017)). We quantified brain entrainment to speech envelope and lips-opening as commonly done in unimodal (Destoky et al. (2019); Luo and Poeppel (2007); Peelle et al. (2012); Bourguignon et al. (2013)) and audio-visual speech studies (Park et al. (2015); Park et al. (2016)). However, differently from previous audio-visual studies, no visual information was ever shown to our participants nor subjects were asked

to recall speech-related visual cues. One key prediction is that if the phenomenon of neural entrainment describes a fundamentally top-down synthetic process, we should find significant entrainment to lips-opening beyond that to acoustic speech. Furthermore, to be considered predictive (Park et al. (2018b)), we expect a temporal dissociation between neural entrainment to attended (speech envelope) and reconstructed (lips-opening) cues, with the latter being relatively anticipated.

1.3 From Single Subject to Conversational Interaction

The majority of experimental paradigms used in cognitive neuroscience are primarily concerned with studying the neural mechanisms of one individual's neurobehavioral processes. Typical experiments isolate humans or animals from their natural environments by placing them in a sealed room where interaction, if present, happens via a computerized program. Subjects are thus reduced to detached observers of situations in which they play no active role in. This dominant focus on single subjects may prevent us to explore the forces that operates between brains while engaged in natural behavioral conditions.

Verbal interaction is an excellent example in which ones behaviour modulates other persons cognitive processes. As two individuals engage in social interaction, they become part of a complex system whose information flow is mediated by visible behavior, prior knowledge, motivations, inferences about the partner's mental states, and history of prior interactions (Schilbach et al. (2013)). Linguistic communication is indeed, among other characteristics, a mind-reading exercise requiring the formulation of hypotheses about the mental states of the speaker (Sperber and Wilson (1998)).

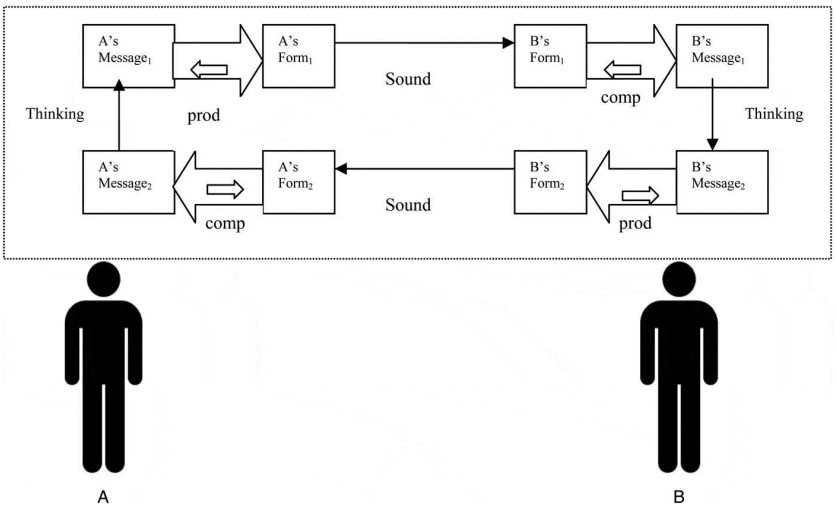

Figure 1.7 *A traditional model of communication between subject A and B. (comp: comprehension; prod: production). Within this model the thick arrows connects message with (linguistic) form. The production arrow represent parallel conversion of single or multiple massages into suitable linguistic form. The internal arrow signifies feedback (e.g., from phonology to syntax) that happens in interactive fashion. The comprehension arrow acts in similar way although the internal arrow could represent feedback from semantics to syntax. All these arrows are capable of handling any type of information (linguistic or nonlinguistic) that occurs during production. For successful communication A's massage must be equal to B's massage. Image adapted from: Pickering and Garrod (2013)*

Although dialogue is undeniably the primary form of language use (Levinson (2016)), previous research has mostly investigated the neural processes subtending either speech production or speech perception, separately and almost ignoring the dynamics of interaction (Price (2012)). Figure 1.7 illustrates a traditional model of communication where production and comprehension are independent of each other. Although this approach has provided fundamental insight on the neural substrates of the speech units for perception and production but the principles and mechanisms that regulate their coordination during natural interaction are still unclear Schoot et al. (2016).

The dynamical process of mutual adaptation which occurs at multiple levels is a key component of natural linguistic interaction that is crucially missing in classical laboratory tasks. One interesting phenomenon during linguistic interaction is that of **Alignment**. Figure 1.8 illustrates an abstract representation of the process of alignment. It shows that interlocutor's linguistic representation interact at multiple levels. The interaction takes place through priming. In simple terms, subjects engaged in a conversation, via a process of automatic imitation tends to accommodate their utterances to their interlocutor at the lexical, phonetic, semantic, and discourse levels simultaneously.

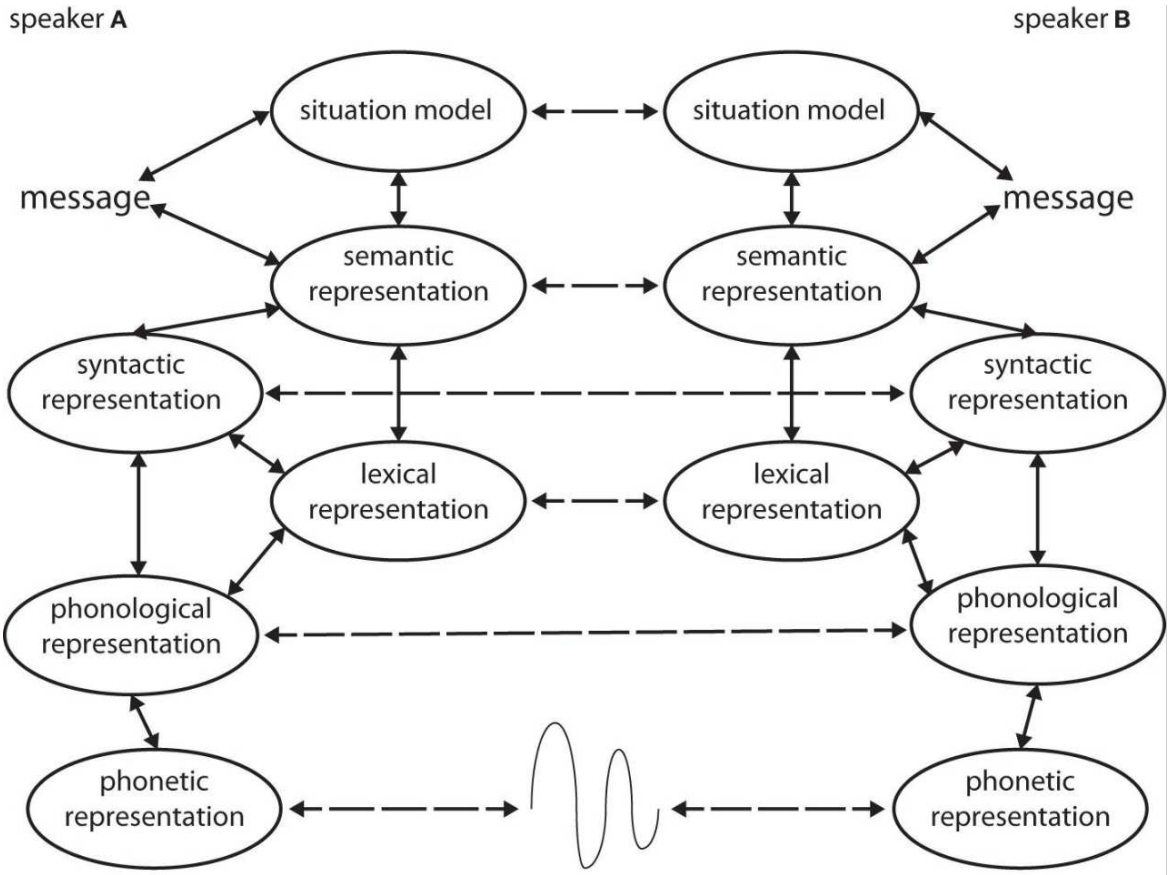

Figure 1.8 *Here speaker A and B represent two interlocutors in a dialog. The circles represent intermediate stages of language production and comprehension. The dashed line indicates alignment channels. Image adapted from: Pickering and Garrod (2004)*

As conversation progress interlocutors simultaneously affect each others mental states. Conversational success is indeed characterized by the shared understanding of the spoken content, speakers' mutual likability, background environment, etc. (Menenti et al. (2012); Garnier et al. (2013)). More interestingly, people involved in a dialogue automatically and implicitly converge at multiple linguistic levels (Bilous and Krauss (1988); Pardo et al. (2010)) as well as with co-verbal bodily gestures (Turner and West (2010)). For instance, agreeing interlocutors tend to copy each other's choices of sounds, words, grammatical constructions as well as the temporal characteristics of speech. Nevertheless, this form of implicit behavioral alignment is still poorly understood, especially regarding its effects on communication efficacy, social and contextual determinants, and neural underpinnings (Stolk et al. (2016)).

1.3.1 Nature of Verbal Alignment

The phenomenon **Alignment** in verbal communication comes from Communication Accommodation Theory developed by Howard Giles (Giles et al. (1987)). This theory concerns with the motivations underlying behavioral changes that people make in their speaking strategies during verbal communication. Specifically, it deals with the individuals' underlying cognitive and affective processes of convergence and divergence through speech.

The nature of convergence depends on many factors such as individual use of language, pronunciation, vocal intensities, pause between words, speaking style, likeness of interlocutor and intimacy of self disclosures (Giles and Smith (1979)). In practice people converge on some of these levels but can diverge on other levels during verbal communication (Coupland et al. (1991)). Attraction (likability, charisma, credibility) as well as perception of the background of interlocutor have an huge impact on convergence (Turner and West (2010)). People with similar personality, beliefs tend to converge more (Pardo et al. (2016)). People with higher desire for social acceptance converge easier to their partner in order to feel a sense of fitting (Coupland et al. (1991)). Individual shift of linguistic behaviors toward interlocutor in order to accommodate them could result in more engaging positive conversation and effectiveness of communication which result in increased mutual understanding, lower interpersonal anxiety and uncertainty.

Divergence on the other hand is the linguistic strategies which results in differences in linguistic characteristics between conversation partners (Coupland et al. (1991)). This helps to create dominance of one individual in two person or group settings. Divergence results in uncomfortable feelings, in the presence of interlocutors. Put simply if we dislike someone we tend to diverge more. For example in 2011, a study on doctor patient communication (Baker et al. (2011)) found that unexpressed differences during communication resulted in less effective treatment. Another example is that when you dont like some one your change your voisce not to sound like them (Eadie (2009)).

Although in the literature the term **Alignment** is often labeled as accommodation, imitation, convergence. Here, without going into the argument of the usage of proper term, we considered as **Convergence** any types of speech characteristics adjustments that are made during verbal interaction and results in a shift towards the speech of interlocutor. Whereas moving away from the speech of the interlocutor is considered as **Divergence** and maintaining one's default linguistic behavior in spite of exposure to the speech of the interlocutor is called **Nochange**.

1.3.2 Problems in quantifying Convergence

Within the context of **Convergence**, most studies focused exclusively on acoustic measures of convergence (e.g. Delvaux and Soquet (2007); Pardo et al. (2010); Nielsen (2011); Babel and Bulatov (2012)) although some also employed perceptual judgements from naive listeners (e.g. Giles (1973); Bourhis and Giles (1977); Namy et al. (2002); Pardo (2006); Babel and Bulatov (2012)). As far as objective acoustic measures are concerned, a large heterogeneity characterizes the variables upon which speech convergence is computed. Among them, duration (Goldinger (1998); Pardo (2013)), speaking rate (Giles et al. (1991); Pardo et al. (2010)), F0 or intonation contour (Babel and Bulatov (2012); Goldinger (1998); Levitan and Hirschberg (2011); Pardo et al. (2010)), intensity (Levitan and Hirschberg (2011); Natale (1975)), voice quality (Levitan and Hirschberg (2011)), vowel spectra (Babel and Bulatov (2012); Delvaux and Soquet (2007); Pardo et al. (2010)), voice onset time (Nielsen (2011); Sanchez et al. (2010); Shockley et al. (2004)), lip aperture (Gentilucci and Bernardis (2007)) and individual phonemic variants (Coupland (1984);Mitterer and Müsseler (2013)).

Most studies have focused on a single feature, while few have examined more than one at the same time (Goldinger (1998); Pardo et al. (2010); Pardo (2013); Bailly and Lelong (2010); Levitan and Hirschberg (2011)). Although these studies show similar results, a great deal of inconsistency and variability across measures does exist (Pardo (2013)). Subjects may converge on multiple attributes simultaneously, or they might converge on some of them and at the same time diverge on other (Bilous and Krauss (1988); Pardo et al. (2012); Pardo et al. (2010); Pardo (2013)). Further complexity is driven by the temporal evolution of convergence during the course of interaction. Research in this area has usually focused on the idea that convergence is linear (i.e. it grows as the conversation proceeds; Suzuki and Katagiri (2007); Natale (1975)). However, subjects do not remain involved to the same degree over the whole course of a conversation (Gregory Jr. et al. (1997); Levitan and Hirschberg (2011); Edlund et al. (2009); Vaughan (2011); Looze and Rauzy (2011); De Looze et al. (2014)) and in fact, speech convergence is more likely to be both a linear and a dynamic phenomenon (Levitan and Hirschberg (2011)). One of the reasons possibly explaining why results on convergence are so variable across studies may also largely depends on task choice. In fact, speech convergence has been analysed both in natural conversational interactions (e.g. Giles (1973); Natale (1975); Coupland (1984); Pardo (2006)) and in more controlled environment, such as asking speakers to repeat single words or utterances after a pre-recorded voice (e.g. Goldinger (1998); Namy et al. (2002); Shockley et al. (2004); Nielsen (2011)).

1.3.3 A New Framework to measure Phonetic Convergence

In **Chapter 3** we develop and validate a computational pipeline to extract convergence at the single word level. Specifically we did not want to force any specific hypothesis on which sub-set of acoustic features to use, but rather exploit the full richness of the acoustic spectrum. In fact, we use Mel-frequency cepstral coefficients (MFCCs) which is a short-term power spectrum representation of sounds. MFCCs are the most widely used in every speech technology application (Kinnunen and Li (2010)). Also some previous studies (Delvaux and Soquet (2007); Aubanel and Nguyen (2010)) have supported the advantages of using MFCC features over phonetic analysis such as formant tracking. Finally, in order to extract un-biases measures of convergence we used a simple yet powerful data driven technique. We use automatic speaker identification techniques, based on GMM (Gaussian Markov Modelling) UBM (Universal Background Model), to model convergence. The GMM-UBM approach provides the structure and parameters to control the behavior of any specific type of data distribution i.e. it is text-independent where there is no prior knowledge of what the speaker will say. The Gaussian components can be considered to be modeling the underlying broad phonetic sounds that characterize a speakers voice. It is also computationally inexpensive and based on a well-understood statistical model. (Reynolds et al. (2000)).

We intended to limit the complexity of the task still retaining the dynamic nature of true dyadic interaction. In **Chapter 3** we describe the development of a new version of the Domino task (Arléo (1997); Bailly and Lelong (2010)), a controlled but yet engaging verbal interaction game, focused on the phonetic level. The Domino task consist in two speaker taking turns in chaining bisyllabic words according to a rhyming rule. The rule is that each speaker has to choose a word whose first syllable is the same as the last syllable previously produced by the other participant. The task shows some resemblance with verbal games that are popular with children throughout the world, and that are referred to as “Grab on Behind”, “Last and First”, or “Alpha and Omega”, in English, and “Shiritori” in Japanese. It shares some characteristics of conversational interactions, though fundamentally focusing on the phonetic aspect. Interestingly, participants speak one at time in alternating turns, whilst allowing us to both control the linguistic material employed by the participants, and to avoid overlaps between the participants’ turns. We performed two studies to investigate two different sides of convergence. One on acoustic side of convergence and another on the articulatory side of convergence.

1.3.4 Neural Correlates of Phonetic Convergence

Although the investigations of convergence at the brain level are still sparse, few interesting studies ignited the exploration of inter-brains neural synchrony during conversation by following a hyper-scanning approach (Hasson et al. (2012)). These studies showed that the speaker's brain activity is spatially and temporally coupled with that of the listener (Silbert et al. (2014)), and that the degree of coupling and anticipation from the listener's side predicts comprehension (Stephens et al. (2010); Kuhlen et al. (2012); Dikker et al. (2014); Liu et al. (2017)). Other studies instead explored how social factors affect the neurobehavioral pattern of communication. For example, it has been shown that neural activity in the left inferior frontal cortex synchronizes between participants during face-to-face interaction (Jiang et al. (2012)) and social role does play an important part in such interpersonal neural synchronization (Jiang et al. (2015)).

While techniques such as fMRI and fNIRS, due to their poor temporal resolution, are most suited to investigate alignment at the motivational, emotional or semantic level, the higher temporal resolution of EEG or magnetoencephalography (MEG) makes these techniques more suitable to investigate the intrinsically, faster, dynamics intervening during speech coordination. One aspect that has recently received attention regards the online negotiation of each-other turns during dialogues (turn-taking) which relies on a complex between-speaker neurobehavioral coordination (Levinson (2016)). For instance, in a task where subjects alternate in pronouncing letters of the alphabet, inter-brain EEG oscillations in the theta/alpha (6-12 Hz) bands were synchronized in temporo-parietal regions as well as linked to behavioral speech synchronization indexes (Kawasaki et al. (2013)). More recently, a dual MEG/EEG study employing a similar number-counting task, reported alpha suppression in left temporal and right centro-parietal electrodes during speech interaction as opposed to a condition of speaking alone (Ahn et al. (2018)). Using a dual-MEG set-up during natural conversations allowed to show that rolandic oscillations in the alpha (~10-Hz) and beta (~20-Hz) bands depended on the speaker's vs. listener's role (Mandel et al. (2016)). In the left hemisphere, both bands were attenuated during production compared with listening. Before the speaker and listener swapped roles, power in the alpha band was briefly enhanced in the listener. These studies have thus begun to explore the intra- and inter-brain oscillatory dynamics underlying one key behavioral coordination aspect during speech interaction, which is our ability to accommodate to the temporal properties of our partners' speech (Jungers and Hupp (2009)).

Beside temporal characteristics, speakers can shape how they speak and listen to speech. In fact, interlocutors adjust both their speech production and perception to their audience. For

example, important adjustments are introduced while speaking to infants (Cooper and Aslin (1990)), to foreigners (Uther et al. (2007)) or under adverse conditions (i.e., hearing loss or environmental noise;(Payton et al. (1994))). Adjustments in speech production consider listeners' perceptual salience and effort in listening (Lindblom (1990)). Listeners adapt to talker characteristics (Nygaard et al. (1994); Bradlow and Bent (2008)), suggesting also great flexibility in speech processing mechanisms (Samuel and Kraljic (2009)). When engaged in a conversation instead, interlocutors align (or converge) their phonetic realizations to each other. Phonetic convergence then, amounts to the gradual and mutual adaptation of our speech targets towards a phonetic space shared by our interlocutor.

In **Chapter 4** we aimed at investigating the neural signature of such dynamic phonetic alignment. We asked pairs of participants to engage in an interactive speech game while dual-EEG was recorded. By this manner, we aimed at investigating interpersonal action-perception loops where one person's action (speech articulation) transforms into the sensory input (speech sound) for the other participant, and vice-versa. To this purpose, we used the Verbal Domino Task (VDT) Bailly and Lelong (2010), a fast-paced and engaging speech game allowing a relatively well controlled interaction and involving a turn-based phonetic exchange. At each turn, speakers are presented with a pre-selection of two written words and have to choose and read out the one that begins with the same syllable as the final syllable of the word previously uttered by their interlocutor (see Figure 4.1C). These constraints make the task different from a natural conversation, but contain the fundamental phonetic interactive component we needed for the present investigation. Before the interactive task, each participant's initial phonetic fingerprint was statistically modeled. The phonetic fingerprint was extracted from the individual acoustic spectral properties. We then computed how well the model of one speaker can identify the speech data of her/his interlocutor, at the single word level. In this sense, identification performance of these models allows to estimate how similar the two speakers are during the interactive task. We adopted a conservative criterion for convergence, which as defined as all instances where both participants adapted their speech properties to get closer to each other. All other cases were treated as non-convergence.

The EEG analysis focused on the oscillatory power modulations during phonetic convergence as compared to non-convergence, in three different epochs of interest. The first epoch, locked to speech production onset, was selected to investigate the preparatory activities in the speaking brain. The second and third epochs targeted the listening brain's activities: the ongoing brain activities prior to listening and those induced by speech listening. We expect that phonetic convergence could be associated with modulations of oscillatory activity in the alpha and beta range, two prevalent rhythmic modes of the brain, that have been

already involved in natural conversation and, in particular, in speech turn-taking (Mandel et al. (2016)). In addition, based on previous studies, we can hypothesize a dissociation between effects in the alpha and beta ranges, depending on the role each participant is playing at any given time point. The speaker role would elicit greater suppression in the alpha range, as often reported in behavioral synchronization tasks (Tognoli and Kelso (2015)). By contrast, the listener role would produce beta desynchronization, which has been associated to sensorimotor transformations during speech listening tasks (Bartoli et al. (2016)) and top-down predictive coding of upcoming sensory signals (Cope et al. (2017)).

1.4 Summary

The overall structure of the thesis takes the form of five chapters including introduction and general discussion. **Chapter 2** provides evidence that neural tracking of speech envelope during passive listening of speech, includes the anticipatory synthesis of missing articulatory speech cues. We will later use this results to discuss the neuro-functional origin of these reconstructive processes, by suggesting that they may reside in the computations run in precentral areas. In **Chapter 3** we investigate a particular phenomenon in verbal interaction, which is phonetic **Convergence**. This chapter deals with the intricacy involved in first designing a suitable task and then in building an algorithm to quantify the word-level dynamics of such kind of joint behavioral phonetic adaptation. This chapter reports two studies, one on the acoustic, the other on the articulatory side of convergence. **Chapter 4** uses the same task and the same algorithm to find the neural markers for Convergence in a dual-EEG study. In conclusion, the last chapter (**Chapter 5**) will draw upon the entire thesis, tying up the various theoretical and empirical strands in order to discuss the main findings of the experiments and their implication to future studies into this area.

To the best of my knowledge this is the first research that brings together the role of motor systems during verbal interaction and its effect on articulatory dynamics. I strongly believe that this project may offer an important opportunity to advance the understanding of the role of motor system during speech-based joint interaction. Moreover, it may provide valuable insights in order to build a speech recognition system that will incorporate motor/articulatory knowledge to improve its performance as well as adapt to the human interlocutor in a human-like fashion.

Chapter 2

Cortical tracking of speech reveals top-down reconstructive processes

A manuscript containing description of the following study has been submitted to:: Plos Biology 2018.

Authors: Sankar Mukherjee[1], Alice Tomassini[1], Leonardo Badino[1], Aldo Pastore[1], Luciano Fadiga[1,2], Alessandro D'Ausilio[1,2]

[1]*Center for Translational Neurophysiology of Speech and Communication, Istituto Italiano di Tecnologia, Ferrara, Italy*

[2]*Section of Human Physiology, University of Ferrara, Ferrara, Italy*

2.1 Abstract

Cortical entrainment to the quasi rhythmic components of audio-visual speech has been shown to be involved in speech comprehension. It has been further suggested that neural entrainment may reflect top-down temporal predictions of sensory signals. However, key aspects of a predictive model are its anticipatory nature and its ability to reconstruct missing information, and it is still not clear whether cortical entrainment does show these two properties. Here we tested specifically these hypotheses by measuring cortical entrainment to the acoustic speech envelope and to lips-opening. Differently from previous studies, we had subjects listening to the acoustic speech signal, while no visual cue was presented. To capture the anticipatory component, we analyzed brain-speech and brain-lips coherence at multiple negative and positive lags. We found significant cortical entrainment in the delta range to the acoustic speech envelope as well as to the (absent) lips-opening. Most importantly, these

two phenomena were temporally dissociated. While entrainment to the acoustic speech peaked around +0.3 s lag (i.e., when EEG followed speech by 0.3 s), entrainment to the lips was significantly anticipated and peaked around 0-0.1 s lag (i.e., when EEG was virtually synchronous to the lips signal). Our results demonstrate that neural entrainment during speech listening is anticipatory and capable of reconstructing missing information related to lips movement.

2.2 Materials and Methods

2.2.1 Participants

Twenty-five healthy native Italian speakers participated in the study (16 females; mean age=23.5; age range=20-28 years). All participants were right-handed and had normal self-reported hearing. Participants were all naïve with respect to the aims of the study and were all paid (€10/h) for their participation. The study and experimental procedures were approved by the local ethics committee (Comitato Etico della Provincia di Ferrara). Participants provided written, informed consent after explanation of the task and experimental procedures in accordance with the guidelines of the local ethics committee and the Declaration of Helsinki.

2.2.2 Stimuli

The stimuli were chosen from the Multi-SPeaKing-style Articulatory corpus (MSPKA; Canevari et al. (2015)), an Italian dataset of simultaneous recordings of speech, articulatory data (lips, jaw and tongue) and corresponding Phonetic labels. Speech was recorded at a sampling rate of 22.5 kHz. Articulators were tracked at a sampling frequency of 400 Hz by means of an electromagnetic articulography system (EMA) using the NDI (Northen Digital Instruments, Canada; Berry (2011)). In the present study, we used data from two 5-degrees-of-freedom (x, y, z positions, pitch and roll) sensor coils glued on the upper lip (UL) and lower lip (LL). For head movement correction, a 6-degrees-of-freedom sensor coil was fixed on the bridge of a pair of glasses worn by the participants. Articulatory and acoustic data were recorded from one female speaker reading news-related sentences and tongue twisters. For this study, we selected 200 sentences from the MSPKA corpus. Stimuli were manually checked to ensure there were no missing or noisy data. All stimuli were then normalized to the same average intensity (71 dB). It should be noted that the sentences had different duration (4.36 $\pm$ 1.73 s; MEAN $\pm$ SD), with an average number of phonemes equal

to 13.32 ± 3.5 phonemes/s (MEAN ± SD) and an average phoneme duration of 0.08 ± 0.02 s (MEAN ± SD).

