Università degli Studi di Ferrara

ISTITUTO ITALIANO DI TECNOLOGIA

DOCTORAL COURSE IN

TRANSLATIONAL NEUROSCIENCE AND NEUROTECHNOLOGIES

CYCLE XXXIV

COORDINATOR Prof. Fadiga Luciano

PERCEPTION, PRODUCTION AND INTERACTION: A COMPREHENSIVE INVESTIGATION ON HUMAN SPEECH

Scientific/Disciplinary Sector (SDS) BIO/09

Candidate

Dott. Pastore Aldo

Supervisor

Prof. D'Ausilio Alessandro

Year 2018/2021

Acknowledgments

Questo manoscritto e tutto il lavoro e la fatica che hanno portato alla sua realizzazione e al conseguimento di questo mio grande obbiettivo di vita sono interamente dedicati a te.

Tante volte avrei voluto interrompere il mio lavoro per non togliere più neanche un istante al nostro tempo insieme. Non hai mai voluto, hai sempre messo davanti l'amore per me a tutto il resto.

Non so se lo rifarei, forse no, non te lo consentirei. Potessi tornare indietro forse queste pagine non esisterebbero, sostituite da una manciata di foto in più.

Ad ogni modo, credo che se ora le tue dita potessero sfogliare queste pagine saresti felice ed orgogliosa. Immagino il tuo sorriso e la tua gioia. Quest'idea mi rincuora, un po'.

Vorrei fossi li seduta in una sedia a guardarmi Venerdì, vorrei poterti abbracciare e sentire la tua allegria e perchè no magari anche qualche imbarazzante esultanza da stadio. Immaginerò tutto.

In qualche modo è come se con la sottomissione ufficiale di questo elaborato si chiudesse qualcosa di molto più grande di un corso di Dottorato. Una parte di vita ed anche di me stesso, che non esisterà più. E che farà in eterno sentire la sua rumorosa assenza.

Senza di te, della tua luce e dei tuoi insegnamenti, molto di tutto ciò che sono e che ho fatto di bello fino ad'ora compreso questo lavoro, di cui vado molto fiero, non esisterebbe.

"La vita non è aspettare che passi la tempesta ma imparare a ballare sotto la pioggia". Lo sentivi così tuo, e lo hai reso anche un po' mio.

Grazie per tutto quello che hai fatto e per tutto quello che sei stata per me. Non smetterò mai di essertene grato.

Ti amo e questo sarà eterno.

Buona lettura cicciona mia.

"Il sole che splende sul mondo si è addormentato.
Ma il suo riflesso gentile illuminerà per sempre la mia via."

Contents

1 General Introduction

1.1 The human speech

Speech is the way humans communicate using spoken language. Through speech we express our thoughts, and this is possible via a series of complex movements of the speech articulators that modify the basic tone of the voice into comprehensible sounds.

In particular, muscles in the mouth, face, neck, chest, and abdomen have to be coordinated by the brain in order to handle one of the most sophisticated form of movement control.

The energy needed to produce the speech is provided, in the form of an airstream, by a bellows-like respiratory activator; the energy is then transformed in the larynx by a phonating sound generator and subsequently voice pattern is shaped in the pharynx; finally speech is formed by the articulators in the mouth.

The evolutionary origins of speech are unknown and subject to much debate and speculation. A variety of animals are able to communicate using vocalizations and trained apes can also use simple sign language. However, humans are the only specie capable to articulate phonemically and syntactically to constitute speech.

1.2 Verbal communication: a joint task

Humans are the only animals in the animal kingdom using an articulated language as a communicative tool. Indeed, speech is the main communicative mean that enables interactions in social contexts and it is used for different purposes ranging from basic needs to more complex concepts.

Communication can be defined as an intentional joint action that has a shared goal between speakers. It is considered successful when two speakers come to the same understanding of different relevant aspects of the world, in other words they align their representation of the situation under discussion (Brown-Schmidt & Tanenhaus, 2008). This is typically achieved as people start off at a very good point by communicating with other people who are largely similar to themselves, both because they process language in similar ways and because they share much relevant background knowledge. This means that they can, in principle, use knowledge about themselves to understand and, in particular, predict their interlocutor.

Dialog as a joint action should be seen as both intentional (conveying meaning) and unintentional (automatic), even though interlocutors might have a shared goal. For example, underlying communication is the automatic process of imitation (Pickering & Garrod, 2004) that takes place both at the level of vocabulary choices or grammatical structures. It has been argued that this effect arises from common coding between interlocutors across production and comprehension, without extensive negotiation between interlocutors or modelling of each other's mental states. The same aspect applies to other modalities, such as posture and gestures (Shockley, Santana, & Fowler, 2003), but also to the tendency of laughing and yawning together (Hatfield, Cacioppo, & Rapson, 1994). This clearly means that interlocutors construct aligned non-linguistic representations.

In social interactions, indeed, people coordinate their actions in order to incrementally and interactively reach their communicative goals. This is made by repetition of communicative behaviour (body postures, eye gaze, words, gestures) in order for interlocutors to tune with each other. This mechanism is named *"alignment"* or *"convergence"*. How alignment happens is still under investigation. Natural communication ranges from gestures to speech, creating a multimodal way of communicating. For example, co-speech gestures are meaningful movements of the hand or arm that accompany speech and support joint problem-solving and coordination (Holler & Wilkin, 2011; Pickering & Garrod, 2004). This is a starting point to investigate alignment, however the interdisciplinary nature of such phenomenon makes it difficult to build a comprehensive framework.

1.2.1 The alignment theoretic framework: Priming and grounding

Theoretical approaches to alignment are divided into two categories giving rise to the priming and grounding dichotomy. Priming accounts define alignment as a individual-level mechanism (Pickering & Garrod, 2004, 2006). Grounding theories instead support the idea that alignment arises from interaction and coordination during joint meaning-making (Brennan & Clark, 1996; Holler & Wilkin, 2011).

Both priming and grounding approaches aimed at disentangling the complex phenomenon of alignment and they are concerned with alignment of behaviour as well as higher level mental representations. However, they diverge in defining the involvement of different levels of linguistic processing. On one hand, priming theorizes that speakers non only align their speaking behaviour but also their linguistic knowledge (Pickering & Garrod, 2006). Moreover, the priming mechanism is believed to happen at multiple linguistic levels ranging from phonetics to semantics. On the other hand, grounding accounts believe that alignment of linguistic representation is not a requisite for alignment at other levels of representation. Thus, grounding framework provides a more flexible relationship between behavioural alignment in various modalities and alignment of conceptual representations, whereas priming accounts see them as causally linked.

1.2.2 The alignment theoretic framework: how does alignment work?

Four main theories have been proposed to explain the mechanism underlying the alignment: (1) the Episodic Theory (ET) (Goldinger,1998) of speech perception and production; (2) the motor theory (MT) of speech perception (Liberman & Whalen, 2000); (3) Communication Accommodation Theory (CAT; Giles & Coupland, 1991); (4) Simulation Theory (ST) (Gambi & Pickering, 2013). Firstly, ET theory posits that anything perceived by the individual leaves a trace in memory that contains detailed phonetic information, such as the speaker's voice characteristics. This echo of the perceptual experience that is left in memory can be shaped by more recent perceptual events, and this affects production of the same word.

Secondly, the Motor Theory of perception argues that perceptual units of speech processing are not defined by speech's acoustic characteristics, but rather by articulatory gestures (Liberman & Whalen, 2000). Galantucci and colleagues (2006) proposed that imitation should be facilitated over non-imitative responses so that, when the same speech unit has been already perceived, speakers produce it faster. This phenomenon seems to be true across different modalities such as vision and audition (Fowler et al., 2003; Galantucci et al., 2009; Jarick & Jones, 2008; Kerzel & Bekkering, 2000;).

Thirdly, the CAT poses that speech convergence arises from the speaker's motivation to make the conversation more likeable to the conversational partner (Giles & Coupland, 1991). In this context, socially-relevant variables like personality traits (Natale, 1975), in-group and out-group differences (Giles, 1973), communicative intentions and affective goals (Giles & Coupland, 1991), social approval (Natale, 1975) are of crucial importance.

Finally, the Simulation Theory aims at summing up all the previously proposed theories in a more integrated framework of speech perception. It additionally argues that forward models of the motor commands map as well sensory consequences of executing such motor commands by making predictions.

According to pioneering work by Wolpert and colleagues (Wolpert & Flanagan, 2001; Wolpert et al., 2003), forward models allow for online control of one's own actions. Researchers in the speech field propose that a similar mechanism might be put into place during speech articulatory control (Guenther et al., 2006; Hickok, 2012; Tian & Poeppel, 2010).

Simulation of other people's speech might be enabled by a forward model running in one's own production system. Pickering and Garrod (2013) proposed that a combination of inverse and forward models might drive speech comprehension. Indeed, once a speech input is received, this input is inversely mapped from its perceptual representation to a production command that is the same command the individual who is listening would use to produce the speech himself.

Once this production command is derived, the listener has to pick the production motor command that is more similar to the motor command recovered via the

inverse model. This leads to a cascade effect that switches on the forward production model and forward comprehension model which together predict the upcoming input. These models depend on the characteristics of the listener's speech production architecture, with the forward speech production model computing predictions about articulatory movements, and the forward comprehension model dealing with acoustic features that would produce that specific combination of articulatory movements. This goes hand in hand with the definition of speech as joint action. Indeed, one of the fundamental aspects of joint actions is that they not only require understanding between partners but also prediction of those actions.

The evidence that mouth articulations, such as lips and tongue, involved in speech comprehension, activate while listening to speech but not during non-speech (Fadiga et al., 2002; Watkins, Strafella, & Paus, 2003) corroborates this hypothesis. Hence, it seems that during speech perception listeners use the production system as part of an emulator that operates in real time.

1.2.3 Acoustic Convergence: tuning two voices towards a common acoustic point

Research in the speech interaction field focused on different aspects of speech accommodation during interaction naming them mainly convergence or alignment. However, a distinction has to be made: one type of accommodation relies on cognitive, physiological, functional, and social constraints (Littlejohn & Foss, 2010), the second one on linguistic and paralinguistic factors (Heldner & Edlund, 2010).

The latter applies to synchronization between two speakers' acoustic features, and takes the form of a synchronized variation of speakers' voices. An example of this phenomenon is the instantaneous diminution of one speaker's voice amplitude that often happens as a consequence of the same volume diminution in the partner's voice.

In the other hand, convergence rests on the first definition of accommodation, and in contrast with synchrony is a long term mechanism arising during the conversation by means of the tuning of two speakers' speech towards a common acoustic point.

A pitfall of convergence research lays upon the fact that it has been always defined as a linear process where convergence is higher as the time passes. However, two speakers do not engage with the same involvement throughout the conversation, but convergence is instead time-varying (De Looze et al., 2014; Levitan & Hirschberg, 2011; Vaughan, 2011). This defines convergence as a more complex and dynamic phenomenon compared to what previous researchers believed, thus laying the bases for further research in the field.

During my research work I approached the problem of measuring the acoustic speech convergence. Due to the mentioned complexity of this time-varying mechanism obtaining a unified quantitative measure remains an unresolved problem.

To deal with the criticalities of the scientific question, i.e. with the lack of a quantitative mathematical definition and the convergence non-linear temporal dynamics, I decided to face the problem using an innovative deep learning approach, based on Siamese neural networks. The Siamese networks are a specific kind of deep learning models particularly well suited for dealing with time-varying data sequences and capable to learn measures of similarity. With this powerful tool, I built a mathematical model that given a couple of speakers is fully capable to compute the *"distance"* between the speaker's speech independently of the words pronounced. In future the model could be improved in order to be fully speaker independent and used to track and the raise and diminution of every couple of voices interacting in a verbal conversation.

1.3 The neural basis of speech processing: two streams and the Wernicke's model

The speech functional anatomy has been a research topic for more than one century. Nonetheless, the neural organization of speech perception and production remains an open field, and the characterization of the anatomic speech areas has been difficult to describe even in gross terms.

The first important speech related brain areas, known as Broca's and Wernicke's areas, have been discovered by studying people who developed speech disorders after severe injuries. Such discoveries were crucial in speech and writing understanding. However, studying speech in healthy subjects were complicated and animal models did not provide useful insights to disentangle this extraordinary complex process.

In 1870's the first hypothesis on how the brain perceives and produces the speech has been formulated, the Wernicke's model was born. The idea that characterizes this first model was intuitive and straightforward: the auditory cortex supports the speech perception, and from there two different speech processing pathway depart (Wernicke, Cohen & Wartofsky, 1874).

The Wernicke's model is now considered obsolete, but it had the intuition to lay the foundation for modern dual streams speech modelling, introducing the speech perception and the speech production pathways.