2.2.3 Experimental Procedure

Subjects sat in a dimly lit, sound and electrically shielded room in front of a screen and two loudspeakers (distance: ~1 m) and with their left and right hands resting on a response box (RB-840 Response Pad, Cedrus Corporation). The experiment was composed of two blocks with a rest period of 2 min in between. Each block consisted of 100 trials that were structured as follows. Each trial started with the appearance of a white fixation cross at the center of the screen. Subjects were required to maintain fixation throughout the duration of the trial. After a variable time (2-2.5 s), the fixation cross changed color (from white to green) and a randomly selected sentence was acoustically presented from the two loudspeakers. When the sentence was completed, the fixation cross turned back white. After a variable pause (1-1.5 s) the fixation cross was replaced by two words displayed above and below the center of the screen. Participants were asked to indicate with a button press which of the two words was spoken in the previously presented sentence. The above/below word was chosen by pressing the up/down key of the response box with the left and right index finger, respectively. Participants were not prompted to respond as quickly as possible, but they were encouraged to select the correct word. However, they were aware that if they did not respond within 5 s, the next trial was automatically run. Stimuli presentation and response acquisition were controlled with Matlab Psychtoolbox-3.

2.2.4 EEG recording

EEG data were recorded continuously during the experiment with a 64-channel system using Ag/AgCl impedance-optimized active electrodes (Brain Products GmbH, Gilching, Germany). Electrodes were placed according to the international 10–20 system and on-line referenced to the right mastoid. Electrooculograms (EOGs) were recorded with four electrodes removed from their original scalp sites (FT9, FT10, PO9, PO10) and placed at the bilateral outer canthi (horizontal eye movements) and below and above the right eye (vertical eye movements, blinks). Electrode impedance was kept <5 kΩ. Data were acquired at a sampling rate of 1000 Hz.

2.2.5 Preprocessing of EEG, speech and articulatory data

All the analyses were performed with the MNE-Python software (Gramfort et al. (2013)). Continuous EEG data were first bandpass-filtered between 0.5 and 40 Hz (two-pass Butterworth filter, third-order). Data were then epoched in shorter segments corresponding to the duration of each trial and visually inspected for bad channels and/or artifacts in the time domain. To identify and remove artifacts related to participants' eye movements and muscle activity, we used Independent Component Analysis (ICA). Noisy channels (not included in the ICA) were then interpolated using a distance-weighted nearest-neighbor approach. Finally, data were re-referenced using a common average reference over all electrodes.

The amplitude envelope of the acoustic speech signal was calculated by adapting a previously described method (Smith et al. (2002)). We used the Chimera toolbox and defined eight frequency bands in a range 100–10000 Hz that are equally spaced on the cochlear map. The speech signal was first filtered within those eight frequency bands (two-pass Butterworth filter, fourth-order). Then, we computed the absolute value of the Hilbert transform for each bandpass-filtered signal. The result of this computation was then averaged across frequency bands, yielding the wideband speech envelope.

Lips-opening was derived from the EMA data (x-y midsagittal coil positions) as the absolute distance between UL and LL in the y-plane, providing a measure of the amount of opening of the lips during speech production. The lips signal was subsequently smoothed by applying an adaptive median filter with a window of 10-50 ms and an elliptic low-pass filter (20 Hz cutoff frequency). Both the EEG and speech signals were down-sampled to match the lower sampling rate of the lips signal (i.e., 400 Hz). Example traces of the raw audio signal, speech envelope and lips-opening are shown in Figure 2.1a.

2.2.6 Spectral analysis

We analyzed EEG data recorded during stimulus presentation (i.e., during sentence listening) and corresponding speech and lips data. To exclude the stimulus-evoked components associated with speech onset (in the EEG signal), we discarded the initial 0.5 s of each sentence (see Figure 2.1b). All data were then cut into 3-s segments. As a consequence, sentences shorter than 3 s were discarded from the analysis (42 out of 200 sentences) while sentences longer than 6 s contributed with more than one segment to the analysis.

All the analyses were then performed on 2-s segments (the first and last 0.5 s of the original 0.3-s segments were used for the analysis based on time-shifting of the EEG signal, see below). The power spectral density (PSD) of both the speech and lips signals was

estimated for frequencies between 0.5 and 12 Hz using the multitaper method (0.5 Hz steps, 2 Hz smoothing; Percival and Walden (1996)).

The cross-spectral density between EEG and speech, EEG and lips as well as speech and lips, was computed on single trials with multitaper frequency transformation. The multitaper method was applied at the following frequency ranges: delta (1-4 Hz, centered at 2.5 Hz; 1.5 Hz smoothing), theta (4-8 Hz, centered at 6 Hz; 2 Hz smoothing) and alpha (8-12 Hz, centered at 10 Hz; 2 Hz smoothing).

Coherence was then computed for each subject, frequency-band (delta, theta and alpha) and EEG channel for temporally aligned signals (0-s delay) as well as by shifting the EEG signal in time by different amounts of delays: from -0.5 s (EEG precedes speech and lips by 0.5 s) up to +0.5 s (EEG follows speech and lips by 0.5 s) in steps of 50 ms (see Figure 2.1b). Speech-lips coherence was also computed at the same delays by time-shifting the speech signal accordingly (see Figure 2.1d).

2.2.7 Statistical analysis

Statistical analysis at the group level was performed using non-parametric cluster-based permutation tests (Maris and Oostenveld (2007)). First, we evaluated statistically both speech-brain and lips-brain coherence at different delays (from -0.5 to +0.5 s in steps of 50 ms) and for the three frequency bands (delta, theta and alpha). To this aim, we generated surrogate data by breaking the original association between the EEG and the speech/lips signals. In practice, for each subject and channel, we randomly reassigned the EEG segments (used to calculate coherence in the original analysis) to speech/lips signals that corresponded to different stimulus segments compared to the original ones. Note that the randomly reassigned speech/lips signals could still be part of the same sentence as the original one (but a different and non-overlapping 2-s segment). This surrogate data does not contain anymore the original temporal dependencies between the EEG and the speech/lips signals and can thus serve as control data. Coherence maps were then calculated for each subject-specific surrogate data in the same way as described above for the original data. Then, for every sample (here defined as [channel, delay]), a dependent-sample two-tailed t value was computed (between the original and surrogate datasets). All samples for which this t value exceeded an a priori decided threshold (uncorrected $p < 0.05$) were selected and subsequently clustered on the basis of spatial and temporal contiguity. Cluster-level statistics was computed by taking the sum of t-values in each cluster. The cluster yielding the maximum sum was subsequently used for evaluating the difference between the two datasets (with the maximum sum used as

test statistic). We permuted the data across the two datasets (swapping the coherence values for the two datasets in a random subset of subjects), and for each random permutation (10000 iterations), we calculated again the test statistics in the same way as previously described for the original data. This procedure generates a surrogate distribution of maximum cluster t-values against which we can evaluate the actual data. The p-value of this test is given by the proportion of random permutations that yields a larger test statistic compared to that computed for the original data.

Besides evaluating both speech-brain and lips-brain coherence against the coherence that can be expected by chance for each pair of signals (as indexed by the surrogate data), we also evaluated whether neural activity is more strongly related to the acoustic speech or to the corresponding lip kinematics at different temporal delays. This analysis was limited to the delta-band (1-4 Hz), as this was the only band where we found significant coherence between EEG and lips (previously described analysis). Group-level statistical analysis was performed again by applying non-parametric cluster-based permutation statistics (described above). Finally, partial (speech-brain and lips-brain) coherence was computed at two different delays, at 0-s and at +0.3-s delay and for both delays we compared the obtained coherence values for the two signals by means of cluster-based statistics (as described above).

2.3 Results

Participants performed the listening task with high accuracy, selecting the correct word in nearly 100% of the trials (98.4 ± 1 %; MEAN±SD), suggesting that they were paying attention during the listening phase and the sentences were perfectly intelligible. Trials in which participants gave the incorrect response were discarded from the analysis.

In line with previous works (Ahissar et al. (2001); Peelle et al. (2012); Golumbic et al. (2012); Park et al. (2016)), both the speech envelope (hereinafter referred to as speech) and the corresponding lips kinematics (i.e., lips-opening; hereinafter lips) are dominated by low-frequency components, as indicated by their power spectra showing distinct peaks in the delta (1-4 Hz) and theta (4-8 Hz) range (Figure 2.1c). Based on the prominent low-frequency content of the speech and lips signals and on previous evidence (Ding and Simon (2013); Gross et al. (2013); Kayser et al. (2015)), we restricted our analyses to three low-frequency bands: delta (1-4 Hz), theta (4-8 Hz) and alpha (8-12 Hz).

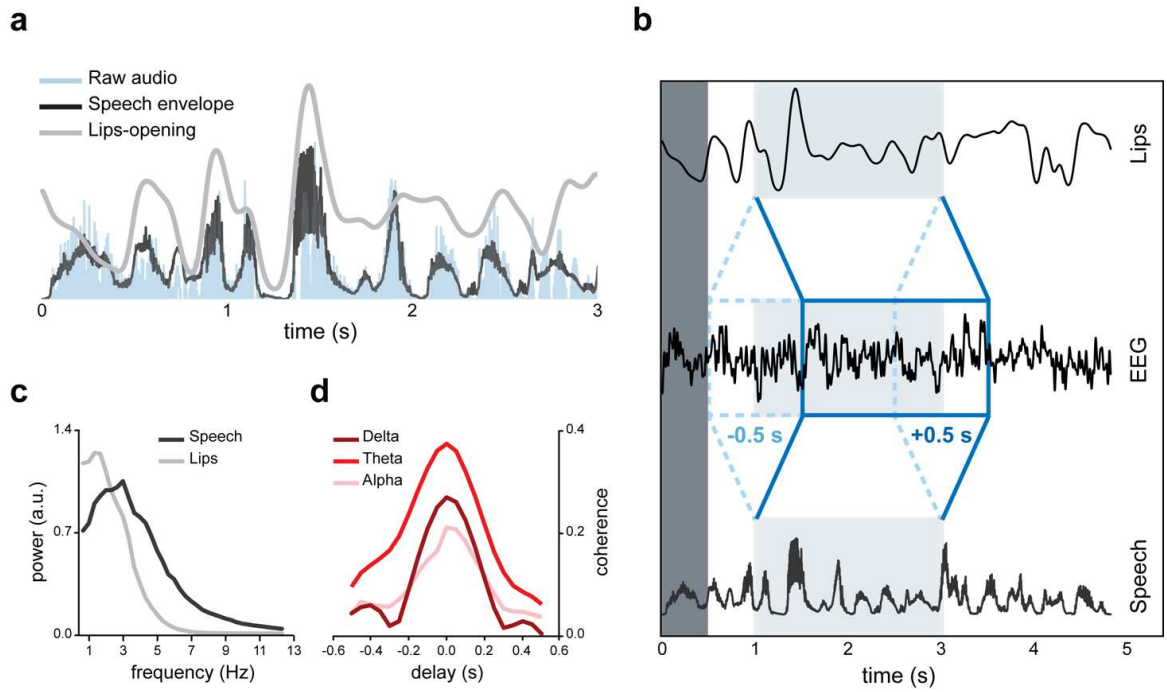

Figure 2.1 *(a) Raw audio signal (light blue), speech envelope (dark grey) and corresponding lips-opening (light grey) for a 3-s segment of an example stimulus. (b) Schematic illustration of speech- and lips-brain coherence computation as a function of delay. Coherence is computed for 2-s stimulus segments (highlighted in light grey) by time-shifting the EEG signal by different amounts of delays from -0.5 s (light blue; EEG precedes speech/lips by 0.5 s) up to +0.5 s (dark blue; EEG follows speech/lips by 0.5 s) in steps of 50 ms. The initial 0.5-s segments (highlighted in dark grey) are discarded from the analysis to exclude any evoked component in the EEG signal associated with stimulus onset. (c) Average power spectra of the speech and lips signals calculated on the entire stimulus dataset. (d) Coherence between the speech and lips signals in the delta (1-4 Hz; dark red), theta (4-8 Hz; red) and alpha (8-12 Hz; pink) bands as a function of delay (i.e., for speech signals that are time-shifted from -0.5 to +0.5 s relative to the lips signals).*

First, we computed speech-brain coherence for temporally aligned signals (zero delay) as well as for EEG signals that were time-shifted (relative to speech signals) by different amounts of delays: from -0.5 s (EEG preceding speech by 0.5 s) up to +0.5 s (EEG following speech by 0.5 s) in steps of 50 ms (see Figure 2.1b). Brain activities are coherent with the acoustic speech at all tested frequency bands (cluster-based statistics against trial-shuffled surrogate data; $p<0.025$; Figure 2.2a&b), although to a different degree (delta=0.038 $\pm$ 0.002, theta=0.035 $\pm$ 0.001, alpha=0.032 $\pm$ 0.0008, coherence mean $\pm$ s.e.; $F_{(2,48)}=6.758$, $p=0.003$; one-way ANOVA for repeated measures with frequency-band [delta, theta, alpha] as within-subject factor; data collapsed across channels and delays), with delta-band coherence being significantly stronger than alpha-band coherence ($p<0.001$; paired-sample t-test, Bon-

ferroni corrected). Despite the difference in strength, coherence shows comparable spatial distribution for all frequency bands, with maximal concentration over bilateral fronto-central sites (see Figure 2.2b; Peelle et al. (2012); Vander Ghinst et al. (2016);Molinaro and Lizarazu (2018); Bourguignon et al. (2018)).

Most interestingly, irrespective of the frequency band, speech-brain coherence is weak and non-significant at 0-s delay and steadily increases at positive delays, reaching maximal values when the EEG follows the acoustic speech by ~0.25 s (i.e., at +0.2/0.3-s delays; Crosse et al. (2015); O'sullivan et al. (2014); Figure 2.2b).

Next, we examined whether brain activities are also coherent with the articulatory signal – the lips – and found statistically significant coherence in the delta band (p=0.003 at 0-s delay; Figure 2.2a&b). Although sharing a similar fronto-central topography, lips-brain coherence seems to be temporally dissociated from speech-brain coherence, being maximal at earlier delays between -0.05 s and +0.15 s (see Figure 2.2b). This suggests that neural entrainment to the (reconstructed) lip movements precede in time the entrainment to the speech envelope.

To test this hypothesis directly we compared speech-brain and lips-brain coherence at the different delays (from -0.5 to +0.5 s). Given that lips-brain coherence was found to be statistically significant only in the delta range, this analysis was confined to this frequency band. A group-level cluster-based permutation test yielded two significant clusters, showing comparable fronto-central topography: one negative, indexing lips-brain coherence larger than speech-brain coherence, spanning from -0.05 s to +0.15 s delay (p=0.017 at 0-s delay), and one positive, indexing speech-brain coherence larger than lips-brain coherence, extending from +0.2 to +0.45 s delay (p<0.0001 at +0.3-s delay; see Figure 2.3b). This significant temporal dissociation further confirms that delta-band brain activities are anticipatorily coupled to (reconstructed) lips oscillatory dynamics while this coupling is later replaced by entrainment to the ensuing acoustic speech signal.

The observed coherence between brain and lips is unlikely to be a mere by-product of the existing coherence between lips and speech. First of all, coherence between the acoustic speech and the lips is maximal (at all delays) in the theta, not in the delta, range (see Figure 2.1d). This might reflect the time scale of syllables production and it has been previously reported (Hauswald et al. (2018)). Secondly, and most importantly, lips-brain and speech-brain coherence show a different dependence from the temporal alignment between the two signals: the former is stronger when the EEG signal is virtually synchronous to the lips signal associated with the sentence production, while the latter is established at later times, when the EEG follows the acoustic speech signal by a few hundreds of milliseconds.

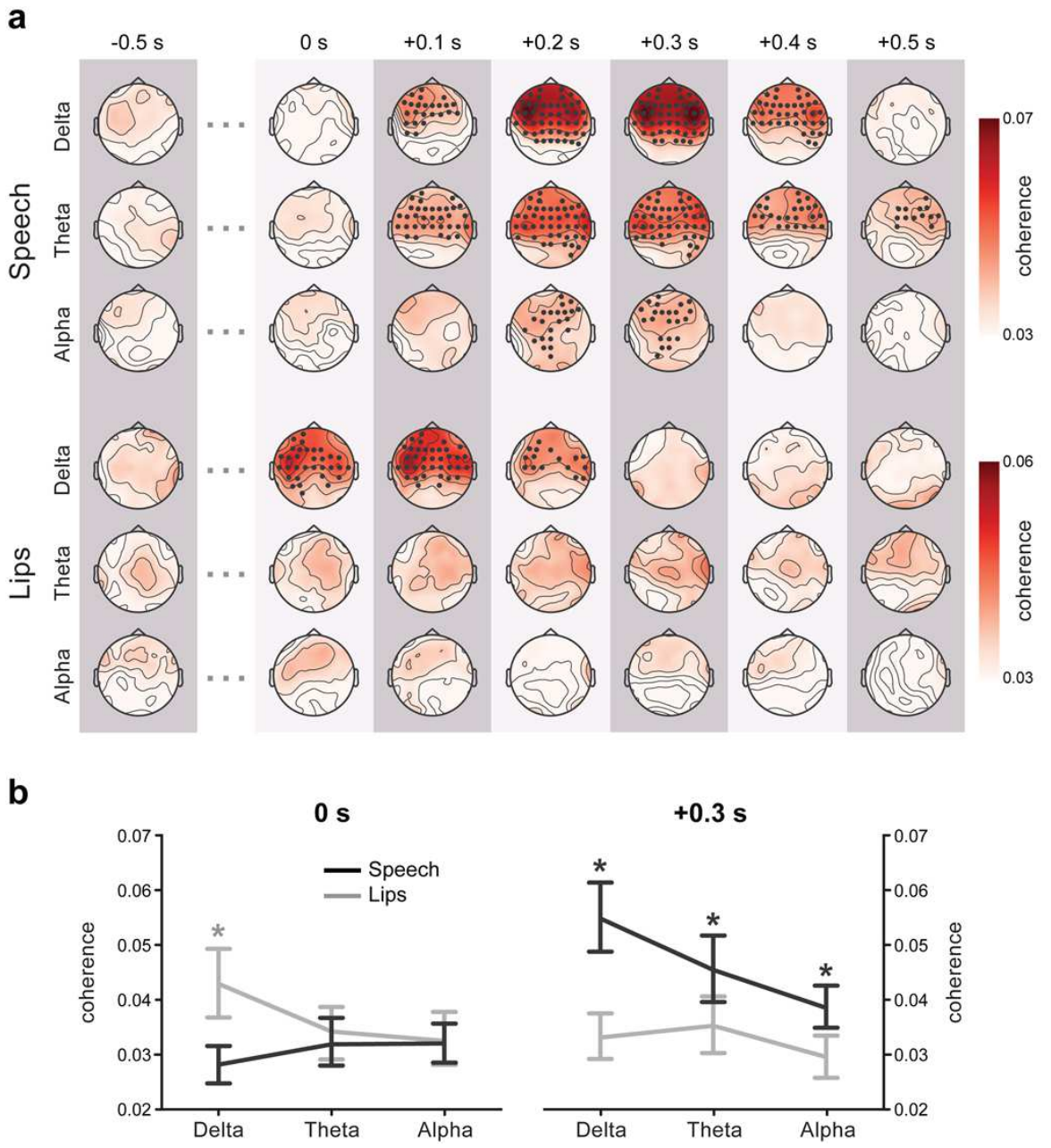

Figure 2.2 *(a) Topographies of speech- (top) and lips- (bottom) brain coherence in the delta (1-4 Hz), theta (4-8 Hz) and alpha (8-12 Hz) bands are shown for -0.5-s delay (leftmost graph) and from 0-s to +0.5-s delays, in steps of 0.1 s. Black dots represent significant channels ($p<0.025$) according to (group-level) cluster-based permutation tests against surrogate data (generated by trial-shuffling). (b) Speech- (black) and lips- (grey) brain coherence averaged across channels is shown at 0-s (left) and at +0.3-s delay (right) as a function of frequency band. Asterisks indicate significant ($p<0.025$) frequency-bands for speech- (black) and lips- (grey) brain coherence, according to cluster-based permutation statistics.*

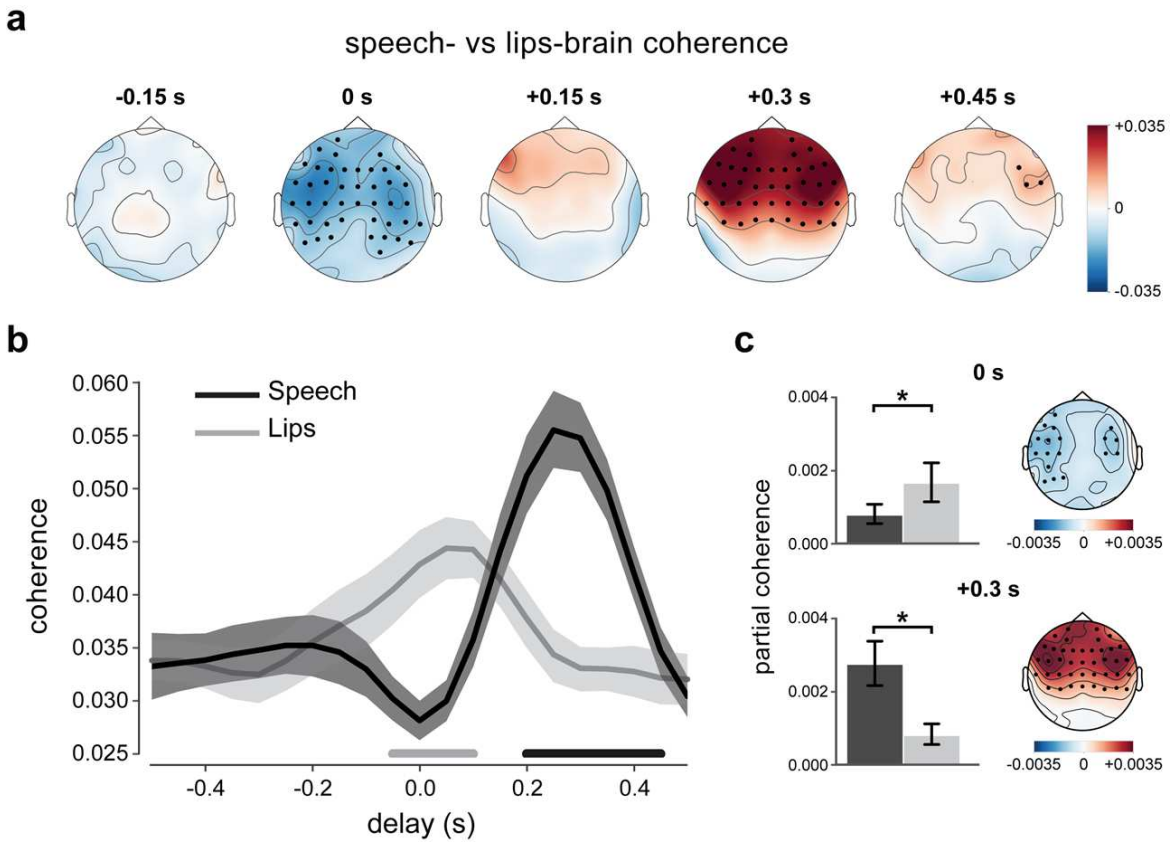

Figure 2.3 *(a) Topographies of the difference between speech- and lips-brain coherence are shown for delays from -0.15 to +0.45 s (in steps of 0.15 s). Black dots represent significant channels (p<0.025) as resulted by (group-level) cluster-based permutation test. (b) Speech- (black) and lips- (grey) brain coherence averaged across channels is shown as a function of delay (from -0.5 to +0.5 s). Horizontal lines indicate significant (p<0.025) delays according to cluster-based permutation test (black: speech>lips; grey: lips>speech). (c) left. Bars show speech- (black) and lips- (grey) brain partial coherence averaged across channels at 0-s (top) and +0.3-s (bottom) delays. Asterisks indicate statistical significance (p<0.025) as resulting from cluster-based permutation statistics. right. Topographies of the difference between speech- and lips-brain partial coherence at 0-s (top) and +0.3-s (bottom) delays. Black dots represent significant channels (p<0.025) according to cluster-based permutation test.*

To further ascertain the relation between the observed lips- and speech-brain coherence we run a partial coherence analysis, i.e., we recomputed coherence at all delays by partialling out the contribution of speech and lips, respectively. This analysis revealed a certain degree of inter-dependency between the two phenomena, as both speech- and lips-brain coherence values were diminished, irrespective of delay; however, it crucially confirmed the previous pattern of results: at 0-s delay, lips-brain (partial) coherence is significantly stronger than

speech-brain (partial) coherence (p=0.019; cluster-based statistics), and the reverse is true at +0.3-s delay, where speech-brain (partial) coherence exceeds lips-brain (partial) coherence (p<0.0001; cluster-based statistics; see Figure 2.3c).

Chapter 3

Behavioral indexes of Phonetic Convergence during Verbal Interaction

When people engage in social interaction, they adjust their speech in order to accommodate to each other. Although convergence is a well-known phenomenon, its quantitative assessment is still an open area of research. In this chapter, first I describe how we create a verbal interaction task and an algorithm to quantify Convergence. Next I present two studies which applied these task and algorithm to find its nature and relationship with articultory counterpart and also its acoustic supra-segmental features.

3.1 A Verbal Interaction task and an Algorithm to quantify Convergence

To circumvent some of the problems that hamper an effective and robust measurement of Convergence, we did not use spontaneous conversations. Rather we used a constrained interaction task that allows better experimental control. Convergence is computed by using an automatic speaker identification technique to quantify subject's effort in moving towards the other speaker. Finally, we implemented a robust method to combine both participants' shift towards each other and explain their behavior over time. The work related to the development of the task and algorithm has been published as Intespeech conference papers in 2017 and 2018 and in Human Brain Mapping journal. Text and figures in Section 3.1 have been adapted from these publications.

3.1.1 The Task

We used the Verbal Domino Task (VDT, Mukherjee et al. (2017); Mukherjee et al. (2018); adapted version from Bailly and Lelong (2010)). The VDT is an interactive, collaborative task jointly performed by two participants who pronounce a sequence of disyllabic words with an alternation between participants across words. Along the sequence, there is a match between the last syllable of each word and the first syllable of the following one, in its pronounced form, written form, or both. The task shows some resemblance with verbal games that are popular with children throughout the world, and that are referred to as "Grab on Behind", "Last and First", or "Alpha and Omega", in English, and "Shiritori" in Japanese. It shares some characteristics of conversational interactions, though fundamentally focusing on the phonetic aspect. Interestingly, participants speak one at time in alternating turns, whilst allowing us to both control the linguistic material employed by the participants, and to avoid overlaps between the participants' turns.

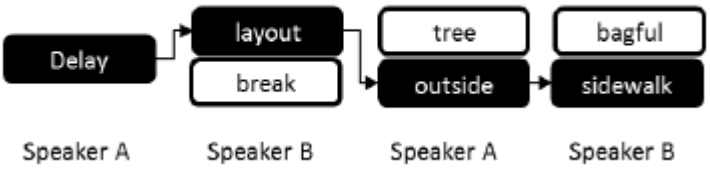

Figure 3.1 *Example of word domino.*

During the task, the two speakers are alternatively presented with two words on a screen and must read aloud the word whose initial syllable coincides with the final syllable of the word previously produced by the other speaker. For example, on having heard Speaker A pronouncing the word "delay", Speaker B is offered to choose between "layoff" and "tissue" and is expected to pronounce "layoff" (see Figure 3.1). The task is collaborative in that participants have to make the right choices for them to reach a common goal, i.e. to jointly go through the domino chain up to the end. The choice of presenting the words visually, rather than allowing self-generation, was required to avoid participants using (a variable amount of) time "searching" for possible candidate words. In fact, this would have added a fundamental memory retrieval component that couldn't be controlled. Word self-generation can also introduce frequent stops in the chain. Moreover, the presentation of two alternatives, instead of a single option, forced participant to actively listen to their partner to select the correct word.

All the experiment conducted here was divided into three main sections. Solo recordings were performed before and after the Duet sessions (Solo_Pre, Solo_Post).

3.1.2 The Algorithm

Convergence Algorithm is consisting of two parts. First, we create speaker dependent models of each speakers and then we quantify Convergence based of these speaker-dependent models predictions.

GMM-UBM

In order to extract unbiased measures of convergence, we used a data driven, text independent, automatic speaker identification technique, based on Gaussian Markov Modelling (GMM) Universal Background Model (UBM). The Gaussian components model the underlying broad phonetic features (i.e. MFCCs) that characterize a speaker's voice and are based on a well-understood statistical model (Reynolds et al. (2000)). In previous work (Bailly and Martin (2014)), a similar method was used to extract phonetic convergence, with some important differences. In Bailly and Martin (2014), the model was trained and tested on phonemes, whereas we applied it at the whole word level. We did not force any hypothesis on what features to use, as we aimed to exploit the full richness of the acoustic spectrum by using Mel-frequency cepstral coefficients (MFCCs) (Aubanel and Nguyen (2010)).

We used the MSR Identity Toolbox (Sadjadi et al. (2013)) for GMM-UBM modelling. First a UBM was trained with the pooled Solo_Pre speech data of all the participants. Then, individual speaker-dependent models were obtained via maximum a posteriori (MAP) adaptation of the UBMs to the Solo_Pre speech data of each speaker separately. The GMM-UBM has multiple hyper-parameters and different settings of these hyper-parameters can affect the performance of speaker-dependent models. A cross-validation technique was used to choose the optimum hyper-parameter settings. Solo_Post speech data were used as a validation set, and each speaker-dependent model's performance was verified against the UBM model.

Phonetic convergence computation

Phonetic convergence is computed on word pairs. For a word pair to be a convergent one, the acoustic properties of the words for the two speakers must be similar to each other, as well as these properties must be significantly different than the speakers original speech.

First of all, speaker-dependent models are grouped together according to their dyads, i.e. if speaker A and speaker B interacted in the experiment, speaker-dependent model A and speaker-dependent model B are used for the following analysis (Figure 3.2A). During the duet, each word (i.e., the MFCCs of the speech) is tested with its corresponding grouped dyad models (Figure 3.2A). The test is performed by using the log-likelihood ratio score (LLR) which allows us to compare how well two statistical models can predict test samples. LLR of samples y_x (y is the MFCCs and x is the speaker identity) during the duet is computed using Equation 1:

$$LLR_{DUET}(y_x) = log\left(\frac{p(y_x|H_A)}{p(y_x|H_B)}\right) \quad (1)$$

where H_A and H_B are the speaker-dependent models of speaker A and B respectively. Now, when x is speaker A, LLR scores are positive (numerator greater than the denominator), whereas if x is speaker B, we get a negative score. The same computation, run on Solo_Pre data, is then used to obtain the distribution of LLR_{PRE} scores which represents the baseline for each subject. In equation 2 H_UBM represents the UBM model.

$$LLR_{PRE}(y_x) = log\left(\frac{p(y_x|H_A)}{p(y_x|H_{UBM})}\right) \quad (2)$$

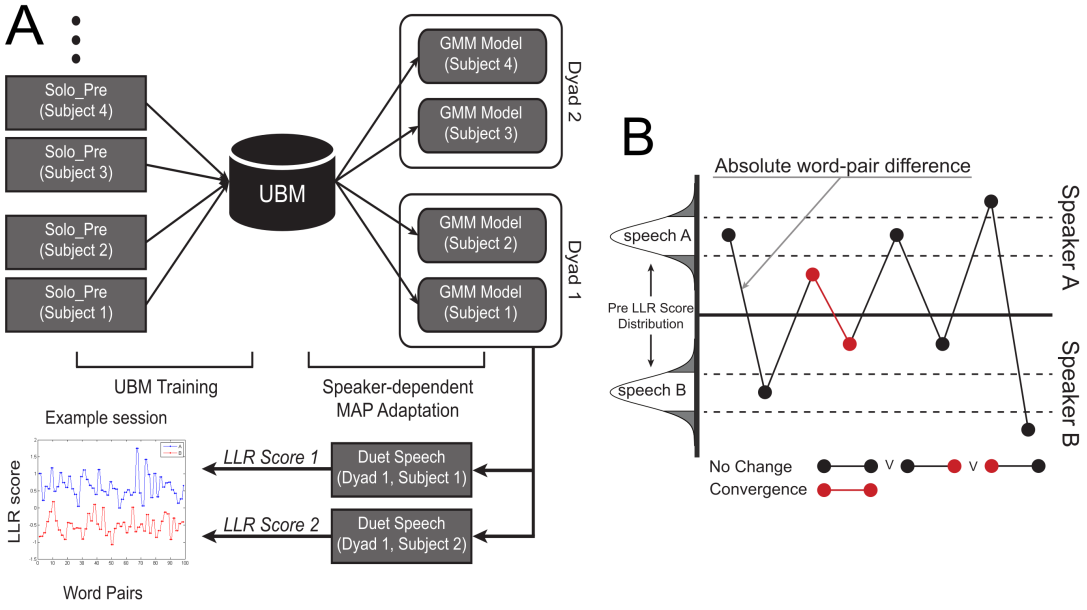

Figure 3.2 *A. Schematic diagram of GMM-UBM modeling and how LLR score of test speech is predicted. B. Graphical depiction of how we selected word pair points, based on how they relate to the distribution of PreSpeech LLR scores and how close they are to each other in the Duet condition.*

Then, in order to define convergence, we set two criteria that must be fulfilled at the same time. In the first one, we evaluate if the speech of both participants in the dyad becomes more similar in two consecutive words. A threshold on the LLR_{PRE} scores distribution allows us to consider only events for which $LLR_{DUET}(y_x)$ becomes close to zero. In this case, since $log_{x\to 1}(x) = 0$, the value of zero means that both speaker-dependent models contribute equally to the prediction. In other words, the test speech is similar to both speakers. When at least two consecutive $LLR_{DUET}(y_x)$ words fulfill this criterion, we consider them as Convergent (see Figure 3.2B). All other words are considered instead as NoChange. The second criterion controls that the convergent words in the Duet are not random phenomena (Ramseyer and Tschacher (2010); Ward and Litman (2007)). We built 48 surrogate pairs (combining participants of the same gender only) from participants who never interacted with each other in the task. We then ran the same computation as before, on the newly built surrogate pairs. This provides the distribution of surrogate consecutive $LLR_{DUET}(y_x)$ difference scores, which we use to threshold the real consecutive $LLR_{DUET}(y_x)$ difference scores and thus define true convergence. The threshold for both criteria was set at 1.5 SD. This threshold was set so that the words considered as Convergent were at the same time 1) extreme values along the continuum and 2) enough represented to be analyzed separately.

3.1.3 Two studies of Convergence

We tested our convergence algorithm with two separate data-sets. First one in English language (see Section 3.2) and the second one in French language (see Section 3.3). These two studies also employed the same VDT task. In one study (see Section 3.2) we looked at the relationship of Convergence with supra-segmental feature (Fundamental frequency (F0)) and in the other one (see Section 3.3) we explored articulatory counterpart of Convergence. In the following two section we have described in details the materials, methods and results of these two studies. the two studies were presented at the Interspeech conference 2017 and 2018, and published as short peer-reviewed papers in the conference proceedings. Also the English study was extended to find the neural markers during convergence which is presented in Chapter 4.

3.2 The Relationship between F0 Synchrony and Speech Convergence in Dyadic Interaction

A manuscript containing description of the following study has been published in: Interspeech 2017 Article.

Authors: Sankar Mukherjee[1], Alessandro D'Ausilio[1,3], Noël Nguyen[2], Luciano Fadiga[1,3], Leonardo Badino[1].

[1]*Center for Translational Neurophysiology of Speech and Communication, Istituto Italiano di Tecnologia, Ferrara, Italy*

[2]*Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France*

[3]*Section of Human Physiology, University of Ferrara, Ferrara, Italy*

3.2.1 Abstract

Speech accommodation happens when two people engage in verbal conversation. In this paper two types of accommodation are investigated – one dependent on cognitive, physiological, functional and social constraints (Convergence), the other dependent on linguistic and paralinguistic factors (Synchrony). Convergence refers to the situation when two speakers' speech characteristics move towards a common point. Synchrony happens if speakers' prosodic features become correlated over time. Here we analyze relations between the two phenomena at the single word level. Although calculation of Synchrony is fairly straightforward, measuring Convergence is even more problematic as proved by a long history of debates on how to define it. In this paper we consider Convergence as an emergent behavior and investigate it by developing a robust and automatic method based on Gaussian Mixture Model (GMM). Our results show that high Synchrony of F0 between two speakers leads to greater amount of Convergence. This provides robust support for the idea that Synchrony and Convergence are interrelated processes, particularly in female participants.

3.2.2 Materials and method

Participants

For this experiment we recruited 16 native Italian speakers (8 males and 8 females, age: mean ± std; 26 years ± 2.3 years). Before the experiment, subjects were asked to self-rate their English knowledge on a 1-10 scale, including speaking fluency (7.19±1.17), reading

(7.87±1.08), writing (7.31±0.95) and understanding (7.56±1.03). We grouped the subjects in 8 dyads (dyad 1 to 8), 4 female-female and 4 male-male. Before the start of the experiment subjects did not know each other and they did not interact with each other.

Domino list

To build the word chain, we first selected disyllabic words from the WebCelex English lexical database (http://celex.mpi.nl/). Then, we sorted these words depending on spoken frequency (Collins Birmingham University International Language Database - COBUILD). The chain was built using a custom-made iterative algorithm using R https://github.com/sankar-mukherjee/SPIC-dommino. The algorithm started from the highest frequency word and then looked for the next highest frequency item that both fulfilled the rhyming criterion and did not already occur in the chain. This frequency criterion was introduced to avoid low frequency and thus very specialized terminology that would have taxed participant second language skills. By this manner, we generated sequences of at least 200 items that were manually checked to exclude those with crude or offensive words. From the list generated, 200 unique bi-syllabic words were selected for the Verbal Domino task. This sequence is freely available online at hyperscanning.

Procedure

The whole experiment was divided into three parts: Pre, Duet and Post. The verbal domino task was played on the Duet portion. 40 words were randomly selected from the 200-word chain. In Pre and Post, subjects had to read these 40 selected words individually. The Pre and Post parts were before and after the Duet respectively, and were used as baselines.

During the Pre and Post parts, subjects had to read aloud the 40 words presented on a screen one at a time. Between-word switching was controlled by a voice trigger. While one subject was performing this task, the other subject waited nearby. Each subject read 40 words in Pre and 40 in Post sections, for an overall 16x80 = 1280 words.

During the Duet part, the verbal domino task started with one word presented on the screen of one of the two subjects (say subject A) while the other partner (say subject B) was presented with a black screen. Then, when subject A read aloud that word, her/his screen immediately went black and subject B was presented with two words on her/his screen. When subject B read the word fulfilling the rhyming criteria, her/his screen went black and two words appeared on the screen of subject A, until the list ended (Figure 3.1). The voice

onset triggered these changes. The whole experiment was monitored by one experimenter. In case of mistakes, subjects were told to stop and start again from the correct word.

We divided the Duet part in 4 sessions. The 200 selected words were divided into two 100-word chains. For the first two duet sessions, subject A initialized the chains, while for the other two duet sessions, subject B initialized the chains. Each subject read 50x4 = 200 words in the whole duet sessions. This resulted in a total of 3200 words. Only 98 errors out of 3200 words were recorded. The Duet task lasted about 25 minutes. Between Pre, Duet and Post parts as well as between the 4 duet sessions, short breaks were introduced to allow the participants to rest.

The subjects' speech was recorded with a 44100 Hz sampling frequency – using two high-quality microphones (AKG C1000S) connected to an external dedicated audio mixer (M-Audio Fast Track USB II Audio Interface). An adaptive energy-based speech detector (Reynolds (1992)) was used for voice onset detection. All operations were implemented through a Psychotoolbox 3 script running in the Matlab environment.

3.2.3 Acoustic analysis

Pre-Processing

All words in which the voice trigger was incorrect (e.g. breathing, stuttering, etc., resulting in premature triggering or no triggering), or the word response was not correct, were removed for the analysis. This resulted in the removal of 98 out of 3200 words for the Duet and 33 out of 1280 for the Pre and Post parts. For each dyad, an average of 387.75±16.36 words were collected. Periods of silence before and after each word were removed using an energy-based Voice Activity Detector. Then 39 dimension (13 static, 13 delta and 13 delta-delta) MFCCs (Mel Frequency Cepstral Coefficients) were extracted every 5ms from 10ms Hanning windows. Finally, MFCCs were z-score normalized to have 0 mean and 1 standard deviation to mitigate the effects of mismatch between microphones and recording environments.

Convergence calculation

See Section 3.1.2 and Section 3.1.2 for the speaker dependent modeling and convergence computation. Here, the confusion matrix for the cross validation set shows that modelling performance is fairly good (Equal error rate (EER) for the training: 2.26%, validation: 10.55%) as shown in Figure 3.3. After the test, we chose 32-component GMMs.

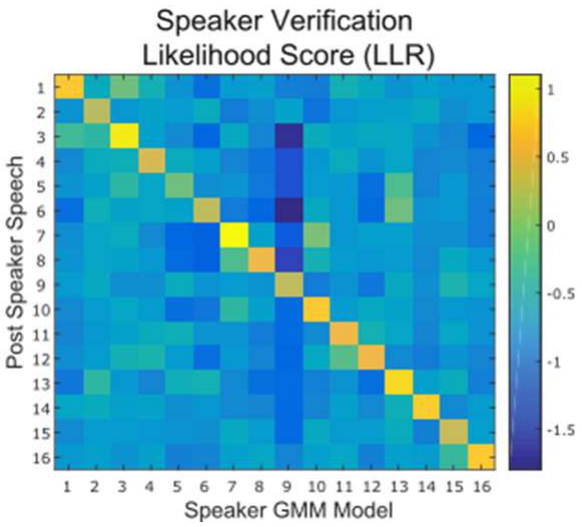

Figure 3.3 *Speaker verification confusion matrix of all the speaker-dependent models against background UBM in the Post data. Here the diagonal positive score line indicates a good MAP adaptation. The diagonal line (top left to bottom right) represents speaker dependent models performance on its their own speaker's speech (which is high compared to the others, suggesting a good speaker dependent model adaptation).*

Synchrony measurement

To measure Synchrony, four prosodic features were extracted: mean fundamental frequency (F0), mean duration, reaction time and mean intensity from each word. Praat software (Boersma and Weenink (2009)) has been used to extract those features. Standard Pearson correlation coefficient $\rho_{xy}\varepsilon[-1,1]$ on two observation sets x and y (belonging to two separate subjects), was computed for each Duet session. This resulted in four correlation coefficients corresponding to each feature for each Duet session.

3.2.4 Results

Convergence results

After fulfilling the two Convergence criteria, the number of Convergence points in each dyad was on average 12.62 % (std 9.02%). The total Convergence points of the whole experiment are shown in Figure 3.4 which indicates that Convergence is sparse. Some dyads had a large amount of convergence while others had a very limited one. Female dyads (FF) converged more than male dyads (MM) (FF 114 and MM 88) which is consistent with previous results (Pardo (2006), Bailly and Martin (2014)).

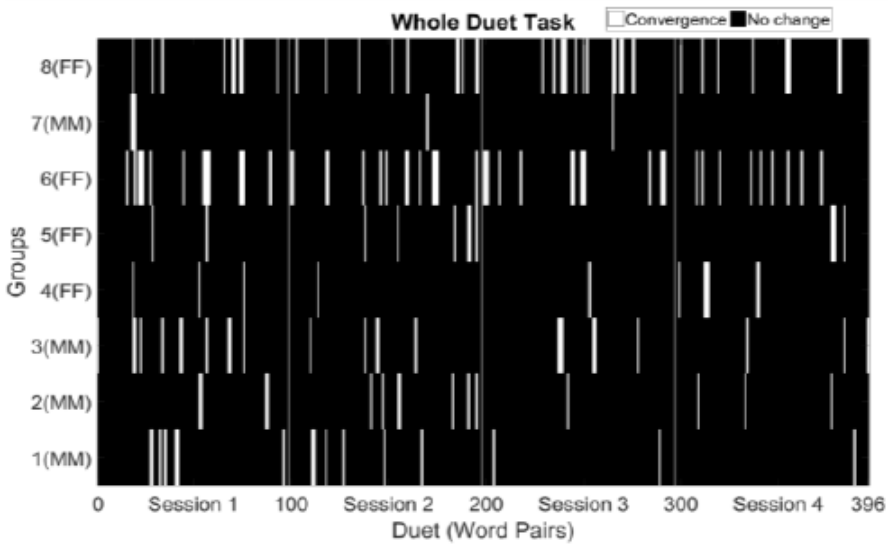

Figure 3.4 *No. of times when dyads converged in the whole experiment. White lines indicate the Convergence moments.*

Results of synchrony parameters during convergence

Differences in mean fundamental frequency (F0), mean duration, reaction time and mean intensity for each word during the Duet sessions between NoChange and Convergence trials are shown in Table 3.1. A two-sided Wilcoxon rank sum tests showed a significant difference in F0 and Intensity between Convergence and NoChange condition.

	Convergence	NoChange	Significance
F0 (Hz)	136.71± 46.58	127.85± 43.98	$P < 0.01$
Intensity(dB)	71.05± 4.29	70.47± 5.07	$P < 0.05$
Duration (ms)	893± 302	838± 232	$P = 0.11$
Reaction Time (ms)	426± 298	427± 262	P=0.69

Table 3.1 *Results of two-sided Wilcoxon rank sum test between Convergence and NoChange condition (mean SD)*

Relation between Convergence and Synchronization

In order to establish the relationship between Synchrony and Convergence, we used Pearson correlation between the number of Convergence points and Synchrony correlation coefficient of each session. This resulted in a highly significant correlation for F0 (Table 3.2). This shows that Synchrony in F0 is associated to Convergence and this result was largely significant, for female-female dyads (Table 3.2).

Features	All dyads		FF dyads		MM dyads	
	CC	sig (P)	CC	sig (P)	CC	sig (P)
F0 (Hz)	0.517	0.002	0.578	0.02	0.199	0.459
Intensity(dB)	0.326	0.07	0.466	0.06	0.064	0.812
Duration (ms)	0.087	0.63	0.336	0.20	-0.293	0.269
Reaction Time (ms)	-0.002	0.99	-0.279	0.29	0.441	0.086

Table 3.2 *Correlation results between Synchrony and Convergence for males (MM) and females (FF) dyads*

3.2.5 Conclusion

Here we show that speech Convergence can be measured using a speaker identification technique during a well constrained task such as the Domino (Bailly and Martin (2014), Mukherjee et al. (2017)) . Importantly, we introduced several analysis features to make the estimation of Convergence more robust. For instance, we tested modelling performance and verified its validity. We also evaluated if Convergence scores were attributable to random fluctuations in the data or were the true effect of dyadic interaction by testing them against surrogate dyads. Results show that the nature of speech Convergence is sparse, i.e., it is not evenly distributed on all the dyads. Some dyads show higher degree of Convergence while others rarely converge at all. A possible factor in this sparseness may be due to subjects' attention, familiarity with the content and their likability towards each other. However small and sparse, Convergence was associated to Synchrony in F0. This is an interesting new addition to the current discussion about the nature of these two complementary aspects of speech accommodation. Our work provides support for the idea that Synchrony and Convergence are interrelated processes, particularly in female dyads. Future work includes testing this speaker identification technique on free flow dialog and in L1-L1 settings.

3.3 Analyzing Vocal Tract Movements during Speech Accommodation

A manuscript containing description of the following study has been published in: Interspeech 2018 Article.

Authors: Sankar Mukherjee[1], Thierry Legou[2], Leonardo Lancia[2], Pauline Hilt[1], Alice Tomassini[1], Luciano Fadiga[1,3], Alessandro D'Ausilio[1,3], Leonardo Badino[1], Noël Nguyen[2].

[1]*Center for Translational Neurophysiology of Speech and Communication, Istituto Italiano di Tecnologia, Ferrara, Italy*

[2]*Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France*

[3]*Section of Human Physiology, University of Ferrara, Ferrara, Italy*

3.3.1 Abstract

When two people engage in verbal interaction, they tend to accommodate on a variety of linguistic levels. Although recent attention has focused on to the acoustic characteristics of convergence in speech, the underlying articulatory mechanisms remain to be explored. Using 3D electromagnetic articulography (EMA), we simultaneously recorded articulatory movements in two speakers engaged in an interactive verbal game, the domino task. In this task, the two speakers take turn in chaining bi-syllabic words according to a rhyming rule. By using a robust speaker identification strategy, we identified for which specific words speakers converged or diverged. Then, we explored the different vocal tract features characterizing speech accommodation. Our results suggest that tongue movements tend to slow down during convergence whereas maximal jaw opening during convergence and divergence differs depending on syllable position.

3.3.2 Materials and method

Domino task

We use The Verbal Domino Task (VDT) (task description see Section 3.1.1). We asked native French speakers to perform a Verbal Domino Task (VDT) (Bailly and Martin (2014), Mukherjee et al. (2017)) with French words (Fig. 3.5B). To build the word chain, we first selected disyllabic words from the Lexique-3 (http://www.lexique.org/) French lexical database. This database was manually checked to exclude crude or offensive words. The

chain was built by using a custom made iterative algorithm, which started from the highest frequency word and then looked for the next highest frequency item, fulfilling the rhyming criteria and no repetitions. In this manner, we generated sequences of 300 unique disyllabic words for the VDT.

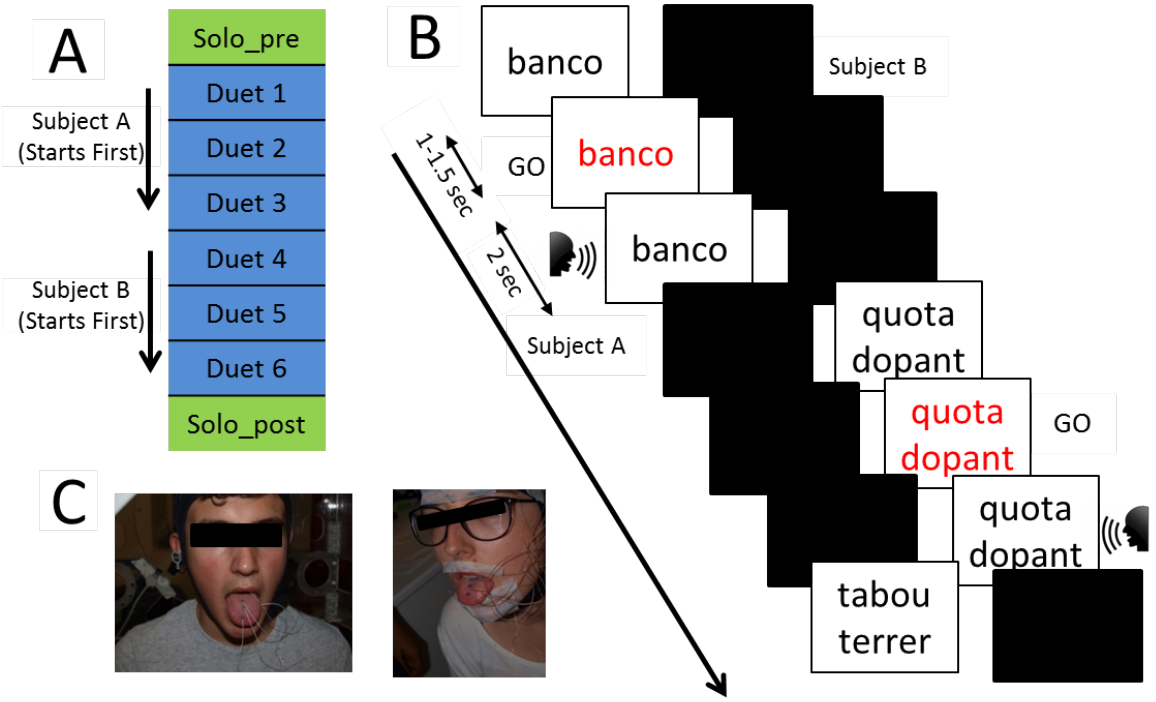

Figure 3.5 *(A) Experimental timeline including the Solo_pre, Solo_post and duet sessions. (B) Sequence of events during the VDT. (C) EMA sensor positioning in one participant.*

Participants

Participants were ten healthy right handed native French subjects (6 males and 4 females, age range 21-26), who did not know each other, composing 5 same-gender dyads. Everyone had self-reported normal hearing. All participants gave informed consent to participate to the experiment. Procedures were approved by the Ethics Committee of the Ferrara University in accordance with the ethical guidelines of the Declaration of Helsinki.