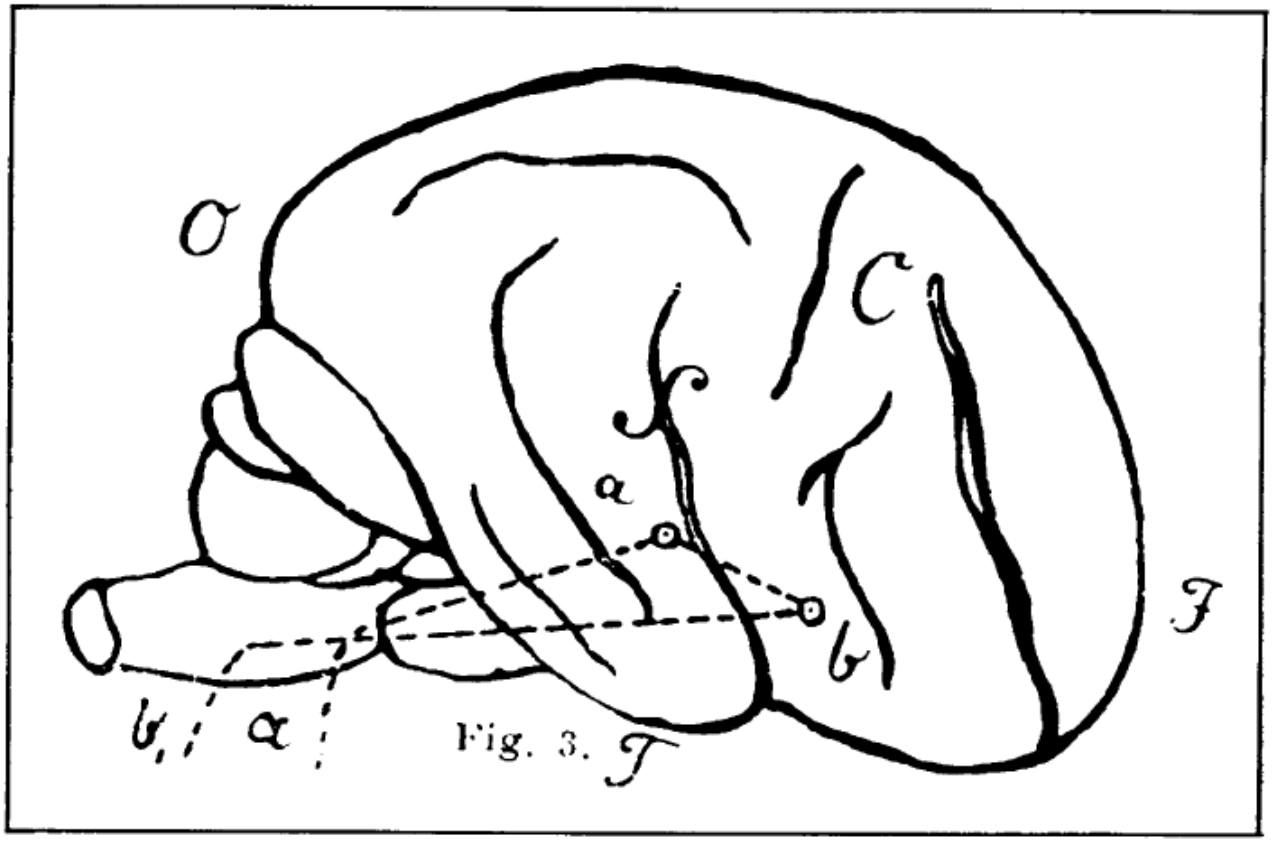

Figure 1. Wernicke's model of speech processing 1874. For speech perception, sounds are sent via the auditory system to the Broadmann area 41, i.e. the primary auditory cortex. This area is then connected to the Wernicke's area, where the words meaning is extracted. For speech production, the words' meaning is sent from the Wernicke's area to the Broca's area, where morphemes are assembled, and a words representation is retained. Finally, the speech motor instructions are sent from the Broca's area to the face related motor cortex and from here to facial motor neurons, in the brainstem, connected to the muscles to activate.

1.4 The speech Perception

The modern dual-stream model of speech processing (Hickok & Poeppel, 2000, 2004, 2007) lays its foundations in older models, such as the classic model of Wernicke. According to this model, speech information is processed alongside two different routes in the brain: the ventral stream and the dorsal stream (see Figure 2 for details of anatomical locations).

The dorsal stream is in charge of translating acoustic speech signals into articulatory-based representations for speech production. On the other hand, the auditory-conceptual ventral stream comprises the superior and middle portions of the temporal lobe and governs the processing speech signals for comprehension. It is bilaterally organised but computationally asymmetric (Abrams et al., 2008; Boemio et al., 2005; Giraud et al., 2007; Hickok & Poeppel, 2007; Zatorre et al., 2002). Indeed, left and right hemispheres seems to differ in their preferred *"sampling rate"* of acoustic stimuli. The left hemisphere is mainly involved in the processing of rapid acoustic changes (faster rate, 25-50 Hz), whereas the right hemisphere prefers lower sampling rates around 4-8 Hz (Poeppel, 2003; Zatorre et al., 2002), thus having a bias in processing spectral frequency information.

In general, the ventral stream governs not only phonological processing, but also lexical-semantic access via the temporal lobe that functions as a computational interface that stores correspondences between phonologic information and conceptual information. (Hickok & Poeppel, 2000, 2004, 2007; Lau et al., 2008).

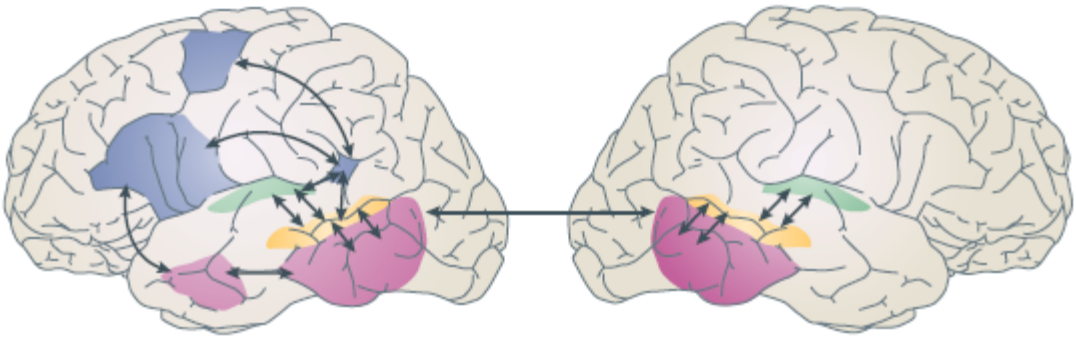

Figure 2. The dual stream model and its anatomical regions (Hickok & Poeppel, 2000, 2004, 2007). Blue regions are the dorsal stream, strongly left dominant, it involves Broca's area in the frontal lobe, a dorsal premotor site and a region at the parieto-temporal boundary (area Spt) thought to be a sensorimotor interface. The pink shaded areas represent the ventral stream. It is bilaterally organized with the more posterior region representing a lexical interface, which links phonological and semantic information. Yellow region is involved in phonological-level processes, and it is directly connected to a region from where both ventral and dorsal streams are originated (green- shaded area). This area is located bilaterally on the dorsal surface of the superior temporal gyrus and is proposed to be active in the early stages of speech processing, in particular in some form of spectro-temporal analysis. The figure is authored by Hickok and Poeppel (2007).

1.4.1 Neural entrainment and the importance of the rhythmic speech features for perception

The quasi-rhythmic structure of the input is crucial for the listening brain, indeed, speech comprehension is possible by speech envelope tracking, via a mechanism called neural entrainment (Ahissar et al., 2001; Ding & Simon, 2014; Luo & Poeppel, 2007; Obleser & Kayser 2019; Peelle & Davis, 2012; Peelle, Gross & Davis, 2013).

The neural entrainment to the speech is characterized by the synchronization of the neural oscillatory activity over the auditory areas to the rhythmicity in the incoming speech signal. The involved cortical regions use the rhythmic properties of the speech to allow the listener to discretise the continuous input into segmented units that undergo the decoding process (Peelle & Davis, 2012).

Research has found neural entrainment to the speech envelope in the delta (1-4 Hz) and theta (4-8 Hz) band (Brohl & Kayser, 2021), reflecting different speech components such as the occurrence of phonemes, words and even slower features such as phrases and the speech prosody (Bourguignon et al., 2013; Ding et al., 2017).

The importance of the speech entrainment to speech comprehension has been demonstrated by several studies that observed the correlation between the entrainment strength and the comprehension performances (Ahissar et al., 2001; Luo & Poeppel, 2007; Peelle et. al. 2013). In general the neural speech entrainment is lower when speech intelligibility is lower as well, as happens in a noisy environment (Ding & Simon 2014; Ghitza, 2012; Kayser et al., 2015; Peelle et. al. 2013; Riecke et al., 2018).

In this context entrainment to non-acoustic rhythmic stimuli, still related to the verbal message, increase its strength: this is the case of the entrainment to the movement of mouth articulators, such as the speaker's lips (Park et al., 2016; Peelle & Sommers, 2015). Entrainment is indeed a phenomenon not only involving the audio signal rather to every signal containing rhythmic properties.

The finding that the auditory cortex and the input rhythmically synchronise, and that this synchronisation correlates with intelligibility and comprehension of the auditory input, has led scientists to implement a neural oscillator model of the auditory cortex (Doelling et al. 2019). It is argued that there is a population of neurons that oscillates at a specific frequency, and that this frequency shifts in order to synchronise with the frequency of the external stimulation only when within the specific range of the oscillator.

In my research work I went through the analysis of the neural entrainment to the speech and in addition I investigated the neural entrainment to non-acoustic speech related signals, i.e. the movements of the speech articulators (tongue, lips and jaw). In particular, the aim of my research in the speech perception field is to investigate the presence of entrainment to speech-related kinematics even if not provided to the listener. The presence of such a neural entrainment would indeed be a fundamental hint of the presence a motor-driven simulation process going on during speech perception. With this aim, I settled down an experiment where

subjects were listening for audio sentences but did not have visual or kinematic speech related information.

Nonetheless, due to the fact that the speech is produced by the movement of the mouth articulators, the possibility that the eventually observed brain to kinematics entrainment is a redundant phenomenon rather than unique was cleared.

1.5 The speech Production

The production of the speech is a mechanism that is composed by multiple unconscious computational steps, starting from the abstract idea to the vocalization of the words. Firstly, the words to pronounce are selected with the correct lexicon and morphology, then organized trough the syntax. Subsequently the phonetic representation of the words is retrieved and the sentence is articulated through the articulations associated with those phonetic properties. (Levelt, 1999).

From a neural point of view, all these computational steps were for the first time formalized in the classic Wernicke's model. Here the speech production was managed by the dorsal stream, and precisely was articulated as follow: the meaning of the words are sent from the Wernicke's area to the Broca's area where morphemes are assembled, and a representation of the words is stored. Words to pronounce are subsequently sent to the facial motor cortex and finally to the motor-neurons into the brainstem that produce the movement commands.

After this classic model of speech production, several models have been described. However, these models considered the central nervous system still in a pure feedforward view, i.e. as a simple generator of the mouth articulators motor trajectories (Ostry et al., 1991; Perrier et al., 1996; Payan and Perrier, 1997; Sanguineti et al., 1997, 1998).

In general, the main problems faced by these models were the short latency of the speech adjustment commands and the fact that some adjustments are task dependent. Latencies are indeed too long to represent a pure local feedback loop

(Kandel et al., 2000) as well as task dependent adjustments result to be hardly explained by the same low-level mechanics (Kelso et al., 1984; Saltzman & Munhall, 1989).

Nowadays, speech production models are overcoming the idea that the central nervous system only produces the feedforward motor commands. Central feedback components are indeed increasingly integrated.

Several experiments indeed demonstrates the importance of the sensory feedback for speech production. For instance, perturbations in the amplitude of the feedback (the listened self-produced speech) or in the noise level lead to automatic compensation of the speech loudness (Lane & Tranel, 1971); mechanical perturbations of the speech articulators are also automatically compensated by the brain in order to reach the correct speech outcome (Abbs and Gracco, 1984; Saltzman et al., 1998; Shaiman & Gracco, 2002).

Additionally, a variety of studies have observed important component of the dorsal stream, i.e the posterior superior temporal gyrus (Zheng, Munhall & Johnsrude, 2009) and the superior parietal temporal area (Buchsbaum et al., 2001; Hickok et al., 2003) implicated in feedback processing specifically related to speech production.

All these evidences led to redefine the idea of the dorsal stream described in the classic dual streams model. In modern models, indeed, the dorsal stream serves for feedback processing related to the speech production (Hickok & Poeppel, 2007; Rauschecker & Scott, 2009; Hickok et al., 2011).

Consequently, the speech production process has been fitted on the idea that the brain uses attended auditory information, comparing the incoming feedback with a prediction derived from efference copies of the motor output, with the resulting prediction error used to track the state of the vocal tract. This modelling of the speech production is called State Feedback Control (Houde & Nagarajan, 2011; Bernstein, 1967) and it states that the auditory information is not only used for speech comprehension but also for speech production.

1.5.1 The Broca's area

It is widely recognized that the Broca's area is an important region for speech processing. Nonetheless, although several studies and more than one century of research, its function is still not completely cleared, and it remains one of the most debated language related area. The French neurologist Paul Broca was the first that described it, observing a patient with a lesion in the posterior part of the frontal convolution of the left hemisphere (Broca, 1861). Precisely, the Broca's area corresponds to Brodmann's areas (BA) 44 and 45 on the Broadmann's classification of cortical areas(see Figure 3, Friederici*, 2011*) of the left hemisphere. Although classically associated with language, the specific functions of Broca's area and exact functional connection to other brain regions are open research questions. (Friederici, 2011).

Nowadays, the crucial relevance of the Broca's area for speech production can be practically assessed by Direct Electrical Stimulation (DES). When applied over this eloquent region, indeed, the so called speech arrest can be induced. This phenomenon consists in the complete interruption of the ongoing speech, by the interruption of orofacial movements and sounds emission (Ferpozzi et al., 2018). It is a reversible phenomenon and allows the exact identification of the Broca's area (Chang et al., 2017). This procedure is very important for research purposes as well as for clinical ones. It is indeed used when a brain surgery has to be performed but language related areas have to be identified and preserved (Mandonnet, Sarubbo, & Duffau, 2017).

1.5.2 The anticipatory speech related High-Gamma activity over Broca's region

High gamma activity (70-150 Hz) has been observed to be engaged when the speech is both perceived and produced (Crone, Boatman, Gordon & Hao, 2001; Crone et al., 2001; Towle et al., 2008).