Procedure

The experiment was divided into three main sections (Figure 3.5A). Solo recordings were performed before and after the Duet sessions (Solo_Pre, Solo_Post). The Solo required subjects to read 60 words to establish a subject-wise baseline. These words were phonetically

balanced and selected from the 300 words chosen for the VDT. During the Solo the other participant was listening to classical music through headphones.

During the Solo, one word at time was presented on a black screen and after a variable delay of (1 - 1.5 s) a GO signal instructed the subject to read it aloud. This random delay was introduced to avoid anticipation and entrainment to the rhythm of presentation. For the same reason, trials presentation was intermingled with random delays (2-2.5 s). Since each subject completed 60 words in the Solo, we collected a dataset of 1200 words. The Solo sessions lasted about 4 minutes.

In Duet, the task started with one word presented on the screen of one subject (Subject A), while the other participant's screen was blank (Subject B). Subject A waited for the GO signal (delay of 1-1.5 s) and had 2 seconds to respond. At the end of the trial, Subjects A's screen went blank and two words appeared on Subject B's screen. Now, Subject B had to choose which word to read aloud as only one was complying with the rhyming rule. This chain of events continued until the end of the list.

The 300 words of VDT were divided into 3 lists of 100 and repeated twice so that the Duet part was composed by six separate sessions. In each session, the two speakers read 50 words each, summing up to 300 words per speaker and thus resulting in a total of 3000 words. The duet sessions lasted about 30 minutes. The VDT was implemented in a Psychotoolbox 3 script running in the Matlab environment.

Speech was recorded by two high-quality microphones (AKG C1000S) and the speech data were digitized and acquired by an acquisition CPU (16 bit, stereo, 22050Hz sampling frequency). Both signal went through an external dedicated amplifier (MMX-11USB 2ch audio mixer) and acquired with a A/D acquisition board (MC measurement computing USB-1608GX-2AO).

Articulatory data was recorded with two EMA systems. The first one was an NDI (Northen Digital Instruments, Canada; sampling frequency, 400 Hz) and the second one was an AG501 (Carstens Medizinelektronik GmbH; sampling frequency, 256 Hz). Seven 5-degrees-of-freedom (5-DOF, x,y,z, pitch and roll) sensor coils were glued on the Upper Lip (UL), Lower Lip (LL), Upper Incisor (UI), Lower Incisor (LI), tongue tip (TT), tongue middle (TD) and tongue back (TB). For head movement correction, a 6-DOF sensor coil was fixed on the bridge of a pair of glasses worn by the participants (Figure. 3.5C).

3.3.3 Pre-Processing

Acoustic Pre-Processing

Incorrect trials (e.g., wrong pronunciation, wrong choice of words, about 3.1%) were excluded from the analysis. Periods of silence were discarded using an energy-based Speech Activity Detector. We then computed MFCCs (Mel Frequency Cepstral Coefficients) by segmenting the data into 25ms frames (10ms overlap) with a Hamming window. The short-time magnitude spectrum, obtained by applying FFT, was passed to a bank of 30 Mel-spaced triangular bandpass filters, spanning from 0 Hz to 3,800 Hz. The output of the 30 filters were transformed into 12 static, 12 velocity and 12 acceleration MFCCs with the 0'th coefficients resulting in 39 MFCC dimensions in total.

Articulatory Pre-Processing

Articulatory data from both EMA systems, was down-sampled at 100Hz. We removed from the dataset all words for which one or more sensors were detached (Convergence: 36.14± 23.36% ; NoChange: 32.27± 16.44%; Divergence: 42± 20.5%). Vocal articulator trajectories (x, y, z positons of the sensor coils) were filtered using an adaptive median filter (10-50ms window) and further smoothed using a 20Hz cutoff elliptic low-pass filter. Coils rotation was ignored. From the x-y midsagittal coil positions we extracted six vocal tract features: lip aperture (LA) (equation 1), lip protrusion (PRO) (equation 2), jaw opening (JO) (equation 3), tongue tip constriction degree (TTCD), tongue blade constriction degree (TBCD) and tongue dorsum constriction degree (TDCD).

$$LA = |UL_y - LL_y| \tag{1}$$

$$PRO = |UL_x - LL_x| \tag{2}$$

$$JO = |UI_y - LI_y| \tag{3}$$

TTCD, TBCD and TDCD are the Euclidean distance of TT, TB and TD to the curve of the palate on the midsagittal plane. To assess how fast these vocal tract features are changing, the velocity of these features was also computed. Since words in the VDT are disyllabic, we expected two local maxima for each word in the jaw opening trajectory, which roughly correspond to the open configuration of the vocal tract for the two vowels. For this reason, we computed the maximum jaw opening of the two syllables separately (JO_Syl_1; JO_Syl_2), and their average (JO_Syl_1&2).

Convergence and Divergence calculation

See Section 3.1.2 for how the GMM-UBM modelling is done. Here a 256-component GMM was chosen as it had the lowest Equal error rate (EER) and showed a good modelling performance of the confusion matrix for the cross validation set (EER=4%; Figure 3.6).

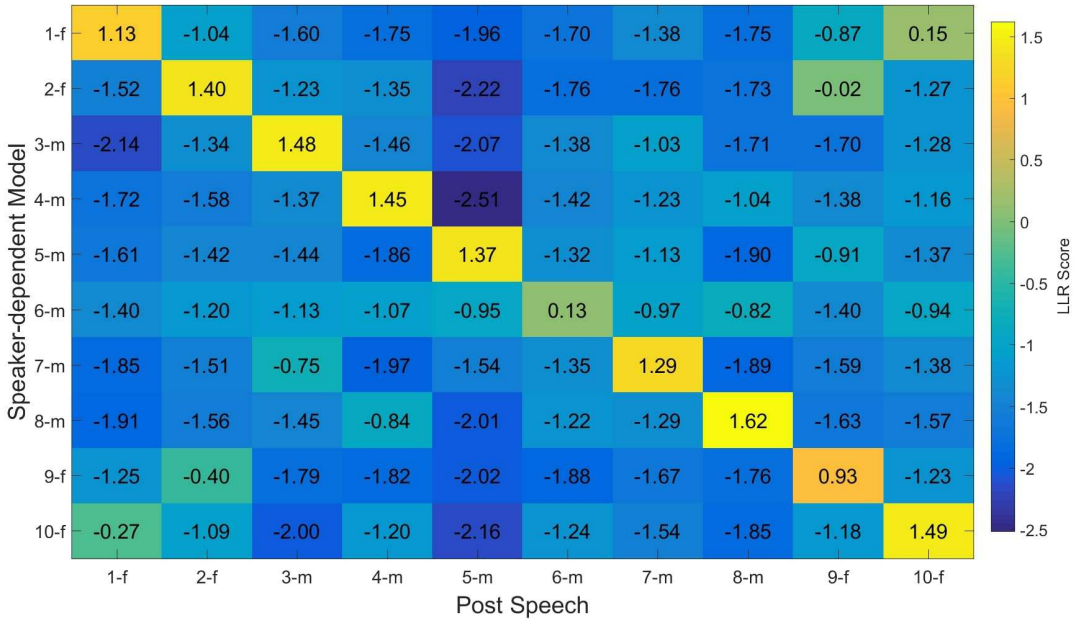

Figure 3.6 *Speaker verification confusion matrix of all the speaker-dependent models against background UBM in the Solo_Post. Here the diagonal positive score line indicates a good MAP adaptation. The numbers are the subtraction between speaker-dependent model and UBM model and is represented by Log-Likelihood ratio (LLR).*

Convergence and Divergence was measured for each word level using the same procedure described in (Mukherjee et al. (2017)). First, we tested each speaker-dependent model on the 60 Solo-pre words. Then we measured the posterior probability score for each word during interaction. We then set a threshold of 1.5 standard deviations (STD) based on the distribution of the prediction scores at baseline [as in Mukherjee et al. (2017)]. If the word was predicted (i.e. minimum posterior probability) by the speaker-own model, we labeled that word as NoChange. If the word was predicted by its partner model we labeled that word as Convergent. Otherwise, when neither own or partner model predicted the word, we labeled that word as Divergent. In this study, the convergence calculation is somewhat different than previous study where we strictly measure symmetric convergence (when both consecutive word become similar to each other). Here we only measure asymmetric convergence (only

one word is well predicted by other model). This is motivated due to small amount of data available in this study.

3.3.4 Results

Convergence and Divergence frequency

The total number of Convergence, Divergence and NoChange during the whole interaction is shown in Figure 3.7. Convergence and Divergence are variable phenomena because some dyads show a large amount of convergence while others much less (Mukherjee et al. (2017), De Looze et al. (2014)). The Female dyads (FF) converged more than male dyads (MM) (FF 25% and MM 7%) which is consistent with previous results (Bailly and Martin (2014), Mukherjee et al. (2017)). A one-way repeated-measures ANOVA with the sessions (6 levels) as within-subject factor did not reveal any significant effect of Convergence ($F(5,45) = 0.78$, $p=0.56$). The same analysis on Divergence showed no significant effects ($F(5,45) = 0.69$, $p=0.63$) indicating that the amount of Convergence and Divergence did not change significantly across the experimental blocks.

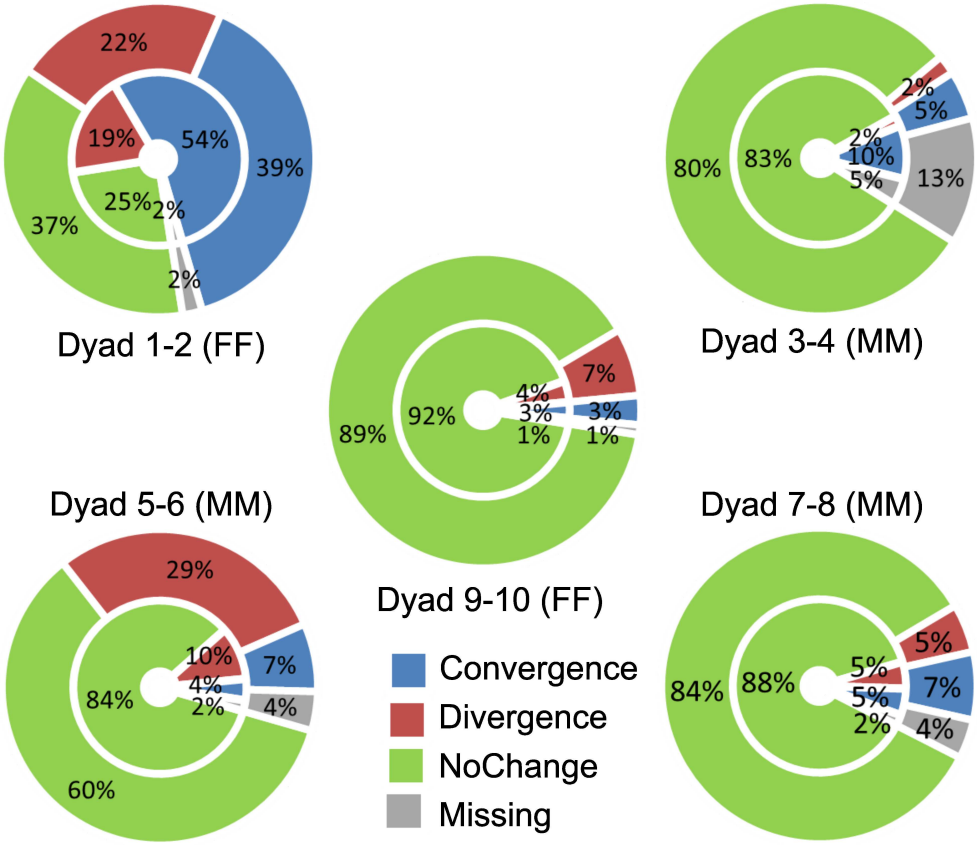

Figure 3.7 *No. of times each participant converged or diverged during the experiment.*

Acoustic features

Four speech acoustic features, F0, F1, F2 and Intensity were extracted from the audio recordings using Praat software (Boersma and Weenink (2009)). First, we averaged within each word and then within subjects. A two-tailed t-test, on z-scored values, was used to explore differences between Convergence and NoChange or Divergence and NoChange. Results show (Table 3.3) that intensity was significantly different ($t(9)= 4.93$; $p<0.0001$) during Convergence compared to NoChange and during Divergence compared to NoChange ($t(9)= -2.81$; $p=0.02$).

	Convergence	NoChange	Divergence	t-test p-value	
	(± STD)	(± STD)	(± STD)	Con-NoCh	Div-NoCh
F0 (Hz)	135±38	134±40	136±36	0.37	0.3
F1 (Hz)	83±22	69±28	80±24	0.56	0.09
F2 (Hz)	272±69	244±84	281±75	0.28	0.11
Intensity (dB)	58±10	57±10	58±10	0.0001	0.02

Table 3.3 *Results of two-sided student t-test (Convergence Vs. NoChange and Divergence Vs. NoChange).*

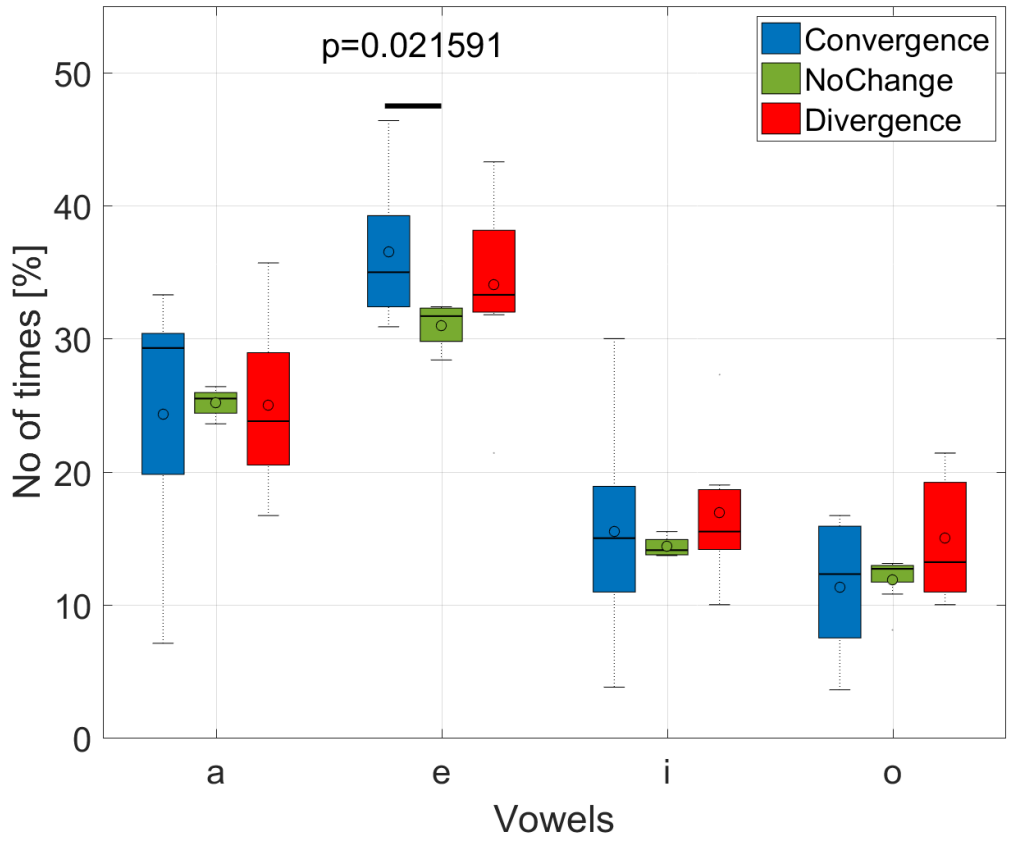

Figure 3.8 *Most frequent vowel distribution [%] in the three conditions for all subjects.*

Vocal Tract features characteristics during accommodation

In the construction of our VDT list we included 11 different vowels (lexical code: a, e, i, o, u, y, O, E, @, 2 for details, see http://www.lexique.org/). However, given that for some subjects we observed relatively few instances of Convergence or Divergence we ended up with a smaller set of vowels in these categories. Therefore, to avoid any biased comparison, when

analyzing articulatory data, we excluded NoChange words containing very rare vowels in Convergence or Divergence. A NoChange word was included if all its vowels were present in at least 5% of the convergent or divergent words. This resulted in four vowels (/a/,/e/,/i/,/o/) whose distribution in the three different conditions is shown in Figure 3.8. A two-tailed t-test on z-scored values was used to explore the differences between Convergence and NoChange or Divergence and NoChange for each vowel. Results show that only for /e/, there was a significant difference between Convergence and NoChange $(t_{(9)} = 2.77; p = 0.022)$. This means that the following analyses on articulatory data were run on all 4 vowels (/a/,/e/,/i/,/o/) as well on (/a/,/i/,/o/).

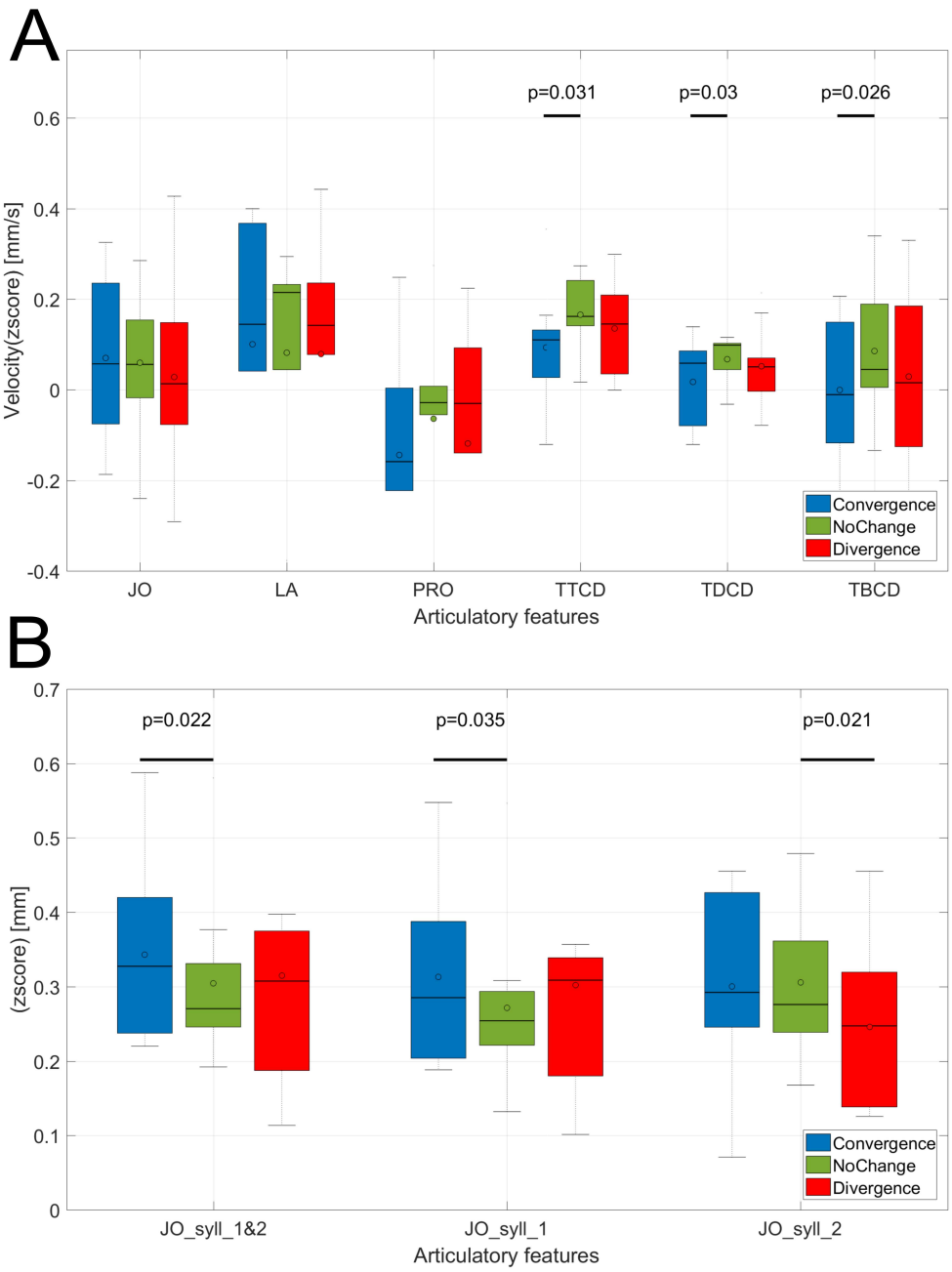

Figure 3.9 *(A) Whole-word articulatory data changes across conditions. The t-tests showed significant differences in the velocity of vocal tract tongue features (TTCD, TDCD, TBCD) between Convergence and NoChange (orizontal lines). (B) Syllable level differences in maximal jaw opening. JO_syll_1&2 and JO_syll_1 are significantly different in Convergence with respect to NoChange whereas JO_syll_2 is significantly different in Divergence with respect to NoChange.*

Vocal tract features were first averaged within each word then within each subject. A two-tailed t-test performed on z-scored values, was used to explore differences between Convergence and NoChange and between Divergence and NoChange. Results showed that velocity of the three vocal tract tongue features were significantly different in Convergence and NoChange conditions ($TTCD: t_{(9)} = -2.55; p = 0.031; TBCD: t_{(7)} = -2.82; p = 0.026; TDCD: t_{(8)} = -2.63; p = 0.03$) (Figure 3.9A) demonstrating that during Convergence speakers move their tongue more slowly than in NoChange.

Moreover, maximal jaw opening was significantly modulated in Convergence Vs. NoChange (Figure 3.9B; JO_syll_1&2: $t_{(9)} = 2.79; p = 0.021$; JO_syll_1: $t_{(9)} = 2.47; p = 0.03$) and maximal jaw opening of the 2nd syllable was significantly different in Divergence Vs. NoChange (JO_syll_2: $t_{(9)} = 2.75; p = 0.022$). Larger values in these features means that during Convergence speakers opened their jaw more than in NoChange, especially in the first syllable. Differently, in the second syllable the pattern of jaw opening was reversed and this was true for Divergence only. The same pattern is observed when removing the /e/ vowel from the dataset. Maximal jaw opening was significantly modulated in Convergence Vs. NoChange conditions (JO_syll_1&2: $t_{(9)} = 3.13; p = 0.012$; JO_syll_1:$t_{(9)} = 2.98; p = 0.015$) and maximal jaw opening of the 2nd syllable was significantly different in Divergence Vs. NoChange (JO_syll_2: $t_{(9)} = 3.23; p = 0.01$).

3.3.5 Conclusion

Speech convergence is the phenomenon by which some participants in a dialogue tend to naturally align with each other in their phonetic characteristics. Here, we demonstrated the robustness of the automatic phonetic convergence detection method we already presented in (Mukherjee et al. (2017)). In fact, as shown in Figure 3.6 and Figure 3.7, our results were similar to those of our previous study. It is worth mentioning that the present dataset is characterized by relevant differences including participants' native language, the language of the word list, the word chain length, the pacing of VDT (self-paced as opposed to externally-paced) and the number of participants.

Besides, we also show preliminary but compelling results indicating that accommodation phenomena occur at the level of articulatory features. When we analyzed average velocity profiles at the whole-word level, we found that speakers slow-down their tongue movements during Convergence. Instead, when separating the two syllables of each word, we observed an interesting pattern of jaw maximal opening. In fact, the first syllable shows larger values during Convergence, whereas the second syllable smaller values for Divergence. Note that

the first syllable is the one shared with the preceding word of the phonetic dyadic context (i.e. the word just uttered by the partner). Most importantly, the VDT rhyming rule forces subjects to focus their attention to the last syllable they heard to match it to the first they have to articulate. Interestingly, the opposite result we found for the second syllable could be explained by the fact that, for the speaker, this syllable does not have to comply with any specific rule. However, due to the variability of accommodation phenomena (De Looze et al. (2014)), results could in part be driven by dyads showing greater effects.

The present work starts exploring the articulatory counterpart of phonetic convergence. Future experiments will need to acquire larger number of dyads and eventually explore if at the single syllable level there are critical articulatory features (Stevens (1989)) which are more or less robust to accommodation phenomena occurring during speech interactions.

Chapter 4

The Neural Oscillatory Markers of Phonetic Convergence during Verbal Interaction

A manuscript containing description of the following study has been published in: Human brain mapping Journal 2018.

Authors: Sankar Mukherjee[1], Leonardo Badino[1], Pauline Hilt[1], Alice Tomassini[1], Alberto Inuggi[4], Luciano Fadiga[1,3], Noël Nguyen[2], Alessandro D'Ausilio[1,3].

[1]*Center for Translational Neurophysiology of Speech and Communication, Istituto Italiano di Tecnologia, Ferrara, Italy*

[2]*Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France*

[3]*Section of Human Physiology, University of Ferrara, Ferrara, Italy*

[4]*Center for Human Technologies, Istituto Italiano di Tecnologia, Genova, Italy*

4.1 Abstract

During a conversation, the neural processes supporting speech production and perception overlap in time and, based on context, expectations and the dynamics of interaction, they are also continuously modulated in real time. Recently, the growing interest in the neural dynamics underlying interactive tasks, in particular in the language domain, has mainly tackled the temporal aspects of turn-taking in dialogues. Besides temporal coordination, an under-investigated phenomenon is the implicit convergence of the speakers towards a shared phonetic space. Here, we used dual electroencephalography (dual-EEG) to record brain

signals from subjects involved in a relatively constrained interactive task where they were asked to take turns in chaining words according to a phonetic rhyming rule. We quantified participants' initial phonetic fingerprints and tracked their phonetic convergence during the interaction via a robust and automatic speaker verification technique. Results show that phonetic convergence is associated to left frontal alpha/low-beta desynchronization during speech preparation and by high-beta suppression before and during listening to speech in right centro-parietal and left frontal sectors, respectively. By this work, we provide evidence that mutual adaptation of speech phonetic targets, correlates with specific alpha and beta oscillatory dynamics. Alpha and beta oscillatory dynamics may index the coordination of the "when" as well as the "how" speech interaction takes place, reinforcing the suggestion that perception and production processes are highly interdependent and co-constructed during a conversation.

4.2 Materials and Methods

4.2.1 Participants

16 healthy participants took part in the task (8 females; age, 26±2.3 years; mean±SD). All participants were right-handed with the exception of one male. They were all native Italian speakers unaware of the purpose of the experiment. We asked dyads of participants to perform the task in English with a view to augmenting the likelihood for these participants to converge towards each other at the phonetic level. Previous studies have shown that the use of a non-native language induces greater phonetic convergence (Gambi and Pickering (2013), Trofimovich and Kennedy (2014)) and greater sensorimotor compensatory activities while listening, compared with the native language (Schmitz et al. (2018)). We asked participants to self-rate their English reading (7.87±1.08), writing (7.31±0.95), understanding (7.56±1.03) and fluency (7.19±1.17, mean±SD) capabilities on a 1-10 scale. Self-reported proficiency in a non-native language is routinely used in experimental psycholinguistic studies, and has been repeatedly shown to be closely related to a large variety of objective measures of proficiency (Marian et al. (2007); Gollan et al. (2012)). Participants were grouped into 8 pairs (pair 1 to 8) before the start of the experiment. Groups consisted in 4 male-male and 4 female-female pairs. Within each pair, speakers will be referred to as A and B in what follows. The participants did not know each other nor interacted before the experiment. The study was approved by the local Ethics Committee and a written informed consent was obtained from the subjects according to the Declaration of Helsinki.

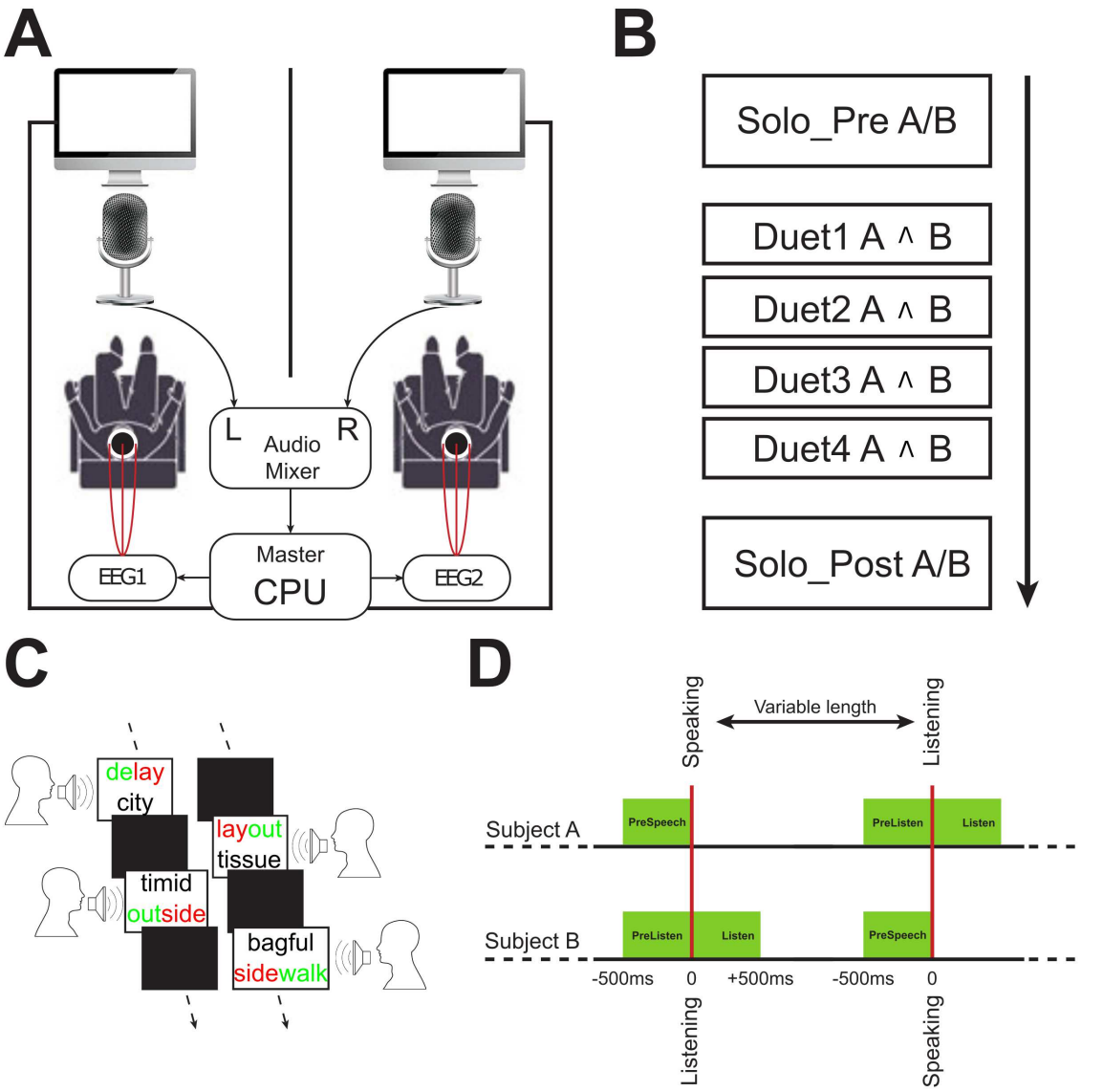

Figure 4.1 *A. Graphical depiction of the experimental setup. B. Schematic illustration of the experimental time line including two solo recordings before and after four duet sessions. Here, ab indicates recording the speech of a and b separately, whereas aB̂ means recording of both together. C. An example of the sequences of words produced during the verbal domino task by the two speakers. D. Schematic illustration of the triggers for the EEG signals (red) and the temporal windows considered for the analyses (green).*

4.2.2 Task and stimuli

We use The Verbal Domino Task (VDT) (task description see Section 3.1.1). We are aware that the VDT shows important differences with respect to a natural conversation, but the

no-overlap constraint is essential to temporally isolate brain activities associated to speaking and listening (see section on EEG data analysis).

During the task, the two speakers are alternatively presented with two words on a screen and must read aloud the word whose initial syllable coincides with the final syllable of the word previously produced by the other speaker. For example, on having heard Speaker A pronouncing the word "delay", Speaker B is offered to choose between "layoff" and "tissue" and is expected to pronounce "layoff" (see Figure 4.1C). The task is collaborative in that participants have to make the right choices for them to reach a common goal, i.e. to jointly go through the domino chain up to the end. The choice of presenting the words visually, rather than allowing self-generation, was required to avoid participants using (a variable amount of) time "searching" for possible candidate words. In fact, this would have added a fundamental memory retrieval component that couldn't be controlled. Word self-generation can also introduce frequent stops in the chain. Moreover, the presentation of two alternatives, instead of a single option, forced participant to actively listen to their partner to select the correct word.

To build the word chain, we first selected disyllabic words from the WebCelex English lexical database (http://celex.mpi.nl/). Then, we sorted these words depending on spoken frequency (Collins Birmingham University International Language Database - COBUILD). The chain was built using a custom-made iterative algorithm in Matlab (Matlab 2016; The MathWorks). The algorithm started from the highest frequency word and then looked for the next highest frequency item that both fulfilled the rhyming criterion and did not already occur in the chain. This frequency criterion was introduced to avoid low frequency and thus very specialized terminology that would have taxed participant second language skills. By this manner, we generated sequences of at least 200 items that were manually checked to exclude those with crude or offensive words. We finally selected one sequence of 200 unique disyllabic words. This sequence is freely available online at hyperscanning.

4.2.3 Procedure

For each pair, we recorded the speech and EEG signals of the two participants simultaneously. They were sitting in a quiet room with the experimenter monitoring the whole experiment.

The experiment was divided into three main sections (Figure 4.1B). Solo recordings were performed before (Solo_Pre) and after (Solo_Post) the Duet session. Solo data were needed to establish a participant-wise baseline. During the Solo task, the other participant wore noise insulating headphones. The Solo task required participants to pronounce 40 words

randomly selected from the 200-word set of the domino chain. Words were presented one after the other on a black screen and subjects had to read them out. Voice onset for each word triggered the appearance of the following word's written form with a random delay (0.6 - 1.5 s). This random delay was introduced to avoid anticipation and entrainment to an external rhythm of presentation. Each Solo session lasted about 2 min. The resulting dataset was made up of a total of 1280 words (including all subjects).

In the Duet session, the task started with one word visually presented on the screen of one of the two participants (e.g., Participant A), while the other participant's screen was blank. When Participant A read the word aloud, her/his screen went immediately blank and two words appeared on Participant's B screen. Participant B chose that of the two words which best fulfilled the rhyming criterion and, as soon as she/he read that word aloud, her/his screen went blank and two other words appeared on Subject A's screen. This cycle was repeated until the end of the list.

The 200-word Verbal Domino chain was divided into 2 lists of 100 words, both repeated twice so that the Duet part was composed by four separate blocks. During Duet blocks 1 and 2, Participant A started the Verbal Domino, whereas for Duet blocks 3 and 4, Participant B was the first. In each of the four blocks, the two participants spoke 50 words each, summing up to 200 words per participant for the Duet part and resulting in a total of 3200 words. The duet session lasted about 25 min.

After each of the four duet blocks, participants could rest for about 2 min. After the resting period, we recorded 1 min of baseline EEG activity by asking participants to fixate a cross at the center of the black screen. The whole experiment was monitored by one experimenter who checked that participants correctly performed the tasks. Whenever one participant chose the wrong word in a Duet session, the task was halted. The experimenter would then tell the other participant which word he/she was to continue with.

4.2.4 Data acquisition

Neural activities were recorded by a dual-EEG recording setup consisting of two Biosemi Active Two systems (Amsterdam, Netherlands), each with 64 channels mounted onto an elastic cap according to the 10-20 international system. The left mastoid was used as online reference. EEG data were digitized at 1024 Hz. Electrodes' sensitivity, as expressed by the Biosemi hardware, was kept below 20 μV.

Three Central Processing Units (CPU) were used for the whole experiment: one master CPU for the control of experimental events, and two other CPUs to acquire EEG data from

the two Biosemi systems. The master CPU controlled the presentation of the stimuli and the detection of the voice onset, and sent triggers to the two other CPUs via a parallel port. All of the operations of the master CPU were controlled with Psychtoolbox-3 running in the Matlab environment.

Speech data was also recorded by the master CPU (16 bits, stereo, 44100 Hz sampling frequency) using two high-quality microphones (AKG C1000S) and an external dedicated amplifier (M-Audio Fast Track USB II Audio Interface). Voice onset was detected using an adaptive energy-based speech detector Reynolds et al. (2000). This detector tracks the noise energy floor of the input signal and labels that signal as speech if any feature vector (computed over a 10-ms window) energy exceeds the current noise floor by a fixed energy threshold. For each participant, and before the experiment, we set up this threshold by controlling the microphone channel amplifier gain (M-Audio Fast Track USB II Audio Interface).

4.3 Automatic analysis of convergence

4.3.1 Acoustic data pre-processing

Trials in which automatic voice trigger was incorrect (e.g., premature triggering by noise or failure of the triggering system because verbal response was not loud enough, wrong choice of words) were excluded from the analysis. This resulted in the removal of 33 out of 1280 Solo words, and 98 out of 3200 Duet words. As a result, we collected an average of 387.75±16.36 (mean±SD) words for each dyad in the Duet session and an average of 68.19±18.76 (mean±SD) words per participant in the Solo session. The number of incorrect word was small 6.375 (SD=5.677) and there was no correlation between number of incorrect word choice and number of convergence points (R = 0.25; p=0.54). The low number of incorrect word choices indicates that the participants had little or no difficulty in performing the task in English.

Acoustic feature extraction was performed as follows. Periods of silence were discarded using an energy-based Speech Activity Detector. We then extracted MFCCs (Mel Frequency Cepstral Coefficients), which are a short-term power spectrum representation of sounds, based on a linear cosine transform of log power spectrum on a nonlinear mel-scale of frequency, widely used in speech technology applications (Kinnunen and Li (2010)). This choice allowed us not to make a priori assumptions about which subset of acoustic features

is associated with between-speaker phonetic convergence, and instead to exploit the entire informational content of the acoustic spectrum.

MFCCs were derived using the speech signal, segmented into 10 ms frames (5 ms overlap) and a Hamming window. The short-time magnitude spectrum, obtained by applying fast Fourier transform (FFT), was passed to a bank of 32 Mel-spaced triangular bandpass filters, spanning the frequency region from 0 Hz to 3800 Hz. The outputs of all 32 filters were transformed into 12 static, 12 velocity, and 12 acceleration MFCCs with the 0'th coefficients resulting in 39 MFCC dimensions in total. Velocity and acceleration features were included to incorporate information about the way the 12 static vectors varied over time. Finally, the distribution of these cepstral features was wrapped (Pelecanos and Sridharan (2001)) per word to the standard normal distribution to mitigate the effects of mismatch between microphones and recording environments.

4.3.2 GMM-UBM

See Section 3.1.2 for how the GMM-UBM modelling is done. Here a 32-component UBM was trained with the pooled Solo_Pre speech data of all the participants (a total of 124068 speech frames).

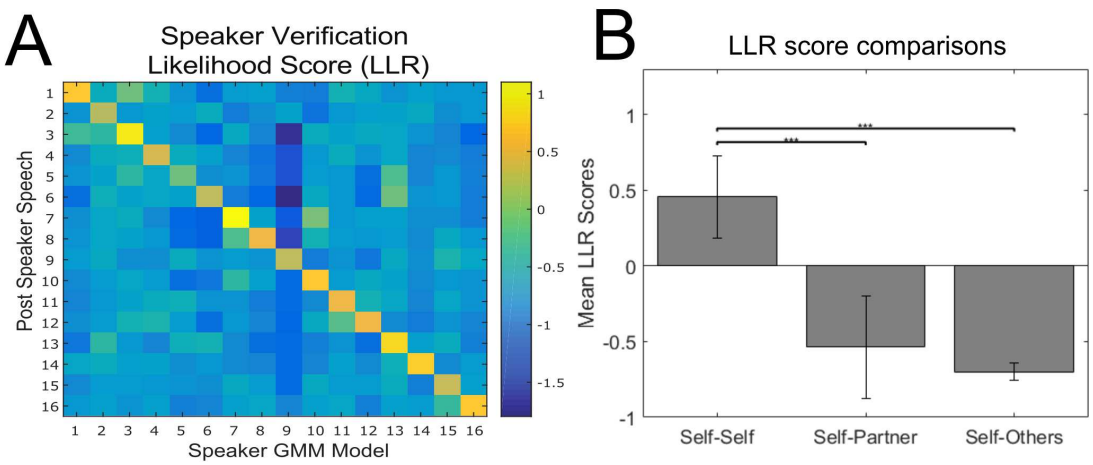

Figure 4.2 *A. Speaker verification confusion matrix of all the speaker-dependent models against background UBM in the Solo_Post. Here the diagonal positive score line indicates a good MAP adaptation. (This figure is same as Figure 3.3). B. The bar plot represents the same data in plot a, only by grouping the data within participant, within dyads, and across dyads. This plot shows the obvious better fit of the model when it is tested on the same speaker (self-self) and far lower performance when testing on another one.*

A cross-validation technique was used to choose the optimum hyper-parameter settings. Solo_Post speech data were used as a validation set, and each speaker-dependent model's performance was verified against the UBM model (Figure 4.2A). Furthermore, to verify if there had been any pre-post change in the acoustic properties of speech, due to the duet interaction, we further grouped the data within participant, within dyad, and across dyads. In principle, if the interaction has been able to affect the phonetic fingerprint of the participants, the cross-validation performance should be better within than across dyads. Differences were verified using Bonferroni corrected paired t-tests.

4.3.3 Phonetic Convergence computation

To see how phonetic convergence was computed see Section 3.1.2.

4.4 EEG data analysis

4.4.1 Preprocessing

EEG data were analyzed using the EEGLAB software (Delorme and Makeig (2004)), the Fieldtrip toolbox (Oostenveld et al. (2011)) and custom-made MATLAB code. EEG data were first bandpass-filtered (two-pass Butterworth filter, fourth-order) between 0.1 and 40 Hz and then down-sampled to 256 Hz. Data recorded during speech production were discarded from the analysis because of strong speech-related artifacts. The remaining EEG data were first visually inspected for bad channels and/or artifacts in the time domain. Noisy channels were interpolated using a distance-weighted nearest-neighbor approach. To identify and remove artifacts related to participants' eye movements, eye blinks, and muscle activity, we used Independent Component Analysis (ICA) according to a consolidated approach (Delorme and Makeig (2004)). Finally, data were re-referenced using a common average reference over all electrodes.

EEG analyses of the Duet condition were constrained by the self-paced structure of the task which, by definition, made the timing of the events of interest (i.e., listening and speaking) not under experimental control. Given that speaking and listening phases alternated at a fast and variable rate, we restricted our analyses to short epochs of 500 ms that allowed us to avoid 1) artifacts due to speech production, and 2) temporal superposition of speech- and listening-related neural processes.

We defined three 500 ms epochs of interest (Figure 4.2D):

- Before speech production (PreSpeech): from -500 to 0 ms relative to (one's own) voice onset.
- Before speech listening (PreListen): from -500 to 0 ms relative to (the partner's) voice onset.
- During speech listening (Listen): from 0 to +500 ms relative to the partner's voice onset.

4.4.2 Time-frequency analysis

Time-frequency representations (TFRs) for the three different epochs (PreSpeech, PreListen, Listen) were calculated using a Fourier transform approach applied to short sliding time windows. All the epochs were zero-padded to avoid edge artifacts and spectral bleeding from contiguous EEG signal possibly contaminated by speech-related artifacts. The power values were calculated for frequencies between 8 and 40 Hz (in steps of 2 Hz) using a Hanning-tapered adaptive time window of four cycles ($\Delta t=4/f$) that was advanced in steps of 50 ms. This procedure results in a frequency-dependent spectral smoothing of $\Delta f=1/\Delta t$. As a consequence of analyzing 500-ms epochs (see above) using four-cycle time windows, the lowest frequency for which we could derive a power estimate (based on the entire epoch) was 8 Hz. In other terms, the relatively fast and self-paced nature of our task did not permit a reliable estimation of the power of slow oscillations (in the delta and theta frequency range).

4.4.3 Statistical analysis

Statistical analysis was performed on the whole-brain oscillatory power (between 8 and 40 Hz). To evaluate statistically whether Convergence and NoChange data showed a difference in oscillatory power, we performed a group-level nonparametric cluster-based permutation test (Maris and Oostenveld (2007)), separately for each epoch of interest (PreSpeech, PreListen, Listen). For every sample (here defined as [channel, frequency, time] triplet), a dependent-sample t value was computed. All samples for which this t value exceeded an a priori decided threshold (uncorrected $p < 0.05$) were selected and subsequently clustered on the basis of temporal, spatial and spectral contiguity. Then, cluster-level statistics was computed by taking the sum of t-values in each cluster. The cluster yielding the maximum sum was subsequently used for evaluating the difference between the two data-sets (with the maximum sum used as test statistic). We randomized the data across the two data-sets, and for each random permutation (10000 iterations), we calculated again the test statistics in the same

way as previously described for the original data. This procedure generates a surrogate distribution of maximum cluster t-values against which we can evaluate the actual data. The p-value of this test is given by the proportion of random permutations that yields a larger test statistic compared to that computed for the original data.

4.5 Results

4.5.1 Behavioral results and GMM-UBM Performance

Turn-taking reaction time (RT) during the Duet sessions, measured as the time elapsed between visual presentation of words and voice onset, did not differ between NoChange (427±262 ms, mean±SD) and Convergence (426±298 ms, mean±SD) trials (Wilcoxon rank sum test: z=-0.46, p=0.64). Turn-taking was self-paced and thus RTs are also a direct measure of the turn-taking pace and the rhythm established by the dyad. This analysis suggests that Convergence and NoChange trials share similar temporal turn-taking dynamics. Furthermore, the Pearson correlation between turn-taking RT and the number of convergence points did not show any significant relationship (R = 0.57; p=0.14).

As far as the GMM-UBM modelling was concerned, we verified each speaker-dependent model's performance against the UBM model (Reynolds et al., 2000). The confusion matrix for the Post speech showed that modelling performance was good. This is measured using the equal error rate (EER) which indicates that the proportion of false acceptances is equal to the proportion of false rejections. The lower the EER value, the higher the accuracy of the classifiers. EER for the training is 2.26% and validation is 10.55% as shown in Figure 4.2A. As visible in Figure 4.2A, the speaker-dependent model's performance was far better when tested on the same subject (self-self). Instead, testing on the Solo_Post of another speaker, led to critically lower performances (self-self vs. self-partner, t(30)=9.1; p<0.00001; self-self vs. self-other, t(30)=16.63; p<0.00001; self-partner vs. self-other, t(30) =1.89; p=0.07). This result suggests that the VDT did not affect the phonetic fingerprint of the participants.

The proportion of convergence points that fulfilled both convergence criteria in each dyad (see section on "automatic analysis of convergence") was on average 12.62±9.02% (mean±SD). The large inter-individual variability is consistent with previous reports showing that convergence may not equally distribute across dyads (Pardo et al. (2017)). It has also been reported that female dyads tend to converge more than male dyads (Pardo (2006); Bailly and Martin (2014)). However, gender differences may potentially be affected by task and psycho-social factors and indeed in our data, female (114 words in total; 22±15.56,

mean±SD) and male dyads (88 in total; 28.5±22.13, mean±SD), did not show any statistical difference in convergence (Wilcoxon rank sum test, p=0.68).

Finally, some studies in this area have modeled convergence as a linear process, i.e., it grows as the conversation proceeds (Suzuki and Katagiri (2007); Natale (1975)). However, subjects do not remain involved to the same degree over the whole course of a conversation, suggesting that convergence can be a dynamic phenomenon (Edlund et al. (2009); De Looze et al. (2014)). The one-way repeated-measures ANOVA with session (Duet1, Duet2, Duet3, Duet4) as within-subject factor did not reveal any significant effect ($F(3,21) = 2.63$, $p=0.08$), offering no conclusive evidence for a change over the experimental blocks.

4.5.2 EEG Results

The comparison between the oscillatory power in the Convergent and NoChange data-sets for the three epochs of interest showed significant results that are summarized in Figure 4.3. Specifically, in the epoch preceding speech onset (PreSpeech), oscillatory power in the alpha/low beta band (9-17 Hz) was attenuated for Convergence compared to NoChange trials ($p=0.035$; see Figure 5). This alpha/low beta power suppression was more pronounced over left anterior scalp sites (F3, F5, F7, FT7, FC5, T7) and during early stages of speech preparation (from -400 to -150 ms relative to speech onset).

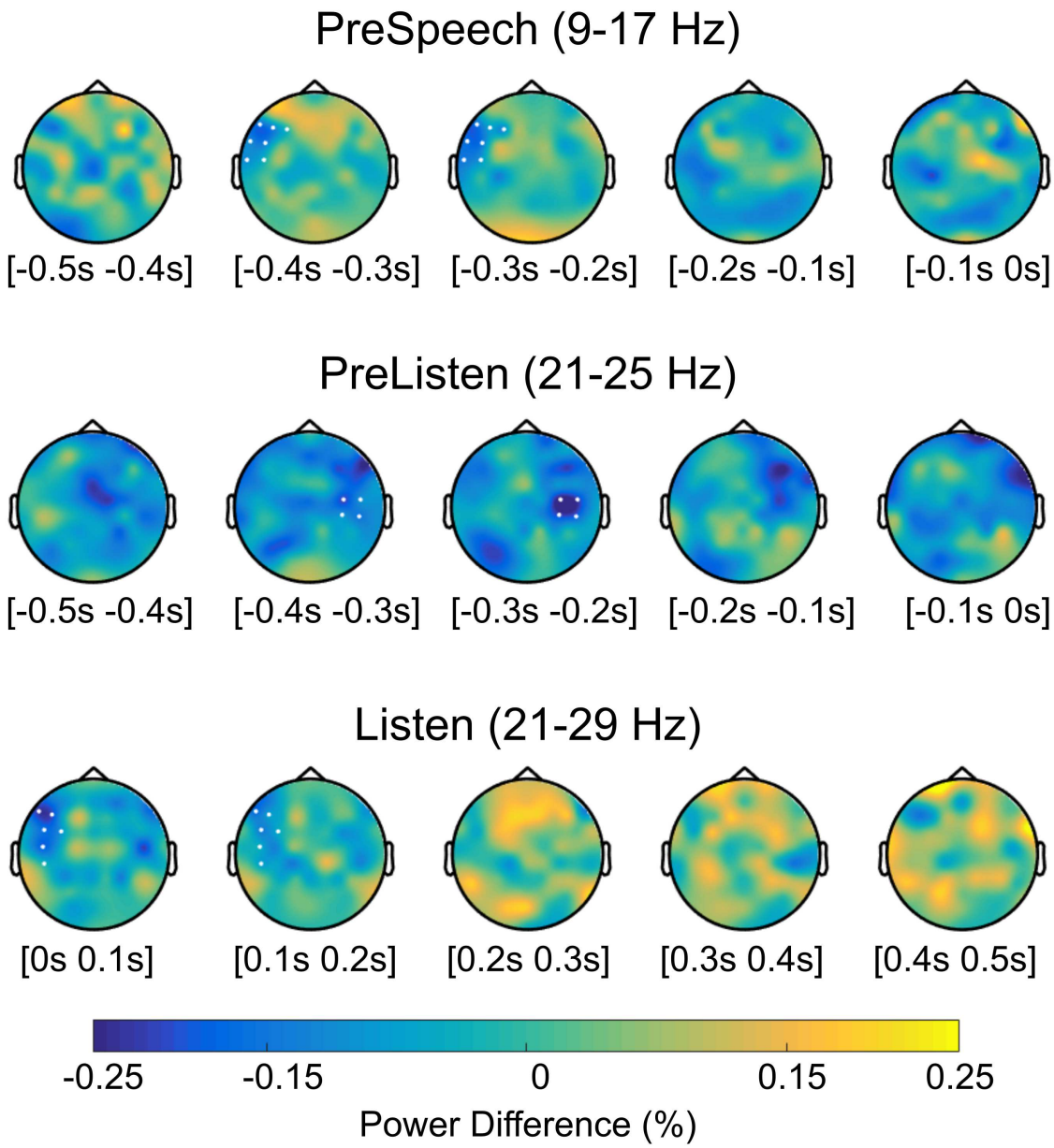

Figure 4.3 *A. Topographical plots of the relative power changes between convergence and NoChange* $\frac{Convergence-NoChange}{NoChange}$ *are shown for the frequency ranges for which the cluster-based permutation test yielded a significant difference. PreSpeech epoch refers to the preparation to speak. PreListen and Listen epochs instead refers to listener's brain activities, respectively, while the partner is speaking and before he or she speaks. Each topographic plot shows the change in power across the two data sets in 100 ms time windows, covering the entire 500 ms epoch of interest. The white dots mark the channels for which significant differences were found.*

The observed power modulation did not depend on the reaction time (possibly indexing task difficulty), since no difference in reaction times was found between the two data-sets (p=0.64, see behavioral results). Moreover, we ensured that trial-by-trial fluctuations in reactions times were not associated with corresponding alpha/low beta-band power fluctuations. In fact, in all turn-taking behaviors, a confounding factor could be related to the temporal aspects of behavioral synchronization to the rhythm of the task (Fujioka et al. (2012)).