Recently, the ElectroCorticography (ECoG), has been used to investigate the functional role of Broca's area, as well as of the motor cortex and Superior

Temporal Gyrus (STG), during word production (Flinker et al., 2015). The advantage of using ECoG respect to standard non-invasive techniques is the possibility to exploit an higher spatial resolution and higher frequency components, such as the high gamma band, inaccessible to non-invasive neural recordings.

In their study Flinker and colleagues (2015) found power activity in the high-gamma band immediately before the speech onset over the Broca's area electrodes. After the speech onset, the activity observed over this region vanished and subsequently arose in the motor cortex area.

During my work I investigated the importance of the high gamma activity arising before the speech onset for the speech production, and the possibility to exploit it to build a brain-machine interface capable to decode the patient's intention to speak. The study that I conducted could be an important step to reach translational application for speech impaired patients and to deeper understand the role of the preparatory high gamma activity over the Broca's area.

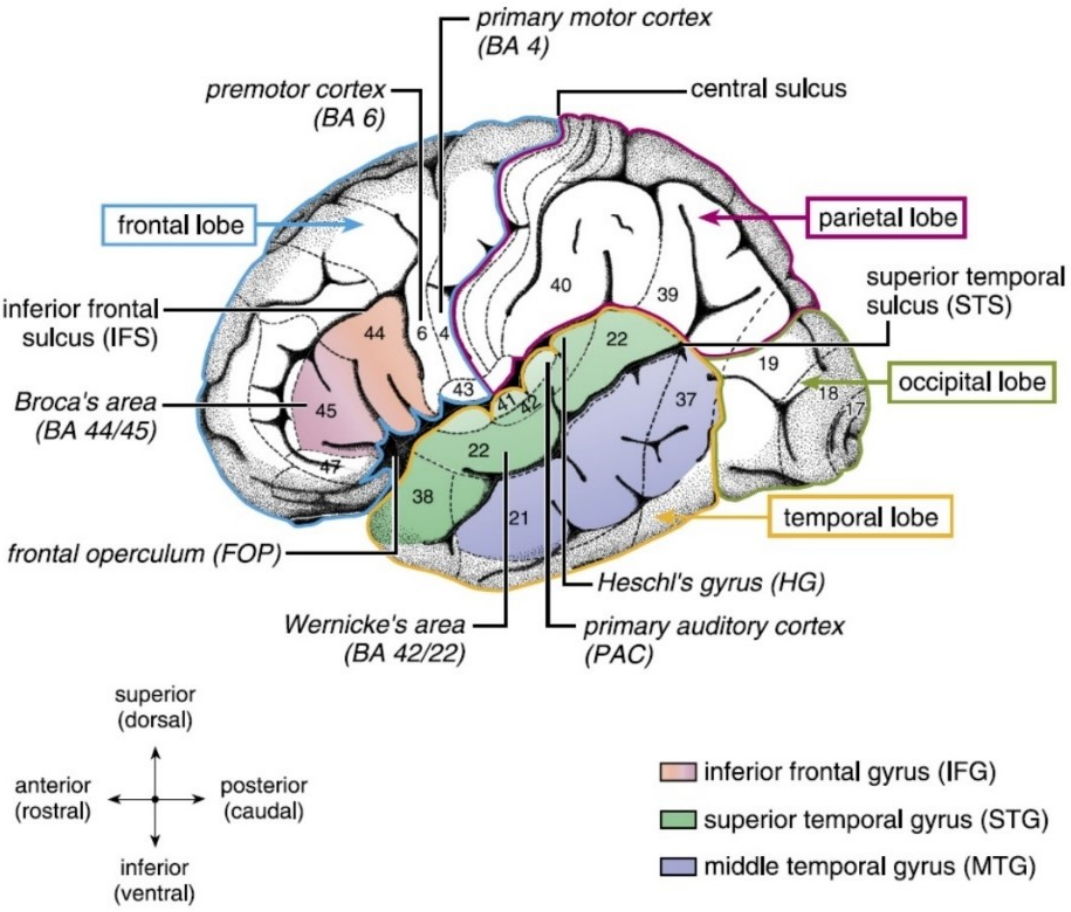

Figure 3. Schematization of areas involved in speech processing with their correspondent Broadmann's classication. The Broca's area corresponds to Broadmann's area 44 and 45. The figure is authored by Friederici (2011).

1.6 References

Abbs, J. H., Gracco, V. L., & Cole, K. J. (1984). Control of multimovement coordination: Sensorimotor mechansims in Speech motor programming. *Journal of motor behavior*, *16*(2), 195-232.

Abrams, D. A., Nicol, T., Zecker, S., & Kraus, N. (2008). Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech. *Journal of Neuroscience*, *28*(15), 3958-3965.

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

Bernstein, N. (1966). The co-ordination and regulation of movements. *The co-ordination and regulation of movements*.

Boemio, A., Fromm, S., Braun, A., & Poeppel, D. (2005). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nature neuroscience*, *8*(3), 389-395.

Bourguignon, M., De Tiege, X., De Beeck, M. O., Ligot, N., Paquier, P., Van Bogaert, P., ... & 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.

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition*, *121*(1), 41-57.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, *22*(6), 1482.

Brennan, S. E., Galati, A., & Kuhlen, A. K. (2010). Two minds, one dialog: Coordinating speaking and understanding. In *Psychology of learning and motivation* (Vol. 53, pp. 301-344). Academic Press.

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.

Bröhl, F., & Kayser, C. (2021). Delta/theta band EEG differentially tracks low and high frequency speech-derived envelopes. *Neuroimage*, *233*, 117958.

Brown-Schmidt, S., & Tanenhaus, M. K. (2008). Real-time investigation of referential domains in unscripted conversation: A targeted language game approach. *Cognitive science*, *32*(4), 643-684.

Buchsbaum, B. R., Hickok, G., & Humphries, C. (2001). Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. *Cognitive science*, *25*(5), 663-678.

Chang, E. F., Breshears, J. D., Raygor, K. P., Lau, D., Molinaro, A. M., & Berger, M. S. (2017). Stereotactic probability and variability of speech arrest and anomia sites during stimulation mapping of the language dominant hemisphere. *Journal of neurosurgery*, *126*(1), 114-121.

Crone, N. E., Boatman, D., Gordon, B., & Hao, L. (2001). Induced electrocorticographic gamma activity during auditory perception. *Clinical neurophysiology*, *112*(4), 565-582.

Crone, N., Hao, L., Hart, J., Boatman, D., Lesser, R. P., Irizarry, R., & Gordon, B. (2001). Electrocorticographic gamma activity during word production in spoken and sign language. *Neurology*, *57*(11), 2045-2053.

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

Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: functional roles and interpretations. *Frontiers in human neuroscience*, *8*, 311.

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews*, *81*, 181-187.

Doelling, K. B., Assaneo, M. F., Bevilacqua, D., Pesaran, B., & Poeppel, D. (2019). An oscillator model better predicts cortical entrainment to music. *Proceedings of the National Academy of Sciences*, *116*(20), 10113-10121.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *European journal of Neuroscience*, *15*(2), 399-402.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491-505.

Ferpozzi, V., Fornia, L., Montagna, M., Siodambro, C., Castellano, A., Borroni, P., ... & Cerri, G. (2018). Broca's area as a pre-articulatory phonetic encoder: gating the motor program. *Frontiers in human neuroscience*, *12*, 64.

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

Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of memory and language*, *49*(3), 396-413.

Friederici, A. D. (2011). The brain basis of language processing: from structure to function. *Physiological reviews*, *91*(4), 1357-1392.

Galantucci, B., Fowler, C. A., & Goldstein, L. (2009). Perceptuomotor compatibility effects in speech. *Attention, Perception, & Psychophysics*, *71*(5), 1138-1149.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic bulletin & review*, *13*(3), 361-377.

Gambi, C., & Pickering, M. J. (2013). Prediction and imitation in speech. *Frontiers in psychology*, *4*, 340.

Garrod, S., & Pickering, M. J. (2009). Joint action, interactive alignment, and dialog. *Topics in Cognitive Science*, *1*(2), 292-304.

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.

Giles, H. (1973). Communicative effectiveness as a function of accented speech.

Giles, H., Coupland, N., & Coupland, I. U. S. T. I. N. E. (1991). 1. Accommodation theory: Communication, context, and. *Contexts of accommodation: Developments in applied sociolinguistics*, *1*.

Giraud, A. L., Kleinschmidt, A., Poeppel, D., Lund, T. E., Frackowiak, R. S., & Laufs, H. (2007). Endogenous cortical rhythms determine cerebral specialization for speech perception and production. *Neuron*, *56*(6), 1127-1134.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological review*, *105*(2), 251.

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and language*, *96*(3), 280-301.

Hartsuiker, R. J., Bernolet, S., Schoonbaert, S., Speybroeck, S., & Vanderelst, D. (2008). Syntactic priming persists while the lexical boost decays: Evidence from written and spoken dialogue. *Journal of Memory and Language*, *58*(2), 214-238.

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1993). Emotional contagion. *Current directions in psychological science*, *2*(3), 96-100.

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, *38*(4), 555-568.

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature reviews neuroscience*, *13*(2), 135-145.

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in cognitive sciences*, *4*(4), 131-138.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*(1-2), 67-99.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature reviews neuroscience*, *8*(5), 393-402.

Hickok, G., Buchsbaum, B., Humphries, C., & Muftuler, T. (2003). Auditory–motor interaction revealed by fMRI: speech, music, and working memory in area Spt. *Journal of cognitive neuroscience*, *15*(5), 673-682.

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron*, *69*(3), 407-422.

Holler, J., & Wilkin, K. (2011). Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of Nonverbal Behavior*, *35*(2), 133-153.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground?. *Cognition*, *59*(1), 91-117.

Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in human neuroscience*, *5*, 82.

Jarick, M., & Jones, J. A. (2008). Observation of static gestures influences speech production. *Experimental brain research*, *189*(2), 221-228.

Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S., Hudspeth, A. J., & Mack, S. (Eds.). (2000). *Principles of neural science* (Vol. 4, pp. 1227-1246). New York: McGraw-hill.

Kappes, J., Baumgaertner, A., Peschke, C., & Ziegler, W. (2009). Unintended imitation in nonword repetition. *Brain and Language*, *111*(3), 140-151.

Kayser, S. J., Ince, R. A., Gross, J., & Kayser, C. (2015). Irregular speech rate dissociates auditory cortical entrainment, evoked responses, and frontal alpha. *Journal of Neuroscience*, *35*(44), 14691-14701.

Kelso, J. S., Tuller, B., Vatikiotis-Bateson, E., & Fowler, C. A. (1984). Functionally specific articulatory cooperation following jaw perturbations during speech: evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(6), 812.

Kerzel, D., & Bekkering, H. (2000). Motor activation from visible speech: evidence from stimulus response compatibility. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(2), 634.

Keysar, B., Barr, D. J., & Horton, W. S. (1998). The egocentric basis of language use: Insights from a processing approach. *Current directions in psychological science*, *7*(2), 46-49.

Kim, M., Horton, W. S., & Bradlow, A. R. (2011). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance.

Lane, H., & Tranel, B. (1971). The Lombard sign and the role of hearing in speech. *Journal of speech and hearing research*, *14*(4), 677-709.

Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics:(de) constructing the N400. *Nature reviews neuroscience*, *9*(12), 920-933.

Levelt, W. J. (1999). Models of word production. *Trends in cognitive sciences*, *3*(6), 223-232.

Levinson, S. C. (2013). Action formation and ascription. In T. Stivers & J. Sidnell (Eds.), The handbook of conversation analysis (pp. 103–130). Wiley-Blackwell.

Levitan, R., & Hirschberg, J. B. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions.

Liberman, A. M., & Whalen, D. H. (2000). On the relation of speech to language. *Trends in cognitive sciences*, *4*(5), 187-196.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*(3), 551-556.

Littlejohn, S. W., & Foss, K. A. (2010). *Theories of human communication*. Waveland press.

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

Mandonnet, E., Sarubbo, S., & Duffau, H. (2017). Proposal of an optimized strategy for intraoperative testing of speech and language during awake mapping. *Neurosurgical review*, *40*(1), 29-35.

Mills, G., & Healey, P. (2008, June). Semantic negotiation in dialogue: the mechanisms of alignment. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue* (pp. 46-53).

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.

Obleser, J., & Kayser, C. (2019). Neural entrainment and attentional selection in the listening brain. *Trends in cognitive sciences*, *23*(11), 913-926.

Ostry, D. J., Flanagan, J. R., Feldman, A. G., & Munhall, K. G. (1991). Human jaw motion control in mastication and speech. In *Tutorials in motor neuroscience* (pp. 535-543). Springer, Dordrecht.

Ostry, D. J., Flanagan, J. R., Feldman, A. G., & Munhall, K. G. (1991). Human jaw motion control in mastication and speech. In *Tutorials in motor neuroscience* (pp. 535-543). Springer, Dordrecht.

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

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

Payan, Y., & Perrier, P. (1997). Synthesis of VV sequences with a 2D biomechanical tongue model controlled by the Equilibrium Point Hypothesis. *Speech communication*, *22*(2-3), 185-205.

Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. Front. Psychol. 3, 320.

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169-181.

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral cortex*, *23*(6), 1378-1387.

Perrier, P., Lœvenbruck, H., & Payan, Y. (1996). Control of tongue movements in speech: The equilibrium point hypothesis perspective. *Journal of Phonetics*, *24*, 53-75.

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

Pickering, M. J., & Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation*, *4*(2), 203-228.

Pickering, M. J., & 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.

Rasenberg, M., Özyürek, A., & Dingemanse, M. (2020). Alignment in multimodal interaction: An integrative framework. *Cognitive Science*, *44*(11), e12911.

Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature neuroscience*, *12*(6), 718-724.

Reitter, D., & Moore, J. D. (2014). Alignment and task success in spoken dialogue. *Journal of Memory and Language*, *76*, 29-46.

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

Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological psychology*, *1*(4), 333-382.