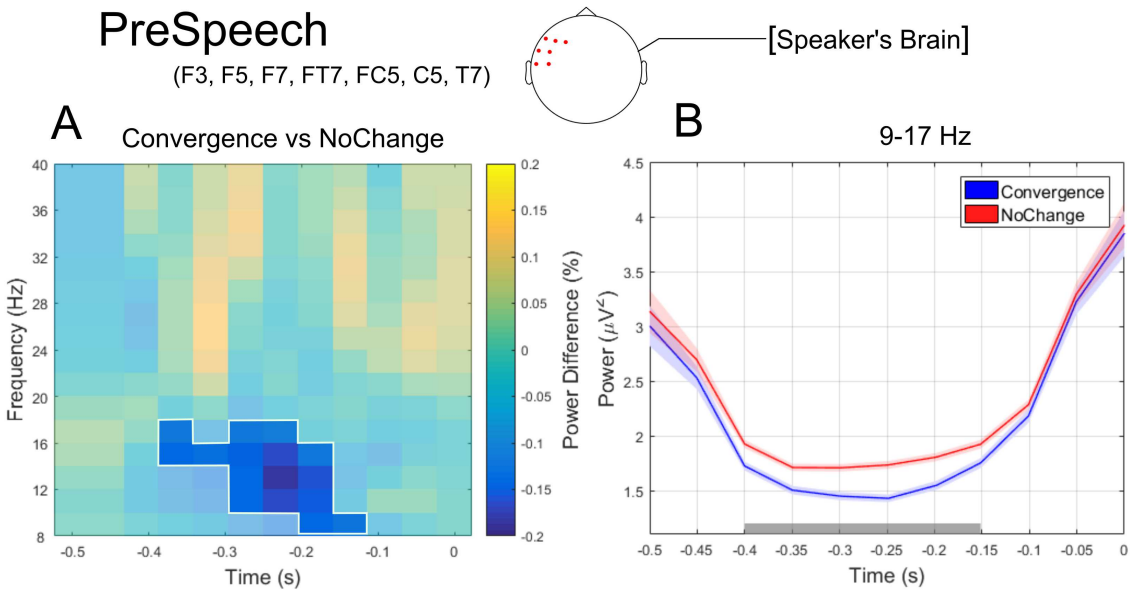

Figure 4.4 *(a) Relative power changes between convergence and NoChange* $\frac{Convergence-NoChange}{NoChange}$ *for the PreSpeech epoch, averaged over leftanterior channels (F3, F5, F7, FT7, FC5, T7; channels showing statistically significant differences according to the cluster-based permutation test) and plotted as a function of time (-0.5-0 s) and frequency (8-40 HZ). The unshaded spectrotemporal region where converge trials show strongest and statistically significant power decrease compared to NoChange trials (i.e., 9-17 Hz frequency range and - 0.4-0.15 s time window). (b) Temporal evolution of the oscillatory power averaged across frequencies ranging from 9 to 17 Hz and left anterior channels F3, F5, F7, FT7, FC5, T7 for convergence (blue line) and NoChange (red line). The gray horizontal line indicates the time points where the cluster-based permutation test revealed a significant difference between data sets. Colored shaded areas indicate standard error of the mean.*

To this end, we calculated the Pearson correlation between single-trial reaction times and oscillatory power averaged across the time points (from -400 to -150 ms), frequencies (from 9 to 17 Hz) and electrodes (F3, F5, F7, FT7, FC5, T7) where we found the strongest power modulations between Convergence and NoChange (see above). Correlation was not significant for both data-sets (Convergence, r = -0.05, p=0.77; NoChange, r =-0.02, p=0.74),

confirming that the oscillatory activity that is modulated by phonetic convergence is not related to the (within-data-set) variability in reaction times.

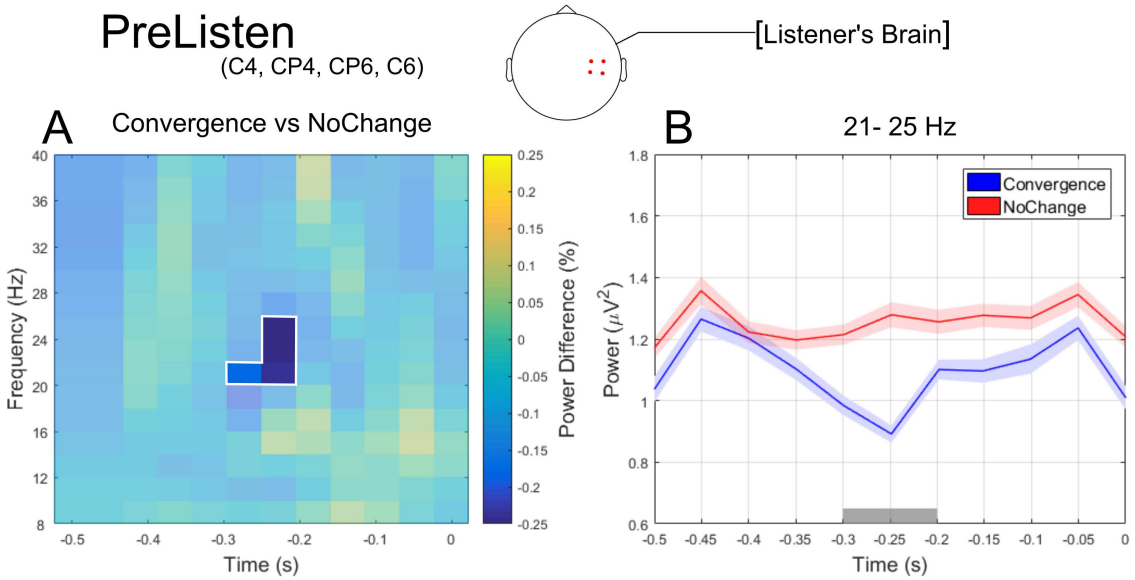

Figure 4.5 *(a) Relative power changes between convergence and NoChange* $\frac{Convergence-NoChange}{NoChange}$ *for the PreListen epoch, averaged over left anterior channels (C4 C6 CP6 CP4; channels showing statistically significant differences according to the cluster-based permutation test) and plotted as a function of time (-0.5-0 s) and frequency (8-40 HZ). The unshaded spectrotemporal region where converge trials show strongest and statistically significant power decrease compared to NoChange trials (i.e., 21-25 Hz frequency range and - 0.3-0.2 s time window). (b) Temporal evolution of the oscillatory power averaged across frequencies ranging from 21 to 25 Hz and right centro-parietal channels C4 C6 CP6 CP4 for Convergence (blue line) and NoChange (red line). The gray horizontal line indicates the time points where the cluster-based permutation test revealed a significant difference between data sets. Shaded areas indicate SE of the mean.*

We investigated the effect of convergence in the listener's brain, before the interlocutor started speaking. With this analysis, we sought to establish whether a specific neural state, as indexed by the ongoing oscillatory power, preceded convergent words. We found a statistically significant difference in oscillatory power across data-sets (p=0.02; Fig. 6). Again, this difference consisted in a reduction of power for the Convergence compared to the NoChange data-set, which was most consistent in the beta band (21-25 Hz), over right centro-parietal electrodes (C4, C6, CP6, CP4) and between -290 and -190 ms relative to the partner's voice onset.

The Pearson correlation showed no significant relationship between beta-band power (averaged across significant frequencies, electrodes and time points) and reaction times at the

single-trial level, indicating that beta power before listening does not co-vary with reaction time (Convergence: r = 0.08, p =0.18; NoChange: r = 0.05, p=0.2).

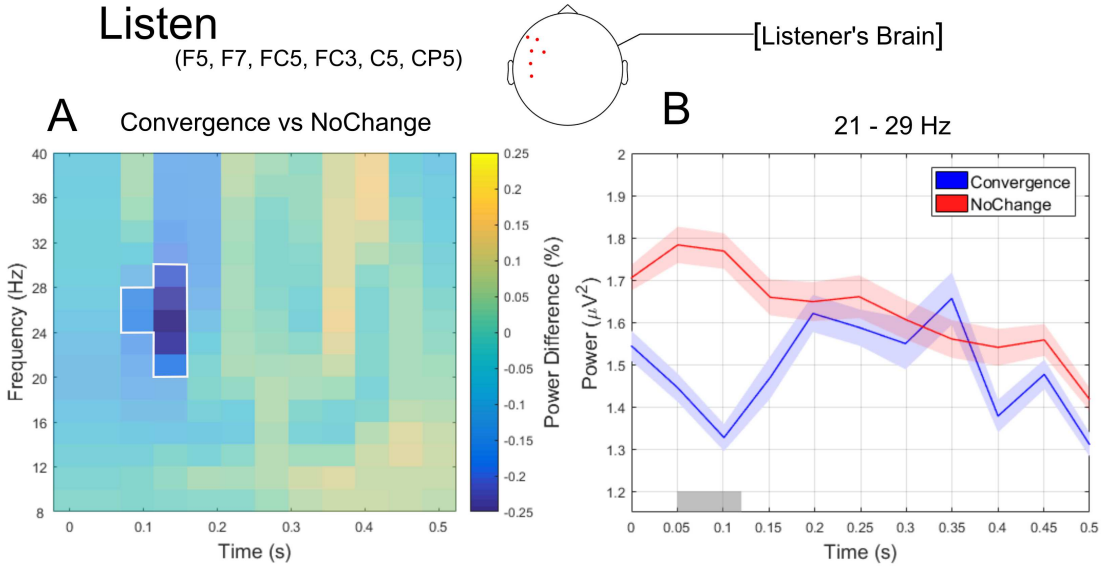

Figure 4.6 *A. Relative power changes between convergence and NoChange* $\frac{Convergence-NoChange}{NoChange}$ *for the Listen epoch, averaged over left anterior channels (F5, F7, FC5, FC3, C5, CP5; channels showing statistically significant differences according to the cluster-based permutation test) and plotted as a function of time (0-0.5 s) and frequency (8-40 HZ). The unshaded spectrotemporal region where converge trials show strongest and statistically significant power decrease compared to NoChange trials (i.e., 21-29 Hz frequency range and + 0.05-0.15 s time window). (b) Temporal evolution of the oscillatory power averaged across frequencies ranging from 21 to 29 Hz and left anterior channels F5, F7, FC5, FC3, C5, CP5 for Convergence (blue line) and NoChange (red line). The gray horizontal line indicates the time points where the cluster-based permutation test revealed a significant difference between data sets. Shaded areas indicate SE of the mean.*

Finally, we also found a significant difference in oscillatory power between Convergence and NoChange in the listener (p=0.03; Figure 4.6). In particular, Convergence trials showed a reduction in power compared to NoChange trials that was most pronounced in the beta band (21-29 Hz), over left frontal electrodes (F5, F7, FC5, FC3, C5, CP5) and just after the partner's voice onset (50-120 ms; i.e., at the very beginning of the listening phase).

Again, we checked whether oscillatory power was related to reaction times in the two data-sets. The Pearson correlation (calculated in the same way as already described for the PreSpeech epoch) yielded no significant relationship between trial-by-trial fluctuations in the beta-band power (in the 21-29 Hz range) and reaction times for both data-sets (Convergence: r = -0.17, p=0.97; NoChange: r = 0.006, p=0.35). As for the other epochs, our results are

not driven by the rhythmic nature of the task and thus by behavioral synchronization in the temporal domain.

Chapter 5

Conclusion

A first novelty of this work is to investigate the role of motor system towards reconstructing missing sensory information. We measure the coherence between brain signals and an articulatory signal (Lip-aperture) which is missing from the subjects sensory environment. Second, we investigate the mutual behavioral adaptations happening during a speech-based interaction. We first create an engaging verbal interaction task and then with a robust algorithm quantify the coordination that arises during interaction. Afterward, this word-level index of mutual adaptation was used to explore the neural correlates associated to this emergent joint phenomenon.

In the following paragraphs, I will briefly summaries the contribution of my studies and the future potential extensions of this work.

5.1 Motor contribution towards reconstructing missing sensory information

Our data shows that neural entrainment during speech listening is reconstructive and predictive. Brain coherence with missing cues (lips-opening) was indeed significant and temporally dissociated from that observed for available acoustic cues (speech envelope). The reconstructive nature of this phenomenon is essential to substantiate its functional role (Crosse et al. (2015)). Indeed, it has been reported that frontal brain activities could modulate the strength of neural entrainment to speech in sensory cortices with a relatively short delay of 50-60 ms (Park et al. (2015)). This has been taken as evidence of a top-down attentional modulation exerted upon a fundamentally bottom-up, sensory-driven process (Park et al. (2018b)).

However, these recent studies do not clarify whether top-down neural processes can synthesize missing sensory signals. Demonstrating a reconstructive capability is the prerequisite to consider neural entrainment as instantiating an analysis-by-synthesis computation (Bever and Poeppel (2010); Crosse et al. (2015)). At the same time, the main functional tenet of an analysis-by-synthesis computation resides also in its anticipatory capability. In fact, bottom-up and top-down neural processing should be characterized by a temporally dissociated coupling with sensory signals (Park et al. (2018b)). Here we specifically test the hypothesis that coherence with speech envelope and with (absent) lip-opening are temporally dissociated. Indeed, in agreement with previous investigations, we show significant speech-brain coherence (delta, theta and alpha frequencies) with a lag of 100-300 ms (Crosse et al. (2015); O'sullivan et al. (2014)) and a fronto central topography (Peelle et al. (2012); Vander Ghinst et al. (2016); Bourguignon et al. (2018)). Instead, in the delta range, lips-brain coherence peaks around 0-100 ms, with a similar topography. Most importantly, the temporal dynamic of neural entrainment significantly differs between lips and speech, even after partialling out the contribution of the other signal.

Remarkably, brain entrainment to absent sensory cues was shown without introducing any explicit or implicit request to focus on the visual description of the acoustically presented sentence. Moreover, behavioral performance was at ceiling, suggesting that only a limited advantage could have been derived from the integration of visual (real or reconstructed) information. These two aspects seem to suggest that reconstruction of missing visual cues was independent from a specific attentional effort and, more importantly, automatically put in place in the absence of a specific behavioral need. Hence, to grasp the functional meaning of this reconstructive process, we should consider the temporal relationship between acoustic and visual speech signals as well as their different informational content.

Mouth opening and speech envelope are temporally modulated in the 2–7 Hz frequency range (Chandrasekaran et al. (2009)), which overlaps with the timescale of syllable production as well as with the rhythmic fingerprint of auditory cortices (Giraud and Poeppel (2012)). These commonalities are suggestive of a potential functional link between rhythmicity in speech articulatory movements, in their acoustic outcome and in the intrinsic activity of sensory cortices. However, mouth opening is the visible part of a far more complex pattern of phono-articulatory configurations needed to produce speech sounds (Catford (1977)). Still, notwithstanding the fact that visual signals are far less informative than acoustic ones (Fisher (1968)), they entrain visual cortex oscillations during silent lip-reading (O'Sullivan et al. (2017); Hauswald et al. (2018); Ozker et al. (2018)).

However, there seems to be an asymmetric relationship between activities in primary visual and auditory cortices during audio-visual speech perception. The auditory cortex tracks the temporal dynamics of visual speech cues, while the tracking of auditory speech cues by the visual cortex is far less pronounced (Megevand et al. (2018)). This asymmetry could be explained by the fact that the acoustic signal is in principle sufficient for understanding the message, while the visual one plays a supportive role, which may gain more relevance only in the presence of a degraded acoustic signal (Soto-Faraco et al. (2007)). At the same time though, the visual signal can be anticipated (Chandrasekaran et al. (2009)), potentially conveying critical predictive value for the decoding of the true objects of perception which reside in the acoustic signal.

This putative predictive power of visual cues in audio-visual speech perception has already been described as a reduction in the latency of Event-Related Potentials (ERPs) evoked by speech listening (Van Wassenhove et al. (2005)). Manipulation of the amount of information carried by the visual signal was also linked to the magnitude of the anticipatory effects (Van Wassenhove et al. (2005); Stekelenburg and Vroomen (2007); Arnal et al. (2009)). Predictive models, in general, envision that prior knowledge biases sensory processing by means of top-down projections (Friston and Frith (2015)). Therefore, it is possible that the anticipated visual signal, although carrying relatively low information, can still bias acoustic sensory processing (Schroeder et al. (2008); Golumbic et al. (2013a)).

We find evidence that speech listening does not only entail neural tracking of speech envelope but also includes the anticipatory synthesis of visual speech cues. The neuro-functional origin of these reconstructive processes may reside in the computations run in the frontal cortex. Indeed, unimodal neural entrainment to speech envelope or mouth opening is driven by neural sources located in the left motor and premotor cortex (Crosse et al. (2015); Park et al. (2018a)). A precentral origin of these top-down modulations is also suggested by the fact that corticobulbar excitability is modulated by passive listening to speech (Fadiga et al. (2002); Watkins et al. (2003); D'Ausilio et al. (2014); Schmitz et al. (2018)) and that transient perturbation of the activity in premotor and motor areas produces somatotopically organized modulation of speech discrimination performance (D'Ausilio et al. (2009); Möttönen and Watkins (2009); Sato et al. (2009); D'Ausilio et al. (2012); Bartoli et al. (2015); Murakami et al. (2018)).

5.2 Neural correlates of Phonetic convergence

Phonetic convergence is the phenomenon by which participants in a dialogue tend to naturally align with each other in their phonetic characteristics (Pardo (2013)). Although convergence (phonetic alignment) is a well-known phenomenon, its quantitative assessment is still an open area of research. Several studies have focused on subjective evaluations (Pardo (2013)), whereas others have used a variety of objective acoustic measures (Goldinger (1998)). Therefore, a great deal of inconsistency and variability still exists among studies (Pardo et al. (2017)). One key novelty of our study is that we implemented a quantitative method to extract phonetic convergence from a game-like task, allowing an engaging, yet relatively constrained, phonetic interaction. Phonetic convergence was computed using a robust and automatic speaker identification technique applied to the full acoustic spectrum, thus reducing the number of a priori hypotheses about which acoustic feature shows alignment (Mukherjee et al. (2017)). This method was designed to specifically evaluate cooperative speech behavior. Indeed, phonetic convergence is not extracted from individual speech characteristics, but is rather computed out of the combination of both speakers' speech production. Therefore, we quantified participants' joint efforts to imitate each other's acoustic targets.

The quantification of phonetic convergence as the result of a joint-action behavior was the prerequisite to investigate its neural markers. Here, convergence was associated to specific oscillatory modulations in the alpha and beta bands. Convergent speech preparation in the speaker's brain was characterized by alpha/low beta power suppression which was most prominent over left fronto-central electrodes and early before speech onset (from -400 ms to -150 ms). Convergence in the listener's brain, instead, showed significant beta suppression peaking over left fronto-central sites just after the partner's speech onset (from 50 to 120 ms). At the same time, phonetic convergence is also characterized by lower power of the ongoing beta rhythm over right centro-parietal electrodes before listening. Overall, these findings suggest that alpha and beta oscillatory dynamics are associated with phonetic convergence.

These results are in line with previous studies reporting modulation of alpha (Kawasaki et al. (2013); Mandel et al. (2016); Pérez et al. (2017); Ahn et al. (2018)) and beta rhythms (Mandel et al. (2016); Pérez et al. (2017)) during speech-based interaction tasks. However, one key aspect differentiates our study with respect to prior hyper-scanning investigation of speech interaction. We used a joint-action behavioral feature as a searchlight for the neural underpinnings of speech coordination. In fact, our EEG analysis was driven by a behavioral index that cannot directly nor independently be controlled by any of the partners during the interaction, i.e. phonetic alignment.

5.2.1 The sensorimotor nature of phonetic convergence

As far as the alpha/low beta effect in the speaker is concerned, we first observe that fronto-central de-synchronization in the upper alpha and lower beta bands, always precedes voluntary movements (Pfurtscheller and Aranibar (1979); Pfurtscheller and Berghold (1989); Leocani et al. (1997)). Interestingly, similar results can be observed across hyper-scanning studies. Tognoli and Kelso (2015) employed an interactive finger movement task and manipulated the subjects' view of each other's hands. The results showed that neural oscillations in the alpha range (the phi complex) were modulated by the control of participants' own behavior in relation to that of the partners. Following this pioneering dual-EEG study, a few others have confirmed the role of alpha oscillations, overlaying sensorimotor regions, in behavioral coordination (Dumas et al. (2010); Konvalinka et al. (2014)). In general, the comparison between interactive and non-interactive behaviors has consistently shown the suppression of alpha range oscillations (Tognoli and Kelso (2015)). However, task differences can produce slightly different topographical maps of alpha/low beta suppression. For instance, a centro-parietal topography in a joint attention task (Lachat et al. (2012)), a frontal one in a finger-tapping task (Konvalinka et al. (2014)), while a central effect was present in a nonverbal hand movement task (Ménoret et al. (2014)). All in all, our fronto-central effect matches similar hyper-scanning results, while its left topography may be explained by the lateralization of the speech production function.

To discuss about the functional meaning of our results, we refer to the fact that a rolandic alpha de-synchronization is usually found during execution, observation or mental imagery of movements, possibly reflecting the activation or release from inhibition of the sensorimotor cortex (Caetano et al. (2007); Cochin (1999); Pfurtscheller and da Silva (1999)). In fact, multi agent action coordination requires that participants produce their own actions while simultaneously perceiving the actions of their partners. Similarly, a speech conversation creates the need for a tight action-perception coupling (Hari and Kujala (2009)). In fact, the central alpha band suppression has been proposed to be an index of action-perception coupling (de Lange et al. (2008); Hari (2006)), and thus sensorimotor information transfer during behavioral coordination. Within this context, our study provides evidence that alpha suppression, extending to the low beta range, is present also during speech interaction, in a task that critically requires coordination of articulatory gestures. More importantly, these EEG features were modulated by the efficacy with which participants jointly (as opposed to independently) managed to coordinate each other while converging towards a shared phonetic space.

Moving to the listener's brain activities, phonetic convergence leads to the suppression of beta oscillations. In general, as for the rolandic alpha, fronto-central beta-band de-synchronization has been related to the activation of the sensorimotor cortices (Salmelin et al. (1995); Parkes et al. (2006)). However, using electrocorticography (ECoG) it was shown that beta event-related de-synchronization (ERD) is more focused and somatotopically specific than alpha ERD (Crone et al. (1998)). In this sense, it has been proposed that the rolandic alpha ERD reflects the unspecific activation of sensorimotor areas, while the beta ERD signals a relatively more focal motor recruitment (Pfurtscheller et al. (1994); Pfurtscheller and da Silva (1999)). More specifically and in line with our findings, somatotopic beta attenuation has also been shown for speech listening (Jenson et al. (2014); Bartoli et al. (2016)). In fact, specific sensorimotor regions recruited during speech production are also activated during speech listening (Fadiga et al. (2002); Watkins et al. (2003); D'Ausilio et al. (2014)) and the perturbation of these sensorimotor centers affects speech discrimination performance (Meister et al. (2007); D'Ausilio et al. (2009); D'Ausilio et al. (2012); Bartoli et al. (2015); Möttönen and Watkins (2009)). The beta ERD we observe after speech presentation may thus be interpreted as supporting the perceptual discrimination processes. The effect is localized in a left fronto-central cluster of electrodes, supporting the claim that top-down sensorimotor predictions can exert a functional contribution to the hierarchical generative models underlying speech perception (Cope et al. (2017)).

5.2.2 Phonetic convergence and predictive coding

Beta ERD was also shown to precede speech listening, though with a right centro-parietal topography. This pattern of lateralization is consistent with a possible attentional role (Petit et al. (2007); Gao et al. (2017)). In agreement with this possibility, it has been proposed that the functional role of the pre-stimulus beta rhythm is to convey motor information (efferent copies) to suppress self-generated sensory stimulations, freeing up resources to respond to external sensory stimuli (Engel and Fries (2010)). The mechanism of action might be that beta rhythms interact with other modality-specific rhythms to anticipate sensory events by boosting neural excitability at specific moments in time when salient stimuli are expected to happen (Arnal et al. (2011)). Such a predictive top-down influence is supposed to play a key role in attentional selection (Lewis et al. (2015); Lee et al. (2013); Bastos et al. (2015); Morillon and Baillet (2017)). Interestingly, the pre-stimulus beta suppression has shown to be involved in predictions about the precision of a specific processing channel thus establishing the attentional context for perceptual processing (Bauer et al. (2014)).