Saltzman, E., Löfqvist, A., Kay, B., Kinsella-Shaw, J., & Rubin, P. (1998). Dynamics of intergestural timing: A perturbation study of lip-larynx coordination. *Experimental Brain Research*, *123*(4), 412-424.

Sanguineti, V., Laboissière, R., & Ostry, D. J. (1998). A dynamic biomechanical model for neural control of speech production. *The Journal of the Acoustical Society of America*, *103*(3), 1615-1627.

Sanguineti, V., Laboissiere, R., & Payan, Y. (1997). A control model of human tongue movements in speech. *Biological cybernetics*, *77*(1), 11-22.

Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis I* (Vol. 1). Cambridge university press.

Schegloff, E. A., & Sacks, H. (1973). Opening up closings. Semiotica, 8(4), 289–327.

Schlangen, D., & Hockey, B. A. (2008, June). Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue (pp. 46–53). In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*.

Shaiman, S., & Gracco, V. L. (2002). Task-specific sensorimotor interactions in speech production. *Experimental Brain Research*, *146*(4), 411-418.

Shockley, K., Santana, M. V., & Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(2), 326.

Tabensky, A. (2001). Gesture and speech rephrasings in conversation. *Gesture*, *1*(2), 213-235.

Tian, X., & Poeppel, D. (2012). Mental imagery of speech: linking motor and perceptual systems through internal simulation and estimation. *Frontiers in human neuroscience*, *6*, 314.

Towle, V. L., Yoon, H. A., Castelle, M., Edgar, J. C., Biassou, N. M., Frim, D. M., ... & Kohrman, M. H. (2008). ECoG gamma activity during a language task: differentiating expressive and receptive speech areas. *Brain*, *131*(8), 2013-2027.

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

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

Watkins, K., & Paus, T. (2004). Modulation of motor excitability during speech perception: the role of Broca's area. *Journal of cognitive neuroscience*, *16*(6), 978-987.

Wernicke, C., Cohen, R. S., & Wartofsky, M. W. (1874). Boston studies in the philosophy of science. *The symptom complex of aphasia: A psychological study on an anatomical basis, D. Reichel, Dordrecht*, *1969*, 34-97.

Wolpert, D. M., & Flanagan, J. R. (2001). Motor prediction. *Current biology*, *11*(18), R729-R732.

Wolpert, D. M., Doya, K., & Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *358*(1431), 593-602.

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: music and speech. *Trends in cognitive sciences*, *6*(1), 37-46.

Zheng, Z. Z., Munhall, K. G., & Johnsrude, I. S. (2010). Functional overlap between regions involved in speech perception and in monitoring one's own voice during speech production. *Journal of cognitive neuroscience*, *22*(8), 1770-1781.

2 Research Project: Speech Production

2.1 Personal contribution

The text and figures in the following paragraphs are retrieved directly from the scientific paper *"Prediction of Speech Onset by Micro-Electrocorticography of the Human Brain"* https://doi.org/10.1142/S0129065721500258 of which I am co-first author together with my colleague Emanuela Delfino, Ph.D. . The data recording in humans was conducted by Dr. Tamara Ius. I designed the machine learning model and paradigm to address the scientific question of detecting the patients' speech onset. The analyses were conceptualized by myself and Emanuela Delfino as well as the implementation of the Matlab code (in which Emanuela Delfino made a major contribution).

2.2 Abstract

Recent technological advances show the feasibility of offline decoding speech from neuronal signals, paving the way to the development of chronically implanted speech brain computer interfaces (sBCI). Two key steps that still need to be addressed for the online deployment of sBCI are, on the one hand, the definition of relevant design parameters of the recording arrays, on the other hand, the identification of robust physiological markers of the patient's intention to speak, which can be used to online trigger the decoding process. To address these issues, we acutely recorded speech-related signals from the frontal cortex of two

human patients undergoing awake neurosurgery for brain tumours using three different micro-electrocorticographic (*μ*ECoG) devices. First, we observed that, at the smallest investigated pitch (600 *μ*m), neighbouring channels are highly correlated, suggesting that more closely spaced would provide some redundant information. Second, we trained a classifier to recognize speech-related motor preparation from high-gamma oscillations (70-150 Hz), demonstrating that these neuronal signals can be used to reliably predict speech onset. Notably, our model generalized both across subjects and recording devices showing the robustness of its performance. These findings provide crucial information for the design of future online sBCI.

2.3 Introduction

Recent advances in neuroprosthetics demonstrated that intelligible speech can be offline synthesized from cortical activity (Anumanchipalli, Chartier & Chang, 2019; Angrick et al. 2019). While this represents an important stepping stone, it still leaves open crucial problems that need to be solved to develop speech brain computer interfaces (sBCI) which can be effectively implanted in patients and work continuously online. (Anumanchipalli, Chartier & Chang, 2019; Angrick et al. 2019; Ortiz-Rosario & Adeli, 2013; Rabbani, Milsap & Cron, 2019; Martin et al. 2019; Hill et al. 2012). Here, we addressed two of them.

The first problem is the need of developing devices that can chronically record brain signals in a reliable manner. Several techniques have been presented in the literature, which differ in their degree of invasiveness and spatiotemporal resolution. Starting from one of the least invasive methodologies, electroencephalography (EEG) probes electrical potential variations by using scalp electrodes. Neural oscillations are collected from large regions of the brain, making it an appropriate method for investigating communication within the brain during speech-related tasks (Zhu, Liu, Ristaniemi & F. Cong, 2020) and a powerful clinical tool to recognize, among several others, speech and auditory deficits(Mozaffari Legha & Adeli, 2019). Nevertheless, when dealing with sBCI applications, electrocorticography (ECoG) — performed by placing grids of

electrodes directly above the cortical surface — can be considered an excel- lent trade-off between several requirements. Indeed, this technique offers the advantage of recording neural activity from distributed brain areas with a spatiotemporal resolution inaccessible to non-invasive methodologies (e.g. EEG) and reduced invasiveness when compared to intracortical devices(Rabbani, Milsap & Cron, 2019; Hong & Lieber, 2019; Szostak, Grand & Constandi-nou, 2017; Chen, Canales & Anikeeva, 2017; Buzsaki, Anastassiou & Koch, 2012). In the case of chronic recordings, as is the case for BCIs, the use of ultra-flexible micro-ECoG (*μ*ECoG) arrays, rather than traditional ECoG grids, has the further advantage of lowering the foreign body reaction, and thus improving long-term performances, as largely demonstrated in animal models(Vomero et al., 2020; Bockhorst et al., 2018, Luan et al. 2017; Weltman, Yoo & Meng, 2016; Minev et al. 2015; Viventi et al. 2011; Kim et al. 2010; Biran, Martin & Tresco, 2005). Indeed, thin *μ*ECoG do conformably adhere to the brain surface with a curvature of a human brain without adding pressure to it (Vomero et al., 2020). Therefore, good signal-to-noise ratio of recorded signals occur. Not only low frequencies can be detected but even spike-like activity can be recorded with these arrays and electrode site diameters in the hundreds of micrometer range (Bockhorst et al., 2018). Nonetheless, while *μ*ECoG arrays represent a promising strategy, several of their design parameters still need to be determined, of which, a crucial one, is the pitch distance between electrodes. Indeed, signal redundancy between neighboring electrodes increases as the electrode pitch decreases(Rogers et al. 2019; Muller et al. 2016). Thus, below a given threshold distance, signals would become highly correlated and the negligible additional information that they provide would not justify their additional design and manufacturing costs. Such threshold is presently unknown, and it needs to be estimated from experimental data, also considering the purpose of the BCI (Muller et al. 2016).

The second problem we addressed here is that presently available speech-decoding devices are designed for an offline use. That is, they synthesize words and sentences from brain signals that are known to be collected during speech-related task(Anumanchipalli, Chartier & Chang, 2019; Angrick et al. 2019). On the contrary, in a prospective real-life online scenario, BCIs would be constantly exposed to the flow of neuronal activations with no additional information as to

whether these signals are related to speech production or not, similarly to inner speech settings (Sereshkeh et al., 2017; Martin et al., 2014, 2016, 2018). Under these circumstances, the device would continuously attempt to convert patterns of neuronal activity into words, with conceivably high computational cost (Anumanchipalli, Chartier & Chang, 2019; Angrick et al. 2019).

In recent years, substantial efforts went into developing new strategies to both get closer to a natural speech scenario and optimize the decoding process(Sereshkeh et al., 2017; Martin et al., 2014, 2016, 2018; Kanas et al., 2014, 2014; Dash et al., 2020; Moses et al., 2019; Guenther et al., 2009). Among the explored options, one includes the identification of speech-preparatory neural signals to reliably detect the speech onset. Previous studies reported that the most accurate speech onset/offset neuronal signals are typically found in the temporal cortex (Kanas et al. 2014, 2014) raising the issue that they might be related to the auditory feedback of the subject's own voice. Unfortunately, such signals, while indeed highly correlated with speech onset/offset (Kanas et al. 2014, 2014; Dash et al. 2020), would not be available for a real- life sBCI deployment which implies the decoding of speech from patients that can no longer produce it. Aiming to complement previous attempts in the field, one should investigate neuronal markers with two crucial characteristics: (1) ability of predicting the speech onset, and thus being able to provide sufficient time to trigger the decoding process, and (2) high correlation with speech preparation processes, and thus being available irrespective of the actual emission of speech. Similar to how a vocal cue is employed to start commonly used virtual assistants (e.g. Google Assistant, Alexa or Siri), such *"neuronal cue"* would serve the purpose of precisely identifying speech intentions and consequently trigger the initiation of the decoding process in time.

One suitable candidate region of the human cortex where to find physiological signals related to speech preparation is the speech arrest in Broca's area. Indeed, experimental evidence shows that direct electrical stimulation(Chang et al., 2017; Mandonnet et al., 2017; Tate et al., 2014) during speech production induces the so-called speech arrest phenomenon, i.e. the complete interruption of ongoing speech (Ferpozzi et al., 2018) in absence of orofacial movements and vocalizations(Gomez-Vilda et al., 2019). This reversible functional arrest identifies

Broca's area, which is known to be active prior to articulation rather than during spoken responses(Flinker et al., 2015). Specifically, results showed an increase of the high-gamma activity immediately before the speech onset or the peak of verbal response(Flinker et al., 2015; Pei et al., 2011).

In this study, we recorded neural activity from Broca's area using innovative dense *μ*ECoG grids (pitch distances = 600, 750 and 2500 *μ*m) acutely implanted in two patients performing speech production tasks. We used signals recorded from the speech arrest area to provide a quantitative estimate of the correlation between electrodes, as a function of their distance. If there would be high correlations between adjacent electrodes this would be a sign for redundancy. If not, signals are independent and only one electrode was selected. This estimate, together with our time-frequency analysis, allowed us to identify the most appropriate frequency band in terms of spatial confinement and strict anticipatory nature with the respect to the speech onset, and thus to select the most robust physiological marker of speech preparation periods.

2.4 Materials and Methods

2.4.1 Subjects

Data were collected from two patients undergoing awake neurosurgery for tumor resection (low-grade glioma). The patients gave their informed consent, and the protocol was approved by the Ethics Committee of Azienda Ospedaliera Universitaria Santa Maria della Misericordia (Udine, Italy) after verification of the Italian Ministry of Health.

2.4.2 Recordings

Device specifications and recording setup were described in previous publications(Vomero et al., 2020; Rembado et al. 2016; Castagnola et al., 2013, 2014). Briefly, three different epicortical arrays were used for the recordings (

Figure 4): the first array (hereinafter Epi) consisted of 64 channels arranged in an 8 8 square grid layout, with a pitch of 600 m between contacts and a contact diameter of 140 m; the second array (Multi Species Array; hereinafter MuSA) consisted of 16 channels arranged in a 4 4 square grid layout, with a pitch of 750 *μ*m between contacts and a contact diameter of 100 *μ*m; the third array (hereinafter EpiBig) consisted of 64 channels arranged in a rectangular grid, with a pitch of 2500 *μ*m between contacts and a contact diameter of 200 *μ*m. As required by the surgical procedures, the devices were sterilized before use. The reference electrodes on the arrays were disconnected. Recordings were performed in a single ended configuration by shorting the reference and ground contact and connecting them to the *dura mater*.

The position of the *μ*ECoG arrays on the cortex was determined based on pre-surgical analyses and intra-operative procedures. Pre-surgical analyses included a functional Magnetic Resonance Imaging (fMRI) session while performing different speech production tasks. Intraoperative procedures consisted in identifying the position of the speech arrest area by means of electrical stimulation (IES). Briefly, using a neuronavigation system (Brainlab) and an IES probe, it was possible to map eloquent areas of the brain and visualize them superimposed to the fMRI scan of the patient. This procedure is typically conducted to identify the exact position of specific regions, such as the speech arrest, and evaluate their relative distance from the tumor. We used the same approach to collect the coordinates of the speech arrest area and of the position of the array once in place, which allowed us to align the MuSA and the Epi devices.

Neural signals were collected before the surgical procedure at a sampling frequency of 3051.8 Hz, while the voices were recorded at 24 kHz. The voice the neural signals were recorded using the same data acquisition equipment; thus, they were automatically synchronized. Delays were therefore constant and identical in all trials with respect to the technical equipment. The onset for each trial was identified manually from the spectrogram of the audio signals computed with the free software Audacity.

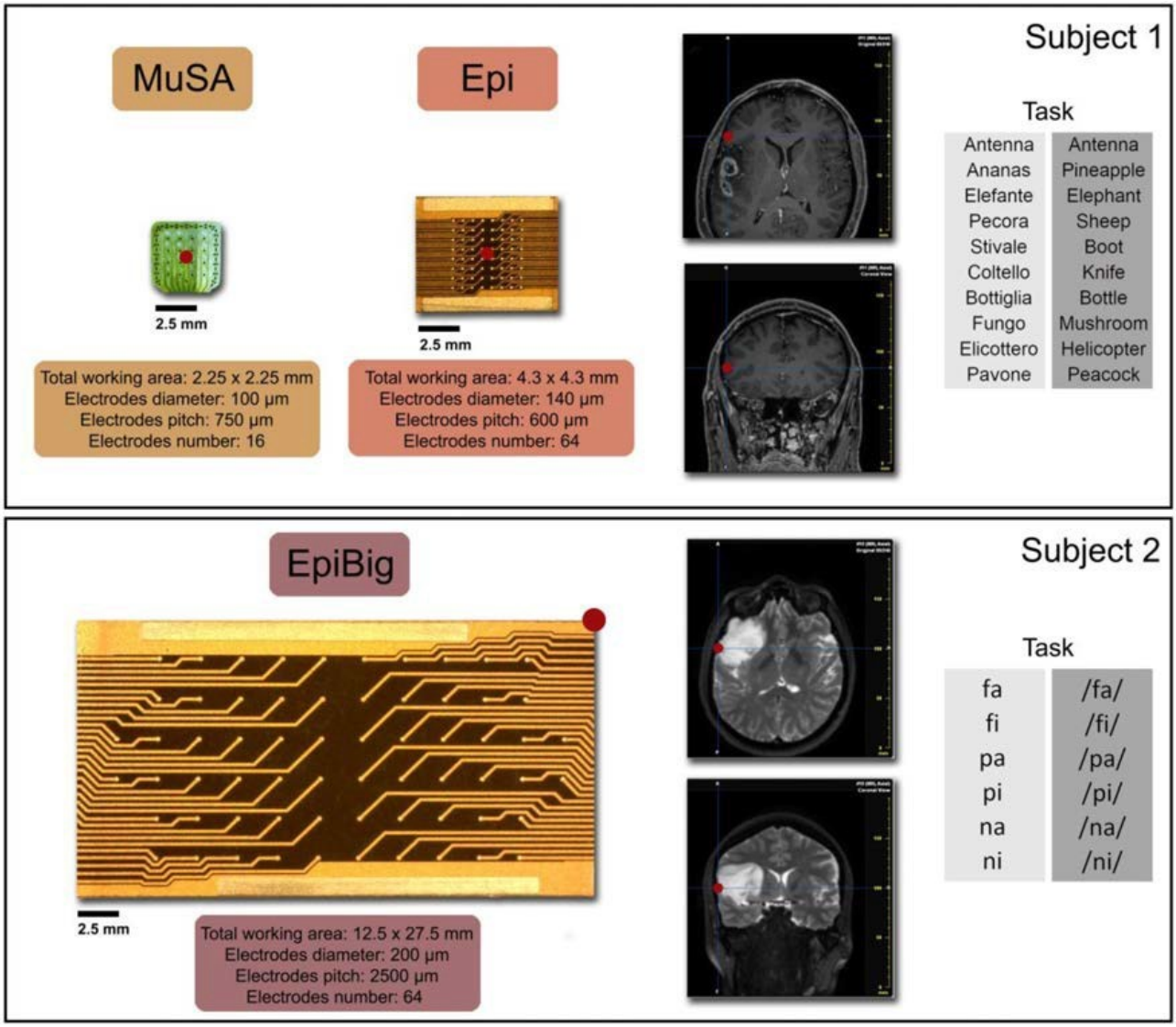

Figure 4. µECoG arrays layout and position over the cortex of subject1 (top) and subject2 (bottom). The top-left panel shows pictures of the Epi and the MuSA µECoG array. The top right panel shows and horizontal and coronal section of the patient's MRI scan. The center of each array (red dot in the left panel) was positioned over the speech arrest area (red dot in the right panel). The bottom-left panel shows a picture of the EpiBig µECoG array. The red dot localizes upper-right corner of the array superimposed to the MRI scan of the patient (horizontal plane and coronal plane). For both subjects, the speech production tasks are reported on the rightmost side of the panels.

2.4.3 Tasks

The first subject (53-year-old male, Italian native speaker, hereinafter subject1) performed two sessions of a naming task, the same conducted during the presurgical fMRI to identify eloquent areas. The task consisted in naming different images shown on a screen. Each session consisted of three blocks which 10 pictures representing Italian nouns were presented. The order of the stimuli is shown in Figure 4 and was not randomized across blocks because of limitations of the equipment available in the surgery room. During the first session, neuronal signals were recorded using the Epi array (Epi dataset, 30 trials) and during the second session data were collected using the MuSA array (MuSA dataset, 30 trials).

The second subject (41-year-old female, Italian native speaker, hereinafter referred as subject2) performed a phoneme production task (see Figure 4). The task consisted in listening to different phonemes and in repeating them. In this task, differently from the naming one, the stimuli were randomized across blocks and neuronal signals were recorded with the EpiBig device (EpiBig dataset, 84 trials).

2.4.4 Characterization of the signalredundancy across electrodes

Data were analyzed in Matlab (version 9.5, Math- works, Inc., Natick, MA) with the aim of characterizing the spatiotemporal dynamic of neural activity related to speech preparation. Ground-truth speech onset times were determined based on the sub- jects' voices recorded during the experiment. We focused our analysis on three frequency bands: beta (15-30 Hz), low-gamma (30-60 Hz), and high-gamma (70-150 Hz) (Muller et al., 2016). Signals were filtered in these three bands, by applying the Matlab function *filtfilt* to minimize phase distortion (8th order Butterworth). We also removed line noise by applying a notch filter at 50 Hz and its harmonics up to 200 Hz. Finally, the filtered data were segmented into trials, spanning from 500 ms before to 500 ms after speech onset, and analysed as follows.

2.4.4.1 Spatial Correlation Analysis

We used a correlation analysis to quantify signal redundancy across electrodes. To this end, we computed the correlation coefficient of the filtered and segmented signals for each pairwise combination of electrodes and averaged it across trials. Then, we averaged the results across electrodes sharing the same distance. The correlation decay was computed from data recorded with the Epi matrix, as this probe possesses the smallest distance between electrodes (600 μm) and thus the highest spatial resolution.

2.4.4.2 Spectrograms

Spectrograms were computed using the Matlab function spectrogram setting a temporal window of 100 ms for low and high-gamma, and 150 ms for the beta band. The overlap between windows was set to 90%. The frequency resolution was set to 1 Hz for low and high gamma bands, and to 0.5 Hz for the beta band. Power spectra were then averaged across trials.

2.4.5 Prediction of speech onset

To identify speech-preparation activities, we first segmented each recording session into *N* non- overlapping intervals, where *N* represents the number of words or phonemes (hereinafter vocalization), according to the task performed. Each interval ranged from 500 ms before an instance of speech onset to 500 ms before the subsequent one. It thus contained only one vocalization. For each interval, we extracted vectors of features from neuronal signals and labelled them either as preparation or non- preparation. Specifically, we labelled as *"preparation"* features extracted in the 500 ms preceding a speech onset event and *"non-preparation"* features extracted in all other time intervals (Figure 5). For each channel, we then trained a support vector machine (SVM) to classify feature vectors based on their assigned labels. Figure 6 reports a diagram of the prediction procedure.

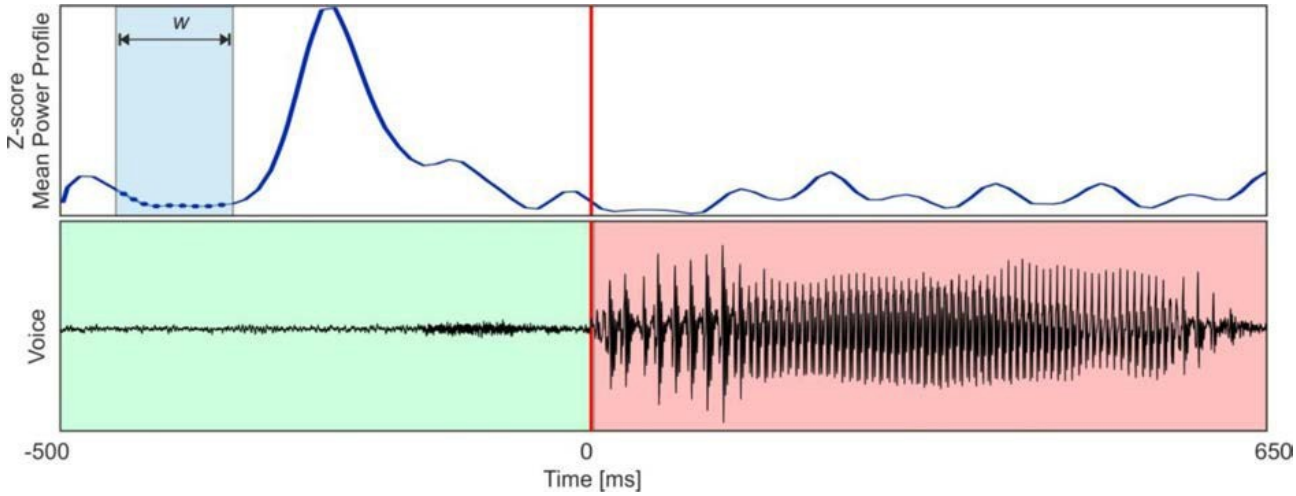

Figure 5. Graphical representation of the feature extraction and labelling procedure. Each consecutive and non-overlapping window (w) of the z-score Mean Power Profile (MPP, (blue line) was considered as an observation. Observations were labelled as preparation within 500 ms before the speech onset (vertical red line) were labelled as preparation (class 0, in green). Observations belonging to the vocalization interval and the following silence were labelled as non-preparation (class 1, in red).

2.4.5.1 Feature extraction

Features for the SVM were extracted from the high- gamma range spectrograms. To this end, we first averaged spectrograms across frequency thus obtaining a single time-varying profile of the power spectral density (hereinafter Mean Power Profile, MPP) for each channel. Since the MPP is the average of the power spectrum values for each bin computed, the time resolution of the MPP is the same of the high-gamma spectrograms. Then, we segmented the data into intervals containing only one vocalization. *Z*-score normalization was applied to compare data recorded from different devices and subjects. Each feature vector consisted of *w* consecutive samples of the *z*-score MPP, with no overlap between consecutive vectors (see Figure 5). Thus, *w* is the parameter that determines how many features are included in each observation. We tested for each dataset and channel independently different window lengths (*w*), specifically: 36, 60, 84, 108, 132, 156 milliseconds.

2.4.5.2 Classification approach

Considering each channel separately, we used a support-vector machine (SVM), which is a super- vised learning method typically used for the classification of observations that cannot be linearly separable in their space. SVMs have been widely used in biomedical research to decode speech or its related features directly from neural signals(Martin et al., 2016; Kanas et al., 2014; Moses et al., 2019). Here, we trained a set of SVM models to classify observations as preparation or not-preparation using the Matlab function fitcsvm for a two-class (binary) problem. This function supports mapping the predictor data using kernel functions.

We used a Gaussian kernel (or Radial Basis Function, RBF), already employed for speech detection from ECoG signals(Kanas et al., 2014, 2014), with a fine kernel scale. The software divides all elements of the predictor matrix by the value of Kernel Scale and applies a Box Constraint that controls the maximum penalty imposed on margin-violating observations, which helps to pre- vent overfitting (regularization). Both values were set to 1 as default value.

In our experiments, speech periods were separated by longer intervals in which the patients remained silent. Our dataset contained thus a significantly greater number of *"non-preparation"* than *"preparation"* feature vectors. To reduce the skew- ness in our data and properly train our classifiers we randomly down-sampled them in order to get balanced classes (Figure 6, Steps 2–5). Then we used a Leave-One interval-Out validation to select the optimal combination of window length *w* and most informative channel. Our classifier was trained using feature vectors belonging to all the intervals except one (Figure 6, *Step* 3) and tested using all feature vectors belonging to the left-out interval. Since the non-preparation class was randomly down-sampled for the training, this procedure was repeated ten times for each left-out interval. Validation performances were obtained by averaging across the 10 randomizations (Figure 6, *Step* 4).