In this framework, the brain acts like a predictive engine (Friston et al. (2011)), aiming at reducing the cost of analyzing the full set of incoming information, by formulating specific perceptual hypotheses that are tested against the temporal flow of sensory evidences (Donnarumma et al. (2017a)). In our turn-taking task, the organization of one's own speech output could bias the subsequent active listening processes by allowing faster or more efficient discrimination of similar acoustic targets. On the other hand, motor activations elicited by speech perception could in turn prime the organization of the immediately following speech planning required in the task. Based on general principles of neural reuse (Anderson (2010)) of action-perception circuits for speech communication (Pulvermüller (2018)) phonetic convergence may depend on the amount of sensorimotor detail extracted while discriminating the speech produced by the partner. In this sense, the degree of neurofunctional sensorimotor overlap between speech perception and production may translate into larger likelihood of motor contagion (Bisio et al. (2010); Bisio et al. (2014); D'Ausilio et al. (2015)).

5.2.3 Predicting the "how" rather than the "when" of speech interaction

Effective prediction however requires task predictability. In our interactive task, the listeners have critical prior information to constrain perceptual analysis. From the listener's point of view, the word spoken by the partner shares one out of the two syllables of the word just produced by herself. The other syllable, the novel one, is contained in one of the two words that the participant can now read on the screen, and that she'll have to pronounce. These task dependencies offer strong anchoring points to predict the dynamics of the ongoing interaction. Importantly, the listeners are forced into a predictive mode of operation regarding the phonetic content (i.e. what syllables I'm going to hear) rather than the timing characteristics of the turn-taking action. This aspect is of particular interest if we consider that previous studies investigated the neural dynamics subtending the estimation of "when" a partner is going to speak in a conversation (Mandel et al. (2016)). In fact, the estimation of this temporal information is fundamental in establishing effective turn taking, as well as supporting word segmentation and parsing sentence-level syntax. However, temporal prediction may not be the only anticipatory mechanism at play during speech interaction. In the present study, we took a different direction by mixing a set of task constraints together with specific computational methods, to investigate how people engage in highly predictive behaviors regarding the (phonetic) "how" component of speech interaction. In this regard, phonological convergence

has been here considered as the tendency to align phono-articulatory tract gestures during the interaction (Mukherjee et al. (2018)).

Here we show that phonetic convergence elicits specific patterns of alpha and beta suppressions that dissociate the speaking, preparing to listen and listening phases. The novelty of the current study arises also from the characterization of phonetic convergence as a dynamic and interactive process. In doing so, these results add to the few recent studies aiming at the investigation of speech and language processes during (quasi)-realistic verbal interactions. In fact, we need to bear in mind that in fast-paced natural dialogues, comprehension and production tend to greatly overlap in time (Levinson and Torreira (2015)). Based on this evidence, it has been suggested that one key issue is to which extent current models of language, developed for isolated individuals (Hickok and Poeppel (2007)), are still valid in interactive contexts (Pickering and Garrod (2013); Schoot et al. (2016)). On one hand, turn-taking involves multi-tasking comprehension and production (Levinson (2016)) and indeed the neural network for language production and comprehension may at least partially overlap (Menenti et al. (2012)). On the other, the now classical neural entrainment to surface auditory features during attentive listening (Luo and Poeppel (2007);Schroeder and Lakatos (2009); Giraud and Poeppel (2012); Ding and Simon (2014); Park et al. (2016)) does not seem to fully explain inter-brain synchronization occurring during conversations (Pérez et al. (2017)). Therefore, when we extrapolate results to the complexity of ecological scenarios, the listener, apart from speech comprehension, may adapts the phono-articulatory properties of speech preparation through substantially incomplete understanding and at the same time may influence the speaker's brain processes through back-channeling.

In conclusion, mutual understanding might be the result of a joint process whereby alignment of situation models is facilitated when interlocutors align their behavioral output (Pickering and Garrod (2004); Schoot et al. (2016)). Also, the fast-paced interactive nature of dialogues suggests that speech and language understanding and production form a shared process that is co-constructed by participants (Donnarumma et al. (2017b)). Along these lines, an emerging trend suggests that a complete grasp of the neural and cognitive processes involved in speech-based communication cannot be achieved without examining more realistic interactions among individuals (Hasson et al. (2012); Pickering and Garrod (2013); Schoot et al. (2016)). However, it is important to highlight that, to investigate the phonetic aspect of linguistic convergence, the current study implemented a series of task constraints to allow a moderate level of experimental control. Eventually, the investigation of whether more realistic and open-ended scenarios result in similar neurobehavioral phenomena will have to be tackled by future studies.

5.3 Future Works

5.3.1 Neural representation of articulatory configurations

We found a relationship between lip-opening with brain responses. We found that this is anticipatory and reconstructive in nature. But here we just looked at one single articulatory feature which, in many natural interactions, is also visible to the listener. In future this work could be extended to look for brain entrainment to those articulators that are never visible to the listener. Apart from lip aperture we recorded other articulatory signals. In fact the same database (i.e. Multi-SPeaKing-style Articulatory corpus; Canevari et al. (2015)) contains also recording of upper and lower incisors, tongue tip, tongue dorsum, tongue back positions (x,y,z components). These raw positional descriptions should however be combined to find a lower dimensionality signal which can represent the whole vocal tract.

Combining these features could be done in two ways. One way could be hypothesis driven via Articulatory Phonology literature (Ohala et al. (1986), Browman and Goldstein (1989), Browman and Goldstein (1992)), which consider that the units of speech production are actions and therefore they are dynamic in nature. In fact macroscopic descriptions of such actions is called articulatory gesture. Example of articulatory gestures are narrowing lips, raising tongue tip etc. A combinations of these articulatory gestures give rise to how human produce sounds. The time-varying trajectories of the individual articulators (such as the lips, tongue, glottis, velum, etc.) can thus represent meaningful descriptions of the vocal tract. In fact this is what we have used in our study e.g. lip aperture. Future work then, will need to combine these raw positional descriptions in a meaningful way. Another way could be data driven approach. By this way we could use different feature extraction techniques in order to extract complex pattern from the x,y,z positions of the EMA sensors. For example, Principal Component Analysis (PCA) which uses an orthogonal transformation in order to extract some linearly uncorrelated variables (or Principal Components) from a set of possibly correlated observed variables. Another possible route could be using acoustic-to-articulatory inversion mapping (Richmond (2006), Wu et al. (2015)), where we use speech signal in the input of a deep neural network and in the output we have articulatory signal. By these way one could extract a lower dimensionality non-linear mixture of the original signals which is also supported by its corresponding acoustic counterpart.

Both of these approaches have advantages and disadvantages. The hypothesis driven has the advantage of being interpretable and of being based on a solid theoretical and empirical background, Such an approach would render interpretation of brain activities potentially easier. At the same time this suffers from human bias as inherently we cannot fathom all

the properties inherent to a system and some will bound to be missed out. On the other hand, the data driven approach could in principle extract complex new features that may describe the system in a much more accurate way that any hypothesis driven approach. But the disadvantage of this is the interpretation of these features, which could be non-trivial.

The next challenge would be to properly measure the strengths of correlations between these features and brain oscillatory dynamics as this essentially amounts to solving a many-to-many mappings. We can use Canonical Correlation Analysis (CCA) or seq2seq modeling to find a meaningful relation with the neural data. CCA tries to find linear combinations between two data sets (here brain signal on all EEG channels and extracted articulatory features) which have maximum correlation with each other (de Cheveigné et al. (2018)). The seq2seq modeling use encoder-decoder framework with deep neural network with brain and extracted features in its input and output respectively (Britz et al. (2017)). The procedure is similar to acoustic-to-articulatory inversion mapping only here we have brain activities instead of acoustic signals.

5.3.2 Extending the framework to a true conversation

Our specialized Verbal Domino Task is made in such a way that it is engaging as well as simple enough to investigate different aspects of verbal interaction such as neural and articulatory counterparts. However, this task is far from a natural conversation where we suspect we could see much more interesting phenomena like dialect-leveling (where speakers mix multiple different dialects), code-switching (where speakers mix different languages), Style-shifting (refers to a single speaker changing style in response to context) etc.

The algorithm used here for the quantification of convergence has been proved to be effective in two different languages e.g. English, French. However, it can be improved so that it can be used in natural conversation scenarios. This would requires first to create a speaker diarization module which is the process of partitioning an input audio stream into homogeneous segments according to the speaker identity (Tranter and Reynolds (2006)). This is relatively straight forward area of research where there have been huge advancements (Anguera et al. (2012)). So the first work would be to extend this task in such a way that speakers can interact with each other in a free flow fashion. My proposal would be to use conservatively sentences instead of words as a transition to natural conversations.

Another area of improvement is to include visual cues, e.g. head nodding (which is a clear example of non-verbal backchanneling during verbal interaction) in order to improve convergence scores accuracy. In fact, we are already building a mobile app which is using

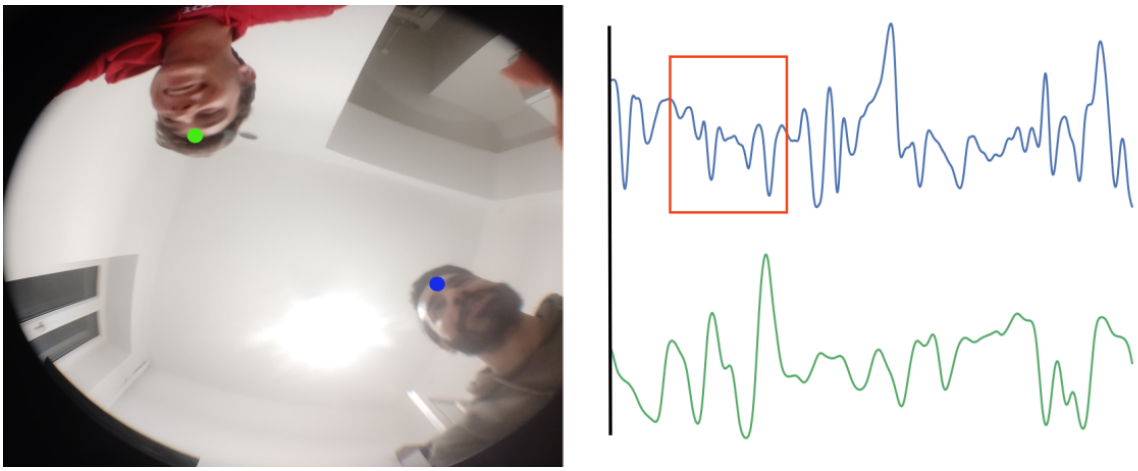

Figure 5.1 *An example of fish eye lens capturing head movements of speakers. Right side of the picture show example signals of speakers head movements.*

the convergence algorithm with an embedded module tracking head movements by using a fish eye lens (Figure 5.1; a special lens, that can be mounted on a smartphone camera and capable of capturing a 180° view). Updates and news on this project will be available in this website in the future https://speechconvergence.com/.

5.3.3 One example application: Second Language Learning

Convergence is in general associated to interaction efficacy, likability (Turner and West (2010)). However little is known about its role in learning how to reach the correct phonetic targets in a second language (L2). In fact, near native pronunciation in L2 is an ability that, dissociates from lexical or syntactic skills (Scovel (1969)), and is never achieved in some individuals (Seliger et al. (1982), Selinker (1972)). In this regard, as far as the teaching strategies to facilitate the acquisition of proper L2 pronunciation, the main approach has been to use the principles of L1 acquisition (i.e. exaggerated input, variety of input, providing visible articulation cues) and to design training protocols to increase personal engagement. In fact, motivation and engagements are key components for successful L2 learning (Gilakjani et al. (2012)). In this sense, modern L2 pedagogy use Computer-Assisted Language Learning methods (CALLs) to include some of the principles of L1-acquisition as well as strategies to boost engagement (Warschauer and Healey (1998)). Historically speaking, L2 learning has poorly benefited from technological innovations (Salaberry (2001)). Usually, CALL tools enhance teaching situations by making them more interesting and interactive for the students. These approaches however, do not offer any specific novel strategy to restructure L1 neural representations to accommodate L2 phonetic specificities. Rather, these tools merely adapt

to the new technology-mediated communication landscape. In fact, even the latest tools still force people to adapt and acquire the L2 phonetic space by means of explicit rule-based learning. Instead, children learn L1 phonetic rules automatically and implicitly through natural interaction. As a follow-up of the work presented in this thesis we will explore the use of (the automatic detection of) phonetic convergence to provide an objective tool to evaluate students proficiency and gradual improvements towards the acoustic target provided by the native speaker.

References

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., and Merzenich, M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 98(23):13367–13372.

Ahn, S., Cho, H., Kwon, M., Kim, K., Kwon, H., Kim, B., Chang, W., Chang, J., and Jun, S. (2018). Interbrain phase synchronization during turn-taking verbal interaction—a hyperscanning study using simultaneous eeg/meg. *Human Brain Mapping*, 39(1):171–188.

Aiken, S. J. and Picton, T. W. (2008). Envelope and spectral frequency-following responses to vowel sounds. *Hearing Research*, 245(1):35–47.

Anderson, M. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and Brain Sciences*, 33(4):245–266.

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., and Vinyals, O. (2012). Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370.

Arléo, A. (1997). Un jeu de dominos verbal: Trois p'tits chats, chapeau d'paille. *Chants Enfantins D'Europe*, pages 33–68.

Arnal, L., Wyart, V., and Giraud, A.-L. (2011). Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience*, 14(6):797–801.

Arnal, L. H., Morillon, B., Kell, C. A., and Giraud, A.-L. (2009). Dual neural routing of visual facilitation in speech processing. *Journal of Neuroscience*, 29(43):13445–13453.

Aubanel, V. and Nguyen, N. (2010). Automatic recognition of regional phonological variation in conversational interaction. *Speech Communication*, 52(6):577–586.

Aubanel, V. and Nguyen, N. (2010). Automatic recognition of regional phonological variation in conversational interaction. *Speech Communication*, 52(6):577–586.

Babel, M. and Bulatov, D. (2012). The role of fundamental frequency in phonetic accommodation. *Language and Speech*, 55(2):231–248.

Bailly, G. and Lelong, A. (2010). Speech dominoes and phonetic convergence. pages 1153–1156.

Bailly, G. and Martin, A. (2014). Assessing objective characterizations of phonetic convergence. pages 2011–2015.

Baker, S. C., Gallois, C., Driedger, S. M., and Santesso, N. (2011). Communication accommodation and managing musculoskeletal disorders: doctors' and patients' perspectives. *Health Communication*, 26(4):379–388.

Bartoli, E., D'Ausilio, A., Berry, J., Badino, L., Bever, T., and Fadiga, L. (2015). Listener-speaker perceived distance predicts the degree of motor contribution to speech perception. *Cerebral Cortex*, 25(2):281–288.

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

Bastiaansen, M. and Hagoort, P. (2006). Chapter 12 oscillatory neuronal dynamics during language comprehension. *Progress in Brain Research*, 159:179–196.

Bastiaansen, M., Oostenveld, R., Jensen, O., and Hagoort, P. (2008). I see what you mean: Theta power increases are involved in the retrieval of lexical semantic information. *Brain and Language*, 106(1):15–28.

Bastiaansen, M., Van Berkum, J., and Hagoort, P. (2002). Event-related theta power increases in the human eeg during online sentence processing. *Neuroscience Letters*, 323(1):13–16.

Bastos, A., Vezoli, J., Bosman, C., Schoffelen, J.-M., Oostenveld, R., Dowdall, J., DeWeerd, P., Kennedy, H., and Fries, P. (2015). Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron*, 85(2):390–401.

Bauer, M., Stenner, M.-P., Friston, K., and Dolan, R. (2014). Attentional modulation of alpha/beta and gamma oscillations reflect functionally distinct processes. *Journal of Neuroscience*, 34(48):16117–16125.

Berry, J. J. (2011). Accuracy of the ndi wave speech research system. *Journal of Speech Language and Hearing Research*, 54(5):1295–1301.

Bever, T. G. and Poeppel, D. (2010). Analysis by synthesis: A (re-) emerging program of research for language and vision. *Biolinguistics*, 4(2-3):174–200.

Bilous, F. and Krauss, R. (1988). Dominance and accommodation in the conversational behaviours of same- and mixed-gender dyads. *Language and Communication*, 8(3-4):183–194.

Binder, J. R., Desai, R. H., Graves, W. W., and 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.

Bisio, A., Sciutti, A., Nori, F., Metta, G., Fadiga, L., Sandini, G., and Pozzo, T. (2014). Motor contagion during human-human and human-robot interaction. *PLoS ONE*, 9(8).

Bisio, A., Stucchi, N., Jacono, M., Fadiga, L., and Pozzo, T. (2010). Automatic versus voluntary motor imitation: Effect of visual context and stimulus velocity. *PLoS ONE*, 5(10).

Boersma, P. and Weenink, D. (2009). Praat: doing phonetics by computer (version 5.1. 05)[computer program]. retrieved may 1, 2009.

Bonhage, C., Meyer, L., Gruber, T., Friederici, A., and Mueller, J. (2017). Oscillatory eeg dynamics underlying automatic chunking during sentence processing. *NeuroImage*, 152:647–657.

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

Bourguignon, M., De Tiege, X., de Beeck, M. O., Ligot, N., Paquier, P., Van Bogaert, P., Goldman, S., Hari, R., and Jousmäki, V. (2013). The pace of prosodic phrasing couples the listener's cortex to the reader's voice. *Human brain mapping*, 34(2):314–326.

Bourhis, R. and Giles, H. (1977). The language of intergroup distinctiveness. *Language, Ethnicity and Intergroup Relations*, pages 119–135.

Bradlow, A. and Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106(2):707–729.

Britz, D., Goldie, A., Luong, T., and Le, Q. (2017). Massive Exploration of Neural Machine Translation Architectures. *ArXiv e-prints*.

Broca, P. (1861). Remarques sur le siège de la faculté du langage articulé, suivies d'une observation d'aphémie (perte de la parole). *Bulletin et Memoires de la Societe anatomique de Paris*, 6:330–357.

Browman, C. P. and Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, 6(2):201–251.

Browman, C. P. and Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, 49(3-4):155–180.

Buzsáki, G. and Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science*, 304(5679):1926–1929.

Caetano, G., Jousmäki, V., and Hari, R. (2007). Actor's and observer's primary motor cortices stabilize similarly after seen or heard motor actions. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21):9058–9062.

Canevari, C., Badino, L., and Fadiga, L. (2015). A new italian dataset of parallel acoustic and articulatory data. In *INTERSPEECH*, pages 2152–2156.

Canolty, R., Edwards, E., Dalal, S., Soltani, M., Nagarajan, S., Kirsch, H., Berger, M., Barbare, N., and Knight, R. (2006). High gamma power is phase-locked to theta oscillations in human neocortex. *Science*, 313(5793):1626–1628.

Catani, M. and ffytche, D. H. (2005). The rises and falls of disconnection syndromes. *Brain*, 128(10):2224–2239.

Catford, J. C. (1977). *Fundamental problems in phonetics*. Midland Books.

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

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979.

Cochin, S. (1999). Observation and execution of movement: Similarities demonstrated by quantified electroencephalography. *European Journal of Neuroscience*, 11(5):1839–1842.

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

Cooper, R. and Aslin, R. (1990). Preference for infant-directed speech in the first month after birth. *Child Development*, 61(5):1584–1595.

Cope, T., Sohoglu, E., Sedley, W., Patterson, K., Jones, P., Wiggins, J., Dawson, C., Grube, M., Carlyon, R., Griffiths, T., Davis, M., and Rowe, J. (2017). Evidence for causal top-down frontal contributions to predictive processes in speech perception. *Nature Communications*, 8(1).

Coupland, J., Coupland, N., and Giles, H. (1991). Accommodation theory. communication, context and consequences. *Contexts of Accommodation. Cambridge & Paris: Cambridge University Press & Editions de la maison des sciences de lâ AZhomme*, pages 1–68.

Coupland, N. (1984). Accommodation at work: Some phonological data and their implications. *International Journal of the Sociology of Language*, 1984(46):49–70.

Crone, N., Miglioretti, D., Gordon, B., Sieracki, J., Wilson, M., Uematsu, S., and Lesser, R. (1998). Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis: I. alpha and beta event-related desynchronization. *Brain*, 121(12):2271–2299.

Crosse, M. J., Butler, J. S., and 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.

D'Ausilio, A., Badino, L., Cipresso, P., Chirico, A., Ferrari, E., Riva, G., and Gaggioli, A. (2015). Automatic imitation of the arm kinematic profile in interacting partners. *Cognitive processing*, 16(1):197–201.

D'Ausilio, A., Bufalari, I., Salmas, P., and Fadiga, L. (2012). The role of the motor system in discriminating normal and degraded speech sounds. *Cortex*, 48(7):882–887.

D'Ausilio, A., Maffongelli, L., Bartoli, E., Campanella, M., Ferrari, E., Berry, J., and Fadiga, L. (2014). Listening to speech recruits specific tongue motor synergies as revealed by transcranial magnetic stimulation and tissue-doppler ultrasound imaging. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1644).

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., and Fadiga, L. (2009). The motor somatotopy of speech perception. *Current Biology*, 19(5):381–385.

de Cheveigné, A., Wong, D. D., Di Liberto, G. M., Hjortkjær, J., Slaney, M., and Lalor, E. (2018). Decoding the auditory brain with canonical component analysis. *NeuroImage*, 172:206–216.

de Lange, F., Jensen, O., Bauer, M., and Toni, I. (2008). Interactions between posterior gamma and frontal alpha/beta oscillations during imagined actions. *Frontiers in Human Neuroscience*, 2(AUG).

De Looze, C., Scherer, S., Vaughan, B., and Campbell, N. (2014). Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction. *Speech Communication*, 58:11–34.

Delorme, A. and 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.

Delvaux, V. and Soquet, A. (2007). The influence of ambient speech on adult speech productions through unintentional imitation. *Phonetica*, 64(2-3):145–173.

Destoky, F., Philippe, M., Bertels, J., Verhasselt, M., Coquelet, N., Vander Ghinst, M., Wens, V., De Tiège, X., and Bourguignon, M. (2019). Comparing the potential of meg and eeg to uncover brain tracking of speech temporal envelope. *NeuroImage*, 184:201–213.

Di Liberto, G., O'Sullivan, J., and Lalor, E. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, 25(19):2457–2465.

Dikker, S., Silbert, L., Hasson, U., and Zevin, J. (2014). On the same wavelength: Predictable language enhances speaker-listener brain-to-brain synchrony in posterior superior temporal gyrus. *Journal of Neuroscience*, 34(18):6267–6272.

Ding, N. and He, H. (2016). Rhythm of silence. *Trends in Cognitive Sciences*, 20(2):82–84.

Ding, N., Melloni, L., Zhang, H., Tian, X., and Poeppel, D. (2015). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, 19(1):158–164.

Ding, N. and Simon, J. (2014). Cortical entrainment to continuous speech: Functional roles and interpretations. *Frontiers in Human Neuroscience*, 8(MAY).

Ding, N. and Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, 109(29):11854–11859.

Ding, N. and Simon, J. Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *The Journal of Neuroscience*, 33(13):5728–5735.

Donnarumma, F., Costantini, M., Ambrosini, E., Friston, K., and Pezzulo, G. (2017a). Action perception as hypothesis testing. *Cortex*, 89:45–60.

Donnarumma, F., Dindo, H., Iodice, P., and Pezzulo, G. (2017b). You cannot speak and listen at the same time: a probabilistic model of turn-taking. *Biological Cybernetics*, 111(2):165–183.

Dumas, G., Nadel, J., Soussignan, R., Martinerie, J., and Garnero, L. (2010). Inter-brain synchronization during social interaction. *PLoS ONE*, 5(8).

Eadie, W. F. (2009). *21st century communication: a reference handbook*, volume 1. Sage.

Edlund, J., Heldner, M., and Hirschberg, J. (2009). Pause and gap length in face-to-face interaction. pages 2779–2782.

Ehrlich, S. and Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, 20(6):641–655.

Engel, A. and Fries, P. (2010). Beta-band oscillations-signalling the status quo? *Current Opinion in Neurobiology*, 20(2):156–165.

Fadiga, L., Craighero, L., Buccino, G., and Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: A tms study. *European Journal of Neuroscience*, 15(2):399–402.

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech, Language, and Hearing Research*, 11(4):796–804.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):815–836.

Friston, K., Mattout, J., and Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, 104(1-2):137–160.

Friston, K. J. and Frith, C. D. (2015). Active inference, communication and hermeneutics. *Cortex*, 68:129–143.

Fujioka, T., Trainor, L., Large, E., and Ross, B. (2012). Internalized timing of isochronous sounds is represented in neuromagnetic beta oscillations. *Journal of Neuroscience*, 32(5):1791–1802.

Gambi, C. and Pickering, M. (2013). Prediction and imitation in speech. *Frontiers in Psychology*, 4.

Gao, Y., Wang, Q., Ding, Y., Wang, C., Li, H., Wu, X., Qu, T., and Li, L. (2017). Selective attention enhances beta-band cortical oscillation to speech under "cocktail-party" listening conditions. *Frontiers in Human Neuroscience*, 11.

Garnier, M., Lamalle, L., and Sato, M. (2013). Neural correlates of phonetic convergence and speech imitation. *Frontiers in Psychology*, 4(SEP).

Gentilucci, M. and Bernardis, P. (2007). Imitation during phoneme production. *Neuropsychologia*, 45(3):608–615.

Geschwind, N. (1970). The organization of language and the brain. *Science*, 170(3961):940–944.

Ghitza, O. (2012). On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum. *Frontiers in psychology*, 3:238.

Gilakjani, A. P., Lai-Mei, L., and Sabouri, N. B. (2012). A study on the role of motivation in foreign language learning and teaching. *International Journal of Modern Education and Computer Science*, 4(7):9.

Giles, H. (1973). Accent mobility: A model and some data. *Anthropological Linguistics*, 15(2):87–105.