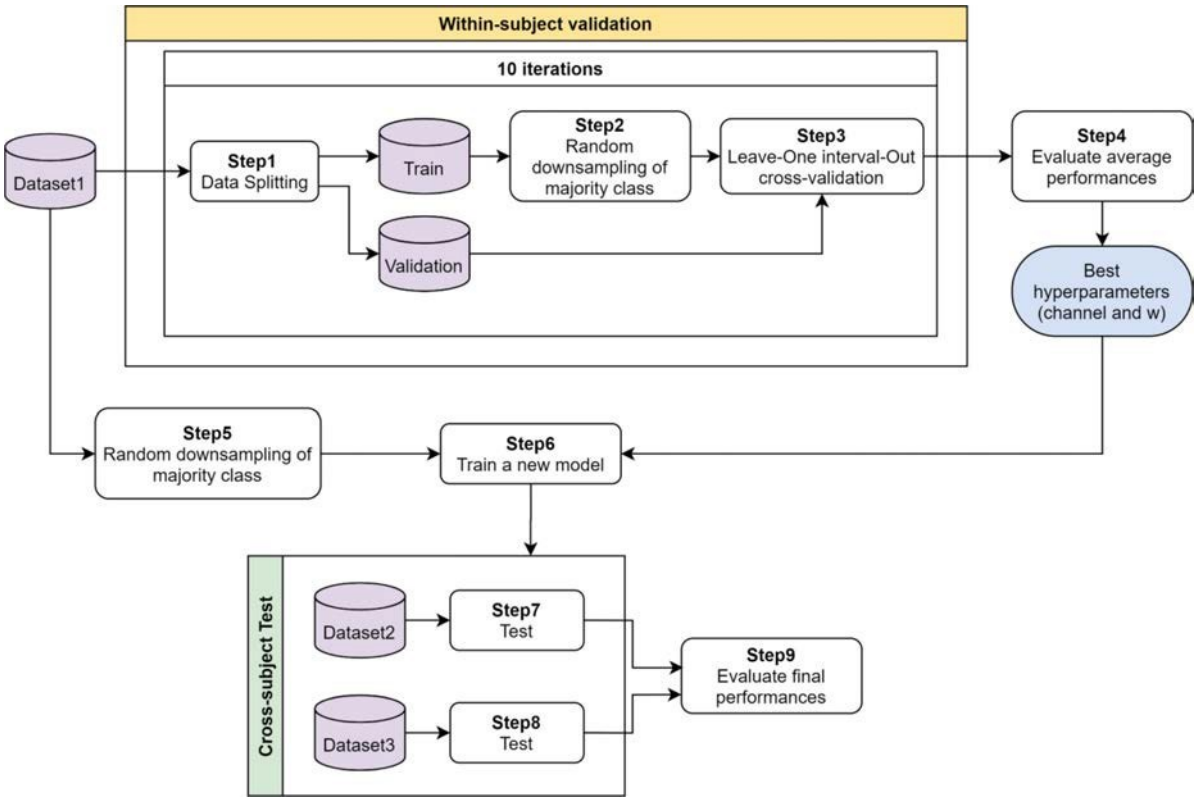

Figure 6. Training procedure of our classifier. (Top) Within-subject validation. (Step 1) Each recording session was segmented into N intervals, where N represents the number of vocalizations. Data where then split into train and validation set. (Step 2) Random down-sampling of the more represented class (i.e. "non-preparation") to train our classifier with balanced classes. (Step 3) Training of the classifier with all intervals except one. The left-out intervals was used for validation. To test the robustness against the random down-sampling, this procedure was iterated 10 times and performances were then averaged. This procedure yielded the optimal hyperparameter for our model. (Step 6) The classifier with the optimal hyperparameters ws trained using the whole daset ans (Steps 7-8) tested using a cross-subject approach.

2.4.5.4 Performance evaluation

To find the optimal value of the window length *w,* we assessed the performance of each model by means of F-score index. This index is defined as the harmonic mean of Precision and Recall and is specifically designed to deal with imbalanced datasets in which one label (i.e. non-preparation) is significantly more represented than the other (i.e. preparation) (Sung, Wong & Kamel, 2009; Saito &

Rehmsmeier, 2015). To estimate an empirical chance level for the F-score, we used a Monte Carlo approach in which we trained our classifier on a data set with shuffled feature labels. The empirical chance level was defined as the average F-score across 10 shuffling (Figure 6, *Steps* 4–9). For each dataset, we identified the best combination of window length w_o and channel number ch_o as that yielding the highest and above chance F-score.

2.4.5.5 Cross-dataset model testing

For the purpose of cross-dataset model testing, we first trained a new model on channel ch_o using window length w_o (Figure 6, *Step* 6). We then tested this model on all channels of the other datasets. This procedure was repeated for all pairwise combinations of datasets (Figure 6, *Steps* 7–8). The number of observations divided by class, for each dataset, before and after the down-sampling of the majority class (*"non- preparation"*) are reported in Table 1.

	Epi	**EpiBig**	**MuSA**
Preparation	150	320	441
Non-Preparation BDs	613	2494	1884
Non-Preparation ADs	150	320	441
Training set dimension	300	640	882

Table 1. Number of observations divided by class, for each dataset, before (BDs) and after (ADs) down-sampling

2.5 Results

2.5.1 Characterization of the signalredundancy across electrodes

As a first step, we studied the degree of signal redundancy between electrodes. Figure 7(a,c) shows the mean correlation coefficients computed from the signals recorded from the Epi array for the three frequency bands (beta: 15-30 Hz; low-gamma; 30-60 Hz; high-gamma; 70-150 Hz). We selected this probe as it has the narrowest pitch (0.6 mm). As expected, there was a clear trend in the spatial extent of the correlations. Specifically, the high- gamma band (Figure 7(c)) exhibited correlations at a narrower spatial scale than the low-gamma band (Figure 7(b)), whose spatial correlations were narrower than those in the beta band (Figure 7(a)). To quantitatively study this trend, we computed the average correlation coefficients as a function of the distance between electrodes. Results in Figure 7(d) show that (1) the correlation coefficient decreases with the increasing distance between electrodes and (2) higher frequencies consistently yield lower correlation coefficients. Notably, at the smallest considered pitch distance of 0.6 mm, signals were highly correlated, and thus redundant, in all three considered frequency bands (correlation coefficient > 0.8). This result suggests a lower bound for the pitch distance, as it shows that values smaller than 0.6 mm would provide low gain in the amount of information provided by nearby recorded signals.

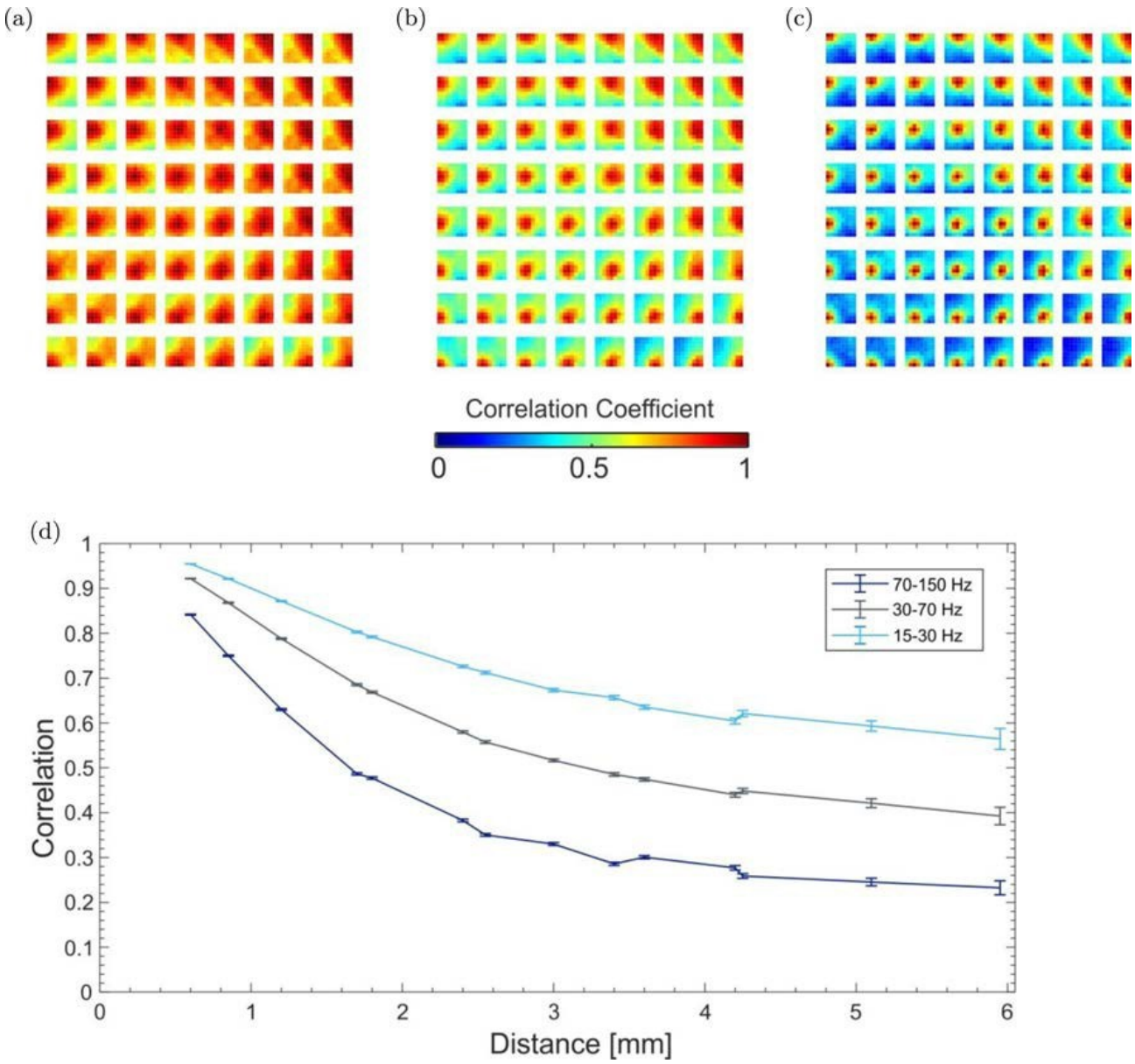

Figure 7. Characterization of the signal redundancy across electrodes performed on the Epi dataset. (a–c) Mean correlation maps of signals in the beta (15-30 Hz, panel (a)), low-gamma (30-60 Hz, panel (b)), and high-gamma (70-150 Hz, panel (c)) frequency bands obtained averaging across trials. Each square of the plot represents the correlation coefficients computed for the electrode in that position against all others. (d) Correlation profiles (mean ± SE) obtained averaging the correlation coefficients of electrodes sharing the same distance for all the tested frequency bands (light blue for beta, grey for low-gamma, and dark blue for high-gamma).

2.5.2 Prediction of speech onset

We next performed a time-frequency analysis of the recorded signals. Figure 8(a,b) shows the average across trials of the high-gamma spectrograms, aligned to the speech onset event. Results were computed from signals recorded from the Epi and MuSA probes implanted in the same patient (Subject 1). Panels A and B show a clear, time- localized increase in power few hundreds of milliseconds before the speech onset event in a subset of neighbouring electrodes (channels enclosed in the rectangular frame). Interestingly, the spatial locations of electrodes in the Epi and MuSA arrays exhibiting such anticipatory activity were overlapping (Figure 8(c)). The increase in power observed in the low-gamma and beta bands was not as equally precise in both time and space (Figure 9).

In this context, time-frequency and correlation analysis have been used to inform the feature selection process. Indeed, spectrograms were used to visualize the time alignment of the band-power increase, while the correlation maps provided a quantitative estimate of the spatial confinement of the signals in the different frequency bands. Results in Figures 7–8 show that the high-gamma modulations were the only ones temporally confined prior to the speech onset and with high spatial specificity. Consequently, we sought to investigate whether such increase in power was a reliable predictor of speech onset on a trial-by-trial basis. To this end, we trained a support vector machine (SVM) to classify a given time bin as belonging to a *"preparation"* or *"non-preparation"* interval based on the spectral features of the signals. Model's hyperparameters were set by a leave-one-out approach and the classifier's performance was assessed by means of an F-score index (see Figures 10-11, and Sec. 2.5 for further details). To provide a quantitative comparison for our model's performance, we used a shuffling procedure to computationally estimate the chance level F-score (henceforth empirical chance level). The spatial distribution of the F-scores obtained when we trained and tested our classifier on the Epi dataset (Figure 10(a)-left). Consistently with the results of Figure 8(a), channels exhibiting a clear anticipatory increase in power in the high-gamma band also yielded classification performance significantly higher than the empirical chance level. This means that spectral features in the high-gamma band can reliably predict speech onset on a trial-by-

trial basis as further demonstrated by the temporal prediction profile shown in Figure 10(b). Here, the voice is aligned to segments detected as preparation by the best-channel classifier (in light green), as well as the ground-truth (in red).

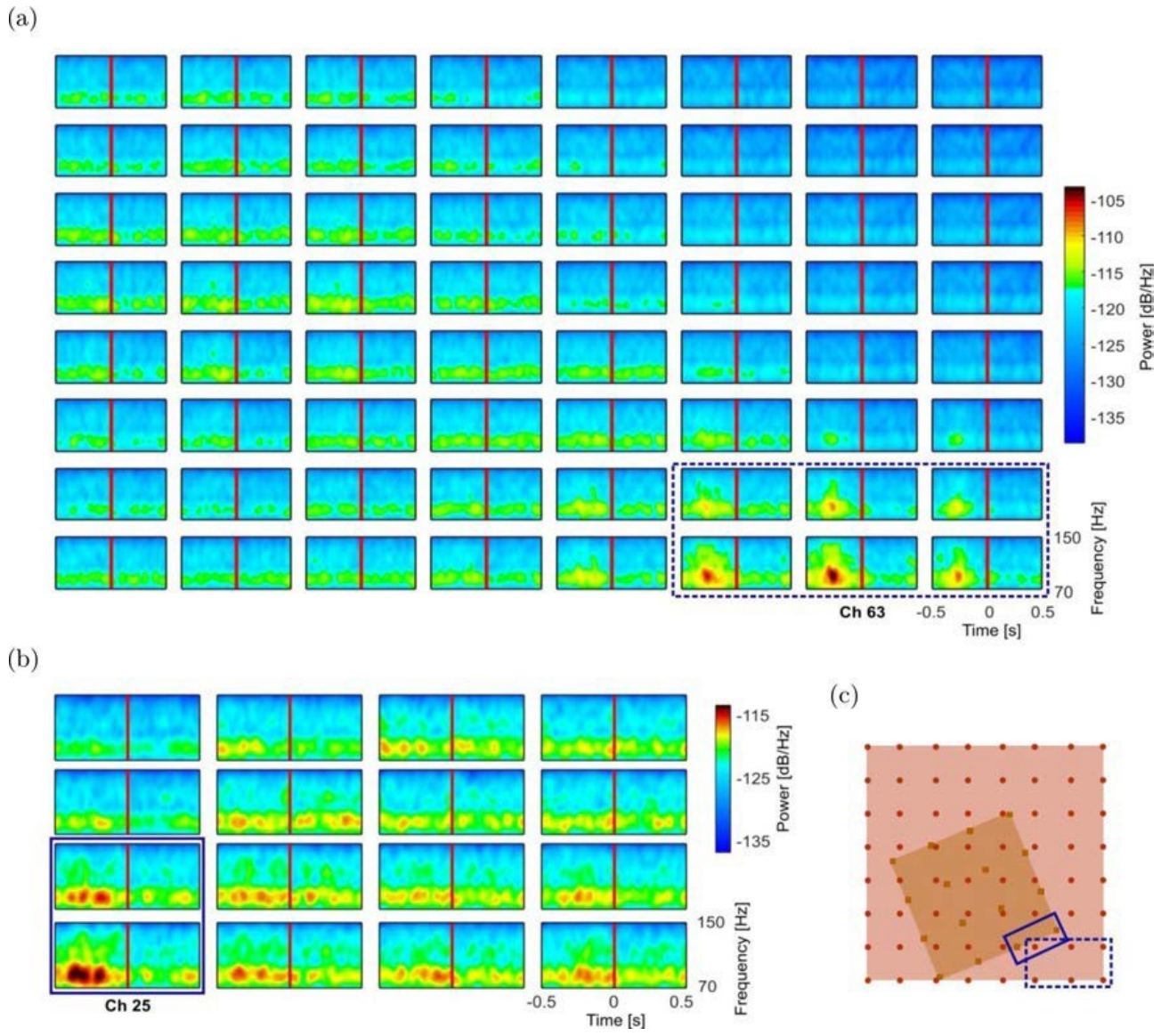

Figure 8. Mean spectrogram maps for the Epi (a) and the MuSA (b) arrays. Data are filtered in the high- gamma band (70–150 Hz) and averaged over trials. (c) Relative orientation on the cortex of the MuSA (light brown) and the Epi (red) devices. Blue rectangles refer to the electrodes highlighted on the spectrogram's plots (dashed line, Epi array; solid line, MuSA array).

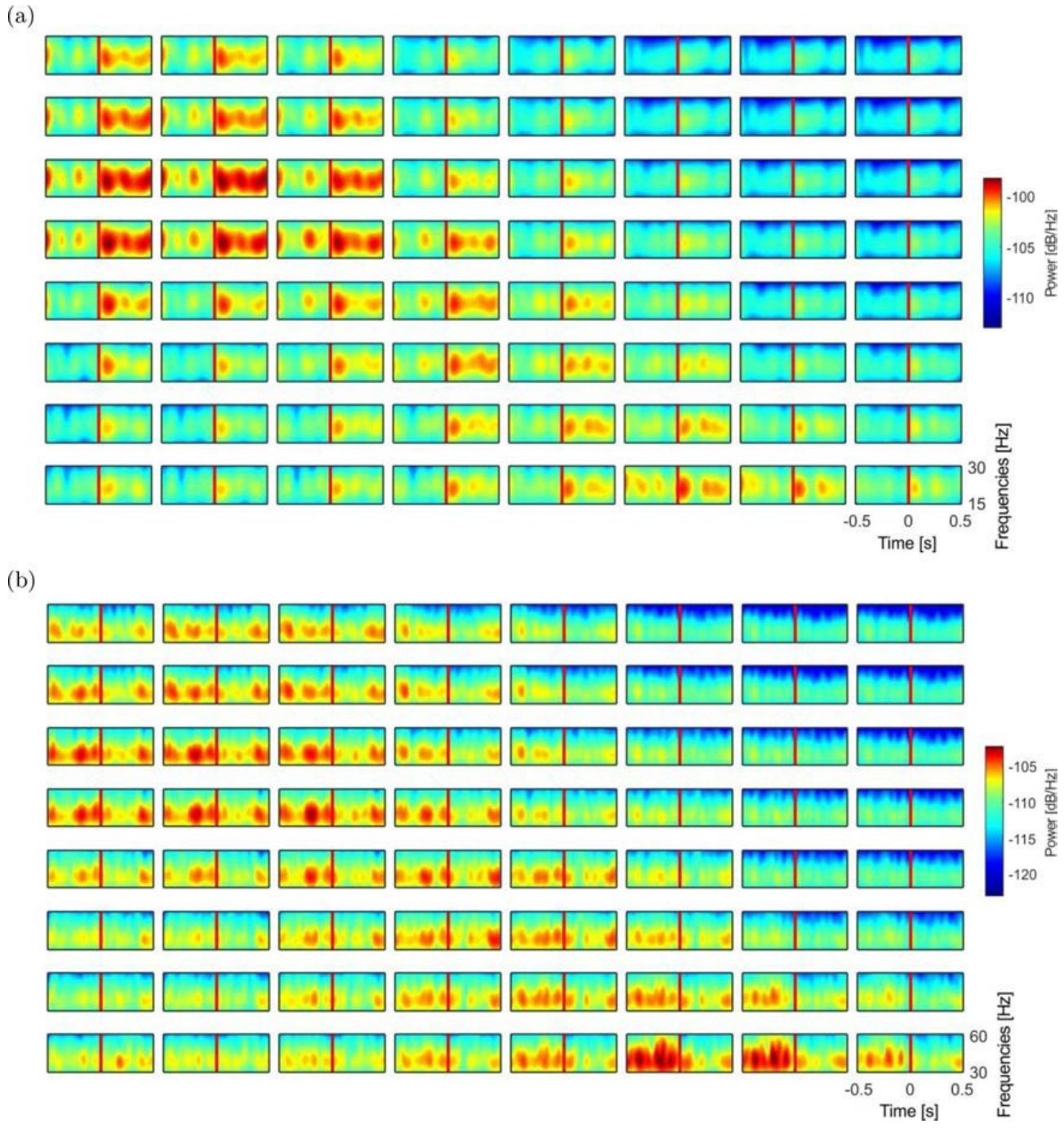

Figure 9. Mean spectrogram maps of the Epi array in the beta (15-30 Hz, panel (a)) and low-gamma (30-60 Hz, panel (b)) frequency bands. Data are averaged over trials aligned to the speech onset (vertical red line).

Neuronal responses can be recorded with a variety of probes and in different subjects. Are the identified spectral features robust with respect to these factors? We investigated this issue by means of a cross-training approach in which we trained and tested our model on all pairwise combinations of datasets, respectively. Figure 10(c) shows the results obtained when our classifier was tested on the Epi dataset and trained on the EpiBig (left panel) and Musa (right panel) datasets (see Figure 11 for all other combinations). Comparison of Figures 10(a) and 10(c) shows that, irrespective of the training dataset, our model yielded higher than the empirical chance level classification performance on electrodes at the same locations of the Epi probe (see Figures10(a,c). This result is particularly notable as the EpiBig and MuSA datasets used for training were recorded from two subjects and with two different types of probes. Taken together, the results of Figure 10 show that activity in the high-gamma band is a reliable marker of speech onset that is robust across subjects and recording devices.

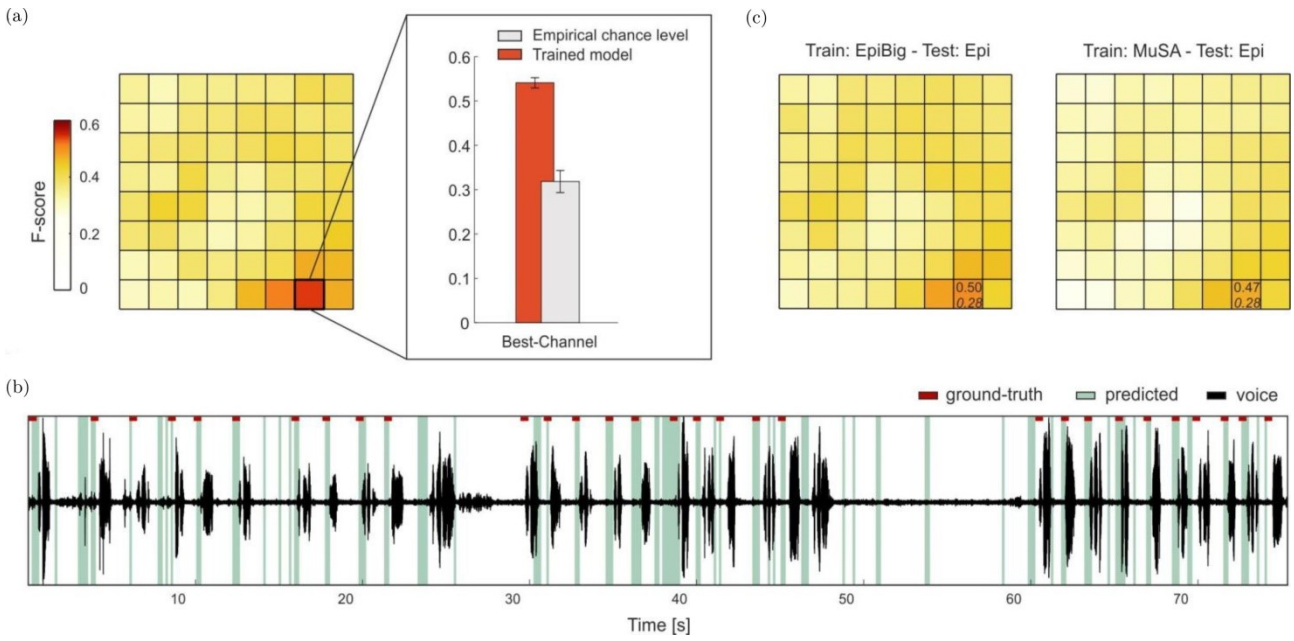

Figure 10. Prediction of speech onset. (a) On the left, the mean of 10 run F-score maps, obtained with the optimal window length tested for high-gamma MPP features of the Epi dataset (subject 1, naming task). On the right, the mean F-score of the best channel (red bar) with its standard deviation, is compared to the empirical chance level (grey bar). The non-random model resulted significantly higher (two sided t-test, P < 0.0001) than the random one. (b) Predicted (light green bars) and ground-truth (red segments) speech preparation profiles are shown aligned with the voice of the subject (black signal). The reported predicted preparation intervals belong to the best channel of the Epi dataset. (c) The mean F-score maps, for the Epi dataset (subject 1, naming task), obtained from the cross-dataset model testing; from left to right respectively the model were trained on the EpiBig (subject 2, phoneme task) and MuSA (Subject 1, naming task) datasets. For the best channel, numeric values indicating the average F-score and the corresponding empirical chance level (italic) are reported. Interestingly, the device area where the models achieved the highest performances overlapped with the one resulting from with-dataset validation, highlighting the robustness of the neural correlates decoded by the different models.

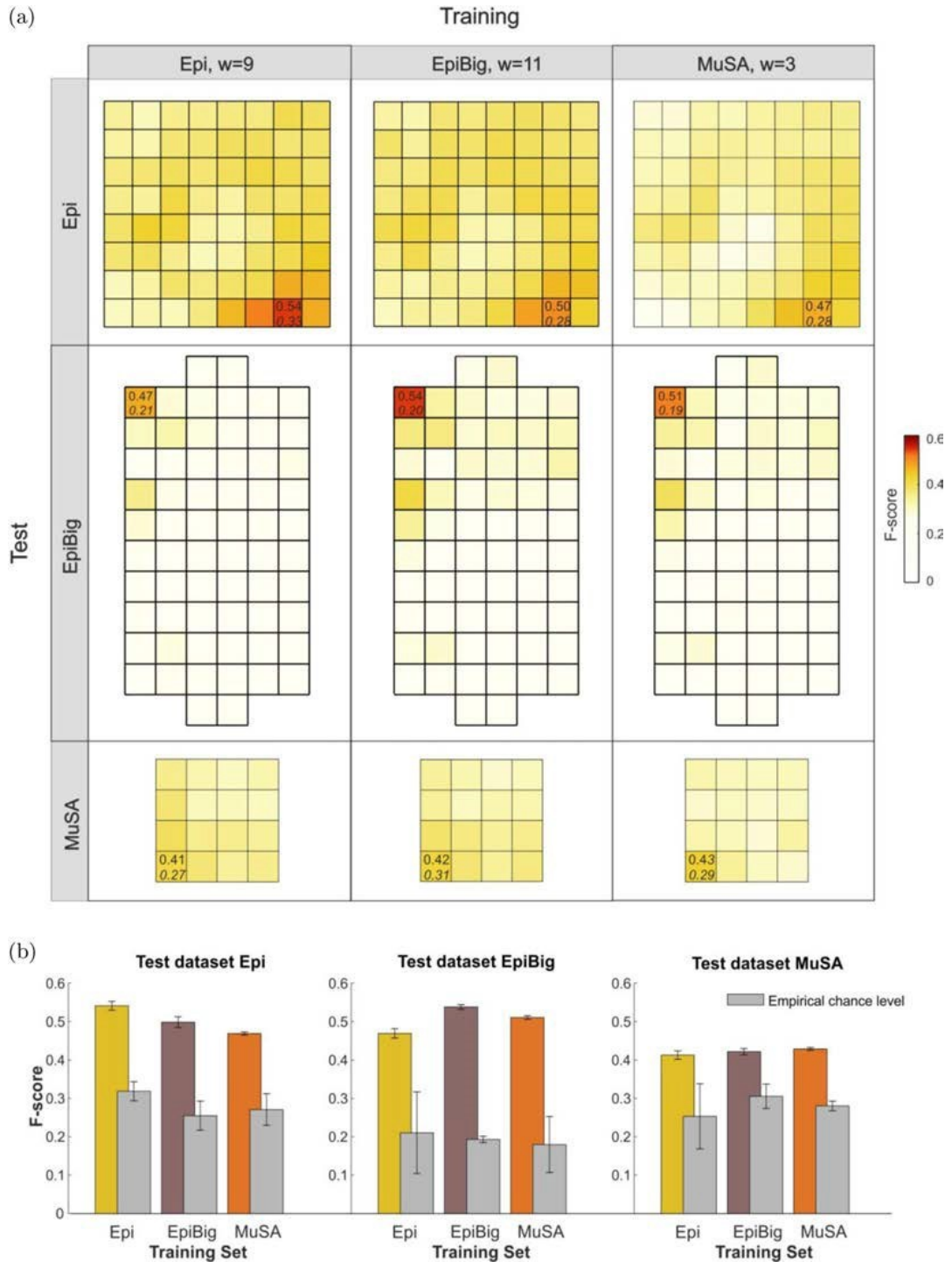

Figure 11. (a) Average F-score maps obtained during cross-dataset model testing of each pairwise combination. Each training was performed considering for each dataset the best channel (cho) and window (wo) obtained during the hyperparameters optimization. Empirical chance level of the best channel is reported in italic. (b) Average F-scores of the best channels compared to the empirical chance level (gray bars). All the within-dataset models were significantly better than the ran- domized ones (diagonal terms, two-sided t-test, P < 0.001). All cross-dataset tests show significantly higher performances than the

randomization test (off-diagonal terms, two-sided t-test, P < 0.001). Data are reported as mean ± SD.