Giles, H., Coupland, N., and Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. *Contexts of Accommodation: Developments in Applied Sociolinguistics*, pages 1–68.

Giles, H., Mulac, A., Bradac, J. J., and Johnson, P. (1987). Speech accommodation theory: The first decade and beyond. *Annals of the International Communication Association*, 10(1):13–48.

Giles, H. and Smith, P. (1979). Accommodation theory: Optimal levels of convergence. language and social psychology, ed. by howard giles and robert n. st clair, 45–65.

Giordano, B. L., Ince, R. A., Gross, J., Schyns, P. G., Panzeri, S., and Kayser, C. (2017). Contributions of local speech encoding and functional connectivity to audio-visual speech perception. *Elife*, 6:e24763.

Giraud, A. L. and Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience*, 15(4):511–517.

Goldinger, S. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychological Review*, 105(2):251–279.

Gollan, T., Weissberger, G., Runnqvist, E., Montoya, R., and Cera, C. (2012). Self-ratings of spoken language dominance: A multilingual naming test (mint) and preliminary norms for young and aging spanish-english bilinguals. *Bilingualism*, 15(3):594–615.

Golumbic, E. M. Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., Simon, J. Z., et al. (2013a). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron*, 77(5):980–991.

Golumbic, E. M. Z., Poeppel, D., and Schroeder, C. E. (2012). Temporal context in speech processing and attentional stream selection: a behavioral and neural perspective. *Brain and language*, 122(3):151–161.

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

Gow Jr, D. W. and Segawa, J. A. (2009). Articulatory mediation of speech perception: a causal analysis of multi-modal imaging data. *Cognition*, 110(2):222–236.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., and Hämäläinen, M. S. (2013). Meg and eeg data analysis with mne-python. *Frontiers in Neuroscience*, 7(267):1–13.

Gregory Jr., S., Dagan, K., and Webster, S. (1997). Evaluating the relation of vocal accommodation in conversation partners' fundamental frequencies to perceptions of communication quality. *Journal of Nonverbal Behavior*, 21(1):23–43.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., and Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biology*, 11(12).

Haarmann, H. and Cameron, K. (2005). Active maintenance of sentence meaning in working memory: Evidence from eeg coherences. *International Journal of Psychophysiology*, 57(2):115–128.

Hagoort, P., Hald, L., Bastiaansen, M., and Petersson, K. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, 304(5669):438–441.

Hari, R. (2006). Chapter 17 action-perception connection and the cortical mu rhythm. *Progress in Brain Research*, 159:253–260.

Hari, R. and Kujala, M. (2009). Brain basis of human social interaction: From concepts to brain imaging. *Physiological Reviews*, 89(2):453–479.

Hasson, U., Ghazanfar, A., Galantucci, B., Garrod, S., and Keysers, C. (2012). Brain-to-brain coupling: A mechanism for creating and sharing a social world. *Trends in Cognitive Sciences*, 16(2):114–121.

Hauswald, A., Lithari, C., Collignon, O., Leonardelli, E., and Weisz, N. (2018). A visual cortical network for deriving phonological information from intelligible lip movements. *Current Biology*, 28(9):1453–1459.

Hickok, G. and Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393.

Hyafil, A., Giraud, A.-L., Fontolan, L., and Gutkin, B. (2015). Neural cross-frequency coupling: Connecting architectures, mechanisms, and functions. *Trends in Neurosciences*, 38(11):725–740.

Jenson, D., Bowers, A., Harkrider, A., Thornton, D., Cuellar, M., and Saltuklaroglu, T. (2014). Temporal dynamics of sensorimotor integration in speech perception and production: Independent component analysis of eeg data. *Frontiers in Psychology*, 5(JUL).

Jiang, J., Chen, C., Dai, B., Shi, G., Ding, G., Liu, L., Lu, C., and Fiske, S. (2015). Leader emergence through interpersonal neural synchronization. *Proceedings of the National Academy of Sciences of the United States of America*, 112(14):4274–4279.

Jiang, J., Dai, B., Peng, D., Zhu, C., Liu, L., and Lu, C. (2012). Neural synchronization during face-to-face communication. *Journal of Neuroscience*, 32(45):16064–16069.

Jungers, M. and Hupp, J. (2009). Speech priming: Evidence for rate persistence in unscripted speech. *Language and Cognitive Processes*, 24(4):611–624.

Kawasaki, M., Yamada, Y., Ushiku, Y., Miyauchi, E., and Yamaguchi, Y. (2013). Inter-brain synchronization during coordination of speech rhythm in human-to-human social interaction. *Scientific Reports*, 3.

Kayser, C., Wilson, C., Safaai, H., Sakata, S., and Panzeri, S. (2015). Rhythmic auditory cortex activity at multiple timescales shapes stimulus–response gain and background firing. *Journal of Neuroscience*, 35(20):7750–7762.

Kerlin, J. R., Shahin, A. J., and Miller, L. M. (2010). Attentional gain control of ongoing cortical speech representations in a "cocktail party". *Journal of Neuroscience*, 30(2):620–628.

Kinnunen, T. and Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 52(1):12–40.

Konvalinka, I., Bauer, M., Stahlhut, C., Hansen, L., Roepstorff, A., and Frith, C. (2014). Frontal alpha oscillations distinguish leaders from followers: Multivariate decoding of mutually interacting brains. *NeuroImage*, 94:79–88.

Kösem, A. and Van Wassenhove, V. (2017). Distinct contributions of low-and high-frequency neural oscillations to speech comprehension. *Language, Cognition and Neuroscience*, 32(5):536–544.

Kuhlen, A., Allefeld, C., and Haynes, J.-D. (2012). Content-specific coordination of listeners' to speakers' eeg during communication. *Frontiers in Human Neuroscience*, (SEPTEMBER).

Kutas, M. and Federmeier, K. (2011). Thirty years and counting: Finding meaning in the n400 component of the event-related brain potential (erp). *Annual Review of Psychology*, 62:621–647.

Kutas, M. and Hillyard, S. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205.

Lachat, F., Hugueville, L., Lemaréchal, J.-D., Conty, L., and George, N. (2012). Oscillatory brain correlates of live joint attention: A dual-eeg study. *Frontiers in Human Neuroscience*, (JUNE 2012).

Lakatos, P., Shah, A., Knuth, K., Ulbert, I., Karmos, G., and Schroeder, C. (2005). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *Journal of Neurophysiology*, 94(3):1904–1911.

Lee, J. H., Whittington, M. A., and Kopell, N. J. (2013). Top-down beta rhythms support selective attention via interlaminar interaction: a model. *PLoS computational biology*, 9(8):e1003164.

Leocani, L., Toro, C., Manganotti, P., Zhuang, P., and Hallett, M. (1997). Event-related coherence and event-related desynchronization/synchronization in the 10 hz and 20 hz eeg during self- paced movements. *Electroencephalography and Clinical Neurophysiology - Evoked Potentials*, 104(3):199–206.

Levinson, S. (2016). Turn-taking in human communication - origins and implications for language processing. *Trends in Cognitive Sciences*, 20(1):6–14.

Levinson, S. and Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6(JUN).

Levitan, R. and Hirschberg, J. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. pages 3081–3084.

Lewis, A. and Bastiaansen, M. (2015). A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. *Cortex*, 68:155–168.

Lewis, A., Schoffelen, J.-M., Schriefers, H., and Bastlaansen, M. (2016). A predictive coding perspective on beta oscillations during sentence-level language comprehension. *Frontiers in Human Neuroscience*, 10(MAR2016).

Lewis, A. G., Wang, L., and Bastiaansen, M. (2015). Fast oscillatory dynamics during language comprehension: Unification versus maintenance and prediction? *Brain and Language*, 148:51–63.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the h&h theory. *Speech Production and Speech Modelling*, pages 403–439.

Liu, Y., Piazza, E., Simony, E., Shewokis, P., Onaral, B., Hasson, U., and Ayaz, H. (2017). Measuring speaker-listener neural coupling with functional near infrared spectroscopy. *Scientific Reports*, 7.

Looze, C. D. and Rauzy, S. (2011). Measuring speakers' similarity in speech by means of prosodic cues: methods and potential. In *Interspeech 2011*, pages 1393–1396.

Luo, H. and Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, 54(6):1001–1010.

Mandel, A., Bourguignon, M., Parkkonen, L., and Hari, R. (2016). Sensorimotor activation related to speaker vs. listener role during natural conversation. *Neuroscience Letters*, 614:99–104.

Marian, V., Blumenfeld, H., and Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (leap-q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, 50(4):940–967.

Maris, E. and Oostenveld, R. (2007). Nonparametric statistical testing of eeg- and meg-data. *Journal of Neuroscience Methods*, 164(1):177–190.

Marslen-Wilson, W. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature*, 244(5417):522–523.

Megevand, P., Mercier, M. R., Groppe, D. M., Golumbic, E. Z., Mesgarani, N., Beauchamp, M. S., Schroeder, C. E., and Mehta, A. D. (2018). Phase resetting in human auditory cortex to visual speech. *bioRxiv*, page 405597.

Meister, I., Wilson, S., Deblieck, C., Wu, A., and Iacoboni, M. (2007). The essential role of premotor cortex in speech perception. *Current Biology*, 17(19):1692–1696.

Menenti, L., Pickering, M., and Garrod, S. (2012). Toward a neural basis of interactive alignment in conversation. *Frontiers in Human Neuroscience*, 6(185):1–9.

Meyer, L. (2018). The neural oscillations of speech processing and language comprehension: state of the art and emerging mechanisms. *European Journal of Neuroscience*, 48(7):2609–2621.

Meyer, L., Henry, M., Gaston, P., Schmuck, N., and Friederici, A. (2017). Linguistic bias modulates interpretation of speech via neural delta-band oscillations. *Cerebral Cortex*, 27(9):4293–4302.

Meyer, L., Obleser, J., and Friederici, A. (2013). Left parietal alpha enhancement during working memory-intensive sentence processing. *Cortex*, 49(3):711–721.

Mitterer, H. and Müsseler, J. (2013). Regional accent variation in the shadowing task: Evidence for a loose perception-action coupling in speech. *Attention, Perception, and Psychophysics*, 75(3):557–575.

Ménoret, M., Varnet, L., Fargier, R., Cheylus, A., Curie, A., des Portes, V., Nazir, T., and Paulignan, Y. (2014). Neural correlates of non-verbal social interactions: A dual-eeg study. *Neuropsychologia*, 55(1):85–97.

Molinaro, N. and Lizarazu, M. (2018). Delta(but not theta)-band cortical entrainment involves speech-specific processing. *European Journal of Neuroscience*, 48(7):2642–2650.

Morillon, B. and Baillet, S. (2017). Motor origin of temporal predictions in auditory attention. *Proceedings of the National Academy of Sciences of the United States of America*, 114(42):E8913–E8921.

Morillon, B., Liégeois-Chauvel, C., Arnal, L., Bénar, C.-G., and Giraud, A.-L. (2012). Asymmetric function of theta and gamma activity in syllable processing: An intra-cortical study. *Frontiers in Psychology*, 3(JUL).

Möttönen, R. and Watkins, K. E. (2009). Motor representations of articulators contribute to categorical perception of speech sounds. *The Journal of Neuroscience*, 29(31):9819–9825.

Mukherjee, S., D'Ausilio, A., Nguyen, N., Fadiga, L., and Badino, L. (2017). The relationship between f0 synchrony and speech convergence in dyadic interaction. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 2341–2345.

Mukherjee, S., Legou, T., Lancia, L., Hilt, P., Tomassini, A., Fadiga, L., D'Ausilio, A., Badino, L., and Nguyen, N. (2018). Analyzing vocal tract movements during speech accommodation. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*.

Murakami, T., Abe, M., Wiratman, W., Fujiwara, J., Okamoto, M., Mizuochi-Endo, T., Iwabuchi, T., Makuuchi, M., Yamashita, A., Tiksnadi, A., Chang, F.-Y., Kubo, H., Matsuda, N., Kobayashi, S., Eifuku, S., and Ugawa, Y. (2018). The motor network reduces multisensory illusory perception. *The Journal of Neuroscience*, 38(45):9679–9688.

Namy, L., Nygaard, L., and Sauerteig, D. (2002). Gender differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology*, 21(4):422–432+454–455.

Natale, M. (1975). Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology*, 32(5):790–804.

Nielsen, K. (2011). Specificity and abstractness of vot imitation. *Journal of Phonetics*, 39(2):132–142.

Nourski, K., Reale, R., Oya, H., Kawasaki, H., Kovach, C., Chen, H., Howard III, M., and Brugge, J. (2009). Temporal envelope of time-compressed speech represented in the human auditory cortex. *Journal of Neuroscience*, 29(49):15564–15574.

Nygaard, L., Sommers, M., and Pisoni, D. (1994). Speech perception as a talker-contingent process. *Psychological Science*, 5(1):42–46.

Ohala, J. J., Browman, C. P., and Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology*, 3:219–252.

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.-M. (2011). Fieldtrip: Open source software for advanced analysis of meg, eeg, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, 2011.

O'Sullivan, A. E., Crosse, M. J., Di Liberto, G. M., and Lalor, E. C. (2017). Visual cortical entrainment to motion and categorical speech features during silent lipreading. *Frontiers in human neuroscience*, 10:679.

O'sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., and Lalor, E. C. (2014). Attentional selection in a cocktail party environment can be decoded from single-trial eeg. *Cerebral Cortex*, 25(7):1697–1706.

Ozker, M., Yoshor, D., and Beauchamp, M. S. (2018). Frontal cortex selects representations of the talker's mouth to aid in speech perception. *eLife*, 7:e30387.

Pardo, J. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America*, 119(4):2382–2393.

Pardo, J. (2013). Measuring phonetic convergence in speech production. *Frontiers in Psychology*, 4(AUG).

Pardo, J., Gibbons, R., Suppes, A., and Krauss, R. (2012). Phonetic convergence in college roommates. *Journal of Phonetics*, 40(1):190–197.

Pardo, J., Jay, I., and Krauss, R. (2010). Conversational role influences speech imitation. *Attention, Perception, and Psychophysics*, 72(8):2254–2264.

Pardo, J., Urmanche, A., Wilman, S., and Wiener, J. (2016). Phonetic convergence and talker sex: It's complicated. *The Journal of the Acoustical Society of America*, 139(4):2105–2106.

Pardo, J., Urmanche, A., Wilman, S., and Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, and Psychophysics*, 79(2):637–659.

Park, H., Ince, R. A., Schyns, P. G., Thut, G., and Gross, J. (2015). Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. *Current Biology*, 25(12):1649–1653.

Park, H., Ince, R. A., Schyns, P. G., Thut, G., and Gross, J. (2018a). Representational interactions during audiovisual speech entrainment: Redundancy in left posterior superior temporal gyrus and synergy in left motor cortex. *PLoS biology*, 16(8):e2006558.

Park, H., Kayser, C., Thut, G., and Gross, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *eLife*, 5(MAY2016).

Park, H., Thut, G., and Gross, J. (2018b). Predictive entrainment of natural speech through two fronto-motor top-down channels. *bioRxiv*, page 280032.

Parkes, L. M., Bastiaansen, M. C. M., and Norris, D. G. (2006). Combining eeg and fmri to investigate the post-movement beta rebound. *NeuroImage*, 29(3):685–696.

Payton, K., Uchanski, R., and Braida, L. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *Journal of the Acoustical Society of America*, 95(3):1581–1592.

Peelle, J. E., Gross, J., and Davis, M. H. (2012). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral cortex*, 23(6):1378–1387.

Pelecanos, J. and Sridharan, S. (2001). Feature warping for robust speaker verification. *Proc. Speaker Odyssey*, pages 213–218.

Percival, D. B. and Walden, A. T. (1996). Spectral analysis for physical applications: Multitaper and conventional univariate techniques. *Technometrics*, 38(3):294–294.

Petersen, S. E. and Fiez, J. A. (1993). The processing of single words studied with positron emission tomography. *Annual review of neuroscience*, 16(1):509–530.

Petit, L., Simon, G., Joliot, M., Andersson, F., Bertin, T., Zago, L., Mellet, E., and Tzourio-Mazoyer, N. (2007). Right hemisphere dominance for auditory attention and its modulation by eye position: An event related fmri study. *Restorative Neurology and Neuroscience*, 25:211–225.

Petten, C. (1993). A comparison of lexical and sentence-level context effects in event-related potentials. *Language and Cognitive Processes*, 8(4):485–531.

Pfurtscheller, G. and Aranibar, A. (1979). Evaluation of event-related desynchronization (erd) preceding and following voluntary self-paced movement. *Electroencephalography and Clinical Neurophysiology*, 46(2):138–146.

Pfurtscheller, G. and Berghold, A. (1989). Patterns of cortical activation during planning of voluntary movement. *Electroencephalography and Clinical Neurophysiology*, 72(3):250–258.

Pfurtscheller, G. and da Silva, F. L. (1999). Event-related eeg/meg synchronization and desynchronization: Basic principles. *Clinical Neurophysiology*, 110(11):1842–1857.

Pfurtscheller, G., Pregenzer, M., and Neuper, C. (1994). Visualization of sensorimotor areas involved in preparation for hand movement based on classification of mu and central beta rhythms in single eeg trials in man. *Neuroscience Letters*, 181(1):43–46.

Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.

Pickering, M. J. and Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36(4):329–347.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech communication*, 41(1):245–255.

Posner, M. and Raichle, M. (1994). Images of mind. scientific american library.

Pérez, A., Carreiras, M., and Duñabeitia, J. (2017). Brain-to-brain entrainment: Eeg interbrain synchronization while speaking and listening. *Scientific Reports*, 7(1).

Price, C. (2012). A review and synthesis of the first 20years of pet and fmri studies of heard speech, spoken language and reading. *NeuroImage*, 62(2):816–847.

Price, C. J. (2010). The anatomy of language: a review of 100 fmri studies published in 2009. *Annals of the new York Academy of Sciences*, 1191(1):62–88.

Pulvermüller, F. (2018). Neural reuse of action perception circuits for language, concepts and communication. *Progress in neurobiology*, 160:1–44.

Ramseyer, F. and Tschacher, W. (2010). Nonverbal synchrony or random coincidence? how to tell the difference. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5967 LNCS:182–196.

Reynolds, D. (1992). A gaussian mixture modeling approach to text-independent speaker identification. *A Gaussian Mixture Modeling Approach to Text-independent Speaker Identification*.

Reynolds, D., Quatieri, T., and Dunn, R. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing: A Review Journal*, 10(1):19–41.

Richmond, K. (2006). A trajectory mixture density network for the acoustic-articulatory inversion mapping. In *Ninth International Conference on Spoken Language Processing*.

Riecke, L., Formisano, E., Sorger, B., Başkent, D., and Gaudrain, E. (2018). Neural entrainment to speech modulates speech intelligibility. *Current Biology*, 28(2):161–169.

Riecke, L., Sack, A., and Schroeder, C. (2015). Endogenous delta/theta sound-brain phase entrainment accelerates the buildup of auditory streaming. *Current Biology*, 25(24):3196–3201.

Sadjadi, S. O., Slaney, M., and Heck, L. (2013). Msr identity toolbox v1.0: A matlab toolbox for speaker recognition research. *Speech and Language Processing Technical Committee Newsletter*.

Salaberry, M. R. (2001). The use of technology for second language learning and teaching: A retrospective. *The modern language journal*, 85(1):39–56.

Salmelin, R., Hámáaláinen, M., Kajola, M., and Hari, R. (1995). Functional segregation of movement-related rhythmic activity in the human brain. *NeuroImage*, 2(4):237–243.

Samuel, A. and Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, and Psychophysics*, 71(6):1207–1218.

Sanchez, K., Miller, R., and Rosenblum, L. (2010). Visual influences on alignment to voice onset time. *Journal of Speech, Language, and Hearing Research*, 53(2):262–272.

Sato, M., Tremblay, P., and Gracco, V. L. (2009). A mediating role of the premotor cortex in phoneme segmentation. *Brain and Language*, 111(1):1–7.

Savariaux, C., Badin, P., Samson, A., and Gerber, S. (2017). A comparative study of the precision of carstens and northern digital instruments electromagnetic articulographs. *Journal of Speech, Language, and Hearing Research*, 60(2):322–340.

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., and Vogeley, K. (2013). Toward a second-person neuroscience. *Behavioral and Brain Sciences*, 36(4):393–414.

Schmitz, J., Bartoli, E., Maffongelli, L., Fadiga, L., Sebastian-Galles, N., and Dâ€™Ausilio, A. (2018). Motor cortex compensates for lack of sensory and motor experience during auditory speech perception. *Neuropsychologia*.

Schoot, L., Hagoort, P., and Segaert, K. (2016). What can we learn from a two-brain approach to verbal interaction? *Neuroscience and Biobehavioral Reviews*, 68:454–459.

Schroeder, C. E. and Lakatos, P. (2009). Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in neurosciences*, 32(1):9–18.

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., and Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends in cognitive sciences*, 12(3):106–113.

Schwartz, J.-L. and Savariaux, C. (2014). No, there is no 150 ms lead of visual speech on auditory speech, but a range of audiovisual asynchronies varying from small audio lead to large audio lag. *PLoS Computational Biology*, 10(7):e1003743.

Scovel, T. (1969). Foreign accents, language acquisition, and cerebral dominance 1. *Language learning*, 19(3-4):245–253.

Seliger, H., Krashen, S., and Ladefoged, P. (1982). Maturational constraints in the acquisition of second languages. *Child-adult differences in second language acquisition*, 2:13.

Selinker, L. (1972). Interlanguage. *IRAL-International Review of Applied Linguistics in Language Teaching*, 10(1-4):209–232.

Shockley, K., Sabadini, L., and Fowler, C. (2004). Imitation in shadowing words. *Perception and Psychophysics*, 66(3):422–429.

Silbert, L., Honey, C., Simony, E., Poeppel, D., and Hasson, U. (2014). Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences of the United States of America*, 111(43):E4687–E4696.

Smith, Z. M., Delgutte, B., and Oxenham, A. J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, 416(6876):87–90.

Soto-Faraco, S., Navarra, J., Weikum, W. M., Vouloumanos, A., Sebastián-Gallés, N., and Werker, J. F. (2007). Discriminating languages by speech-reading. *Perception & Psychophysics*, 69(2):218–231.

Sperber, D. and Wilson, D. (1998). *The mapping between the mental and the public lexicon*, page 184–200. Cambridge University Press.

Stekelenburg, J. J. and Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience*, 19(12):1964–1973.

Stephens, G., Silbert, L., and Hasson, U. (2010). Speaker-listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences of the United States of America*, 107(32):14425–14430.

Stevens, K. (1989). On the quantal nature of speech. *Journal of Phonetics*, 17:3–45.

Stolk, A., Verhagen, L., and Toni, I. (2016). Conceptual alignment: How brains achieve mutual understanding. *Trends in Cognitive Sciences*, 20(3):180–191.

Sumby, W. H. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The journal of the acoustical society of america*, 26(2):212–215.

Suzuki, N. and Katagiri, Y. (2007). Prosodic alignment in human-computer interaction. *Connection Science*, 19(2):131–141.

Szymanski, F., Rabinowitz, N., Magri, C., Panzeri, S., and Schnupp, J. (2011). The laminar and temporal structure of stimulus information in the phase of field potentials of auditory cortex. *Journal of Neuroscience*, 31(44):15787–15801.

Tognoli, E. and Kelso, J. (2015). The coordination dynamics of social neuromarkers. *Frontiers in Human Neuroscience*, 9(OCTOBER).

Tranter, S. E. and Reynolds, D. A. (2006). An overview of automatic speaker diarization systems. *IEEE Transactions on audio, speech, and language processing*, 14(5):1557–1565.

Trofimovich, P. and Kennedy, S. (2014). Interactive alignment between bilingual interlocutors: Evidence from two information-exchange tasks *. *Bilingualism: Language and Cognition*, 17(4):822–836.

Turner, L. H. and West, R. (2010). Communication accommodation theory. *Introducing communication theory: Analysis and application*, 4.

Uther, M., Knoll, M., and Burnham, D. (2007). Do you speak e-ng-l-i-sh? a comparison of foreigner- and infant-directed speech. *Speech Communication*, 49(1):2–7.

Van Wassenhove, V., Grant, K. W., and Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences*, 102(4):1181–1186.

Vander Ghinst, M., Bourguignon, M., de Beeck, M. O., Wens, V., Marty, B., Hassid, S., Choufani, G., Jousmäki, V., Hari, R., Van Bogaert, P., et al. (2016). Left superior temporal gyrus is coupled to attended speech in a cocktail-party auditory scene. *Journal of Neuroscience*, 36(5):1596–1606.

Vaughan, B. (2011). Prosodic synchrony in co-operative task-based dialogues: A measure of agreement and disagreement.

Vigneau, M., Beaucousin, V., Herve, P.-Y., Duffau, H., Crivello, F., Houde, O., Mazoyer, B., and Tzourio-Mazoyer, N. (2006). Meta-analyzing left hemisphere language areas: phonology, semantics, and sentence processing. *Neuroimage*, 30(4):1414–1432.

Wang, X., Lu, T., and Liang, L. (2003). Cortical processing of temporal modulations. *Speech Communication*, 41(1):107–121.

Ward, A. and Litman, D. (2007). *Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora*.

Warschauer, M. and Healey, D. (1998). Computers and language learning: An overview. *Language teaching*, 31(2):57–71.

Watkins, K., Strafella, A., and Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, 41(8):989–994.

Weiss, S., Mueller, H., Schack, B., King, J., Kutas, M., and Rappelsberger, P. (2005). Increased neuronal communication accompanying sentence comprehension. *International Journal of Psychophysiology*, 57(2):129–141.

Wernicke, C. (1974). Der aphasische symptomenkomplex. In *Der aphasische Symptomencomplex*, pages 1–70. Springer.

Wu, Z., Zhao, K., Wu, X., Lan, X., and Meng, H. (2015). Acoustic to articulatory mapping with deep neural network. *Multimedia Tools and Applications*, 74(22):9889–9907.

Zoefel, B., Archer-Boyd, A., and Davis, M. H. (2018). Phase entrainment of brain oscillations causally modulates neural responses to intelligible speech. *Current Biology*, 28(3):401–408.