2.6 Discussion

In this study, we used dense *μ*ECoG arrays to record the epicortical signals from awake patients undergoing brain surgery. We first investigated the spatiotemporal specificity of the neural activity during speech production. Results in Figure 7 show that, with an electrode pitch of 600 *μ*m, the correlation between neighbouring electrodes is greater than 0.8 in all investigated frequency bands. We next used a machine-learning based approach to show that high-gamma frequency signals (70-150 Hz) recorded from the speech arrest area are a reliable predictor of speech onset. These results are important in view of transitioning from offline to online speech brain computer interfaces (BCIs). A previous study reported the correlation profiles in epicortical recordings between electrode pairs with a pitch of 4 mm. The correlation trends showed increases with decreasing distance and that, at the minimum investigated electrode pitch, signals in the low-gamma and high-gamma bands are still largely uncorrelated, although differences and similarities might be affected by local anatomy, electrodes impedance, as well as the physical properties of the measured electric field (Muller et al., 2016). These results suggested that arrays with electrode pitch smaller than 4 mm were promising solutions for increasing signal resolution at high frequencies. While valuable, this study provided how- ever no lower bound for the electrode pitch.

Here, we leveraged recent advances in array design (Vomero et al., 2020; Rembado et al., 2016; Castagnola et al., 2013; Castagnola et al., 2014) to experimentally assess, for the first time, the redundancy between the activities of submillimeter spaced electrodes in the beta, low-gamma, and high-gamma frequency bands. Results in Figure 7 show that at a pitch distance of 600 *μ*m the correlation coefficient between neighbouring electrodes is greater than 0.8 in all investigated frequency bands. This result is of fundamental relevance for the design of future probes. Indeed, it suggests that this distance should be considered a lower bound for the pitch between electrodes as more densely

spaced electrodes would accrue low additional information at the cost, how- ever, of higher design, manufacturing and computational costs. In addition to being spatially confined, high-gamma neural modulations in Broca's area are known to be specifically elicited by language production(Ferpozzi et al., 2018; Pei et al., 2011).

Results in Figure 8 show that these modulations have a clear anticipatory nature, as they consistently increase few hundreds of milliseconds before speech onset. To deploy an effective *online* speech BCI, the detection of the speech onset is of crucial importance. Indeed, without such knowledge, an online speech BCI would constantly attempt to convert neuronal activations into words, even during periods of silence, with consequently higher error rates than in a controlled task that would render the BCI practically useless. A reliable detector of speech preparation would allow instead to trigger the decoding procedure in advance, and only when neuronal signals are effectively related to speech encoding, and bypassing the auditory feedback.(Kanas et al., 2014, 2014; Dash et al., 2020) Indeed, our findings provide neuronal markers that are predictive of speech onset, and highly correlated with speech preparation processes, thus present irrespective of the actual emission of speech. This predictive biomarker could play a key role in view of a real-time sBCI since it would allow to trigger the decoding process. Recent studies confirmed that speech can be offline synthesized starting from ECoG signals(Anumanchipalli, Chartier & Chang, 2019; Angrick et al. 2019) but the deployment of analogous online models in clinical application has not been achieved yet.

Here, aiming to support the transition from offline to online speech BCIs, we trained a support- vector machine classifier to recognize speech-related motor preparation on a per-channel basis. The performances obtained during validation confirmed that the high-gamma activity was indeed well-suited (Figures 10 and11). More importantly, especially for translational applications, the spatial maps of the averaged F-scores were highly consistent when the classifier was tested on data recorded from different patients with different devices, executing different experimental tasks (Figures 10 and11). Indeed, the best-performing channels obtained during cross- dataset model testing were spatially coherent with those found during the within-dataset validation. This result demonstrates that the model

was able to generalize both across different probes and patients. This is of critical relevance if we imagine that in real-life settings, patients would be using a chronically implanted BCI when they already have lost speech production abilities (i.e. no labeled training data would be available(Martin et al., 2018)).

In this study, we aimed to demonstrate the feasibility of speech onset detection in a clinical con- text. That is in a condition where few trials are typically available, preventing thus the use of complex models. Nevertheless, improvement and better tuning of the decoding algorithms are crucial points that should be continuously pursued. While significantly higher-than-chance performances have been already obtained with our per-channel paradigm, in the future it would be worth exploring the possibility of pooling single channel classifiers by means of *"mixture of experts"* approach. Indeed, although our correlation analysis indicates that most of the information is shared between neighbouring channels (correlation coefficient is higher than 0.8 for the high-gamma band), a multi-channel paradigm could significantly improve the decoding performances. Future studies will also have to include more subjects and optimize the algorithm selection, potentially exploring more powerful machine learning approaches(Ahmadlou & Adeli, 2010; Rafiei & Adeli, 2017; Pereira et al., 2020; Rokibul Alam, Siddique & Adeli, 2020; Hirschauer, Adeli & Buford, 2015) and methods which are able to better deal with imbalanced datasets(Manohar, 2021).

2.7 Conclusion

To the best of our knowledge, this study is the first one using acutely implanted *μ*ECoG grids to investigate speech onset in Broca's area.

Some methodological advancements allowed us to find two novel and, in our view, important results. First, electrodes separated by shorter distances than 600 *μ*m would likely provide, at least when data is analysed in the frequency domain, a lot of redundant information so as not to justify their design and manufacturing costs. To establish whether 600 *μ*m represents a lower bound for the electrode pitch or whether a multi-electrode approach would lead to better results, further

investigations are necessary. Second, high-gamma oscillations represent a reliable signature of speech onset that is robust across both recording devices and subjects. These results provide critical information for the design of future real-time speech BCI that are suitable for chronic long-term implant.

2.8 References

Ahmadlou, M., & Adeli, H. (2010). Enhanced probabilistic neural network with local decision circles: A robust classifier. *Integrated Computer-Aided Engineering*, *17*(3), 197-210.

Alam, K. M., Siddique, N., & Adeli, H. (2020). A dynamic ensemble learning algorithm for neural networks. *Neural Computing and Applications*, *32*(12), 8675-8690.

Angrick, M., Herff, C., Mugler, E., Tate, M. C., Slutzky, M. W., Krusienski, D. J., & Schultz, T. (2019). Speech synthesis from ECoG using densely connected 3D convolutional neural networks. *Journal of neural engineering*, *16*(3), 036019.

Anumanchipalli, G. K., Chartier, J., & Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. *Nature*, *568*(7753), 493-498.

Biran, R., Martin, D. C., & Tresco, P. A. (2005). Neuronal cell loss accompanies the brain tissue response to chronically implanted silicon microelectrode arrays. *Experimental neurology*, *195*(1), 115-126.

Bockhorst, T., Pieper, F., Engler, G., Stieglitz, T., Galindo-Leon, E., & Engel, A. K. (2018). Synchrony surfacing: Epicortical recording of correlated action potentials. *European Journal Of Neuroscience*, *48*(12), 3583-3596.

Buzsáki, G., Anastassiou, C. A., & Koch, C. (2012). The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. *Nature reviews neuroscience*, *13*(6), 407-420.

Castagnola, E., Ansaldo, A., Maggiolini, E., Angotzi, G. N., Skrap, M., Ricci, D., & Fadiga, L. (2013). Biologically compatible neural interface to safely couple nanocoated electrodes to the surface of the brain. *ACS nano*, *7*(5), 3887-3895.

Castagnola, E., Ansaldo, A., Maggiolini, E., Ius, T., Skrap, M., Ricci, D., & Fadiga, L. (2014). Smaller, softer, lower-impedance electrodes for human neuroprosthesis: a pragmatic approach. *Frontiers in neuroengineering*, *7*, 8.

Chang, E. F., Breshears, J. D., Raygor, K. P., Lau, D., Molinaro, A. M., & Berger, M. S. (2017). Stereotactic probability and variability of speech arrest and anomia sites during stimulation mapping of the language dominant hemisphere. *Journal of neurosurgery*, *126*(1), 114-121.

Chen, R., Canales, A., & Anikeeva, P. (2017). Neural recording and modulation technologies. *Nature Reviews Materials*, *2*(2), 1-16.

Dash, D., Ferrari, P., Dutta, S., & Wang, J. (2020). NeuroVAD: Real-time voice activity detection from non-invasive neuromagnetic signals. *Sensors*, *20*(8), 2248.

Ferpozzi, V., Fornia, L., Montagna, M., Siodambro, C., Castellano, A., Borroni, P., ... & Cerri, G. (2018). Broca's area as a pre-articulatory phonetic encoder: gating the motor program. *Frontiers in human neuroscience*, *12*, 64.

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

Gómez-Vilda, P., Gómez-Rodellar, A., Vicente, J. M. F., Mekyska, J., Palacios-Alonso, D., Rodellar-Biarge, V., ... & Rektorova, I. (2019). Neuromechanical modelling of articulatory movements from surface electromyography and speech formants. *International journal of neural systems*, *29*(02), 1850039.

Guenther, F. H., Brumberg, J. S., Wright, E. J., Nieto-Castanon, A., Tourville, J. A., Panko, M., ... & Kennedy, P. R. (2009). A wireless brain-machine interface for real-time speech synthesis. *PloS one*, *4*(12), e8218.

Hill, N. J., Gupta, D., Brunner, P., Gunduz, A., Adamo, M. A., Ritaccio, A., & Schalk, G. (2012). Recording human electrocorticographic (ECoG) signals for

neuroscientific research and real-time functional cortical mapping. *JoVE (Journal of Visualized Experiments)*, (64), e3993.

Hirschauer, T. J., Adeli, H., & Buford, J. A. (2015). Computer-aided diagnosis of Parkinson's disease using enhanced probabilistic neural network. *Journal of medical systems*, *39*(11), 1-12.

Hong, G., & Lieber, C. M. (2019). Novel electrode technologies for neural recordings. *Nature Reviews Neuroscience*, *20*(6), 330-345.

Kanas, V. G., Mporas, I., Benz, H. L., Sgarbas, K. N., Bezerianos, A., & Crone, N. E. (2014). Joint spatial-spectral feature space clustering for speech activity detection from ECoG signals. *IEEE Transactions on Biomedical Engineering*, *61*(4), 1241-1250.

Kanas, V. G., Mporas, I., Benz, H. L., Sgarbas, K. N., Bezerianos, A., & Crone, N. E. (2014, August). Real-time voice activity detection for ECoG-based speech brain machine interfaces. In *2014 19th International Conference on Digital Signal Processing* (pp. 862-865). IEEE.

Kim, D. H., Viventi, J., Amsden, J. J., Xiao, J., Vigeland, L., Kim, Y. S., ... & Rogers, J. A. (2010). Dissolvable films of silk fibroin for ultrathin conformal bio-integrated electronics. *Nature materials*, *9*(6), 511-517.

Luan, L., Wei, X., Zhao, Z., Siegel, J. J., Potnis, O., Tuppen, C. A., ... & Xie, C. (2017). Ultraflexible nanoelectronic probes form reliable, glial scar–free neural integration. *Science advances*, *3*(2), e1601966.

Mandonnet, E., Sarubbo, S., & Duffau, H. (2017). Proposal of an optimized strategy for intraoperative testing of speech and language during awake mapping. *Neurosurgical review*, *40*(1), 29-35.

Manohar(2021). SMOTE Synthetic Minority Over-Sampling Technique) (https://www.mathworks.com/matlabcentral/fileexchange/38830-smote-synthetic-minority-over-sampling-technique), MATLAB Central File Exchange. Retrieved on February 22, 2021.

Martin, S., Brunner, P., Holdgraf, C., Heinze, H. J., Crone, N. E., Rieger, J., ... & Pasley, B. N. (2014). Decoding spectrotemporal features of overt and covert speech from the human cortex. *Frontiers in neuroengineering*, *7*, 14.

Martin, S., Brunner, P., Iturrate, I., Millán, J. D. R., Schalk, G., Knight, R. T., & Pasley, B. N. (2016). Word pair classification during imagined speech using direct brain recordings. *Scientific reports*, *6*(1), 1-12.

Martin, S., Iturrate, I., Millán, J. D. R., Knight, R. T., & Pasley, B. N. (2018). Decoding inner speech using electrocorticography: Progress and challenges toward a speech prosthesis. *Frontiers in neuroscience*, *12*, 422.

Martin, S., Millán, J. D. R., Knight, R. T., & Pasley, B. N. (2019). The use of intracranial recordings to decode human language: Challenges and opportunities. *Brain and language*, *193*, 73-83.

Minev, I. R., Musienko, P., Hirsch, A., Barraud, Q., Wenger, N., Moraud, E. M., ... & Lacour, S. P. (2015). Electronic dura mater for long-term multimodal neural interfaces. *Science*, *347*(6218), 159-163.

Moses, D. A., Leonard, M. K., Makin, J. G., & Chang, E. F. (2019). Real-time decoding of question-and-answer speech dialogue using human cortical activity. *Nature communications*, *10*(1), 1-14.

Mozaffarilegha, M., & Adeli, H. (2019). Visibility graph analysis of speech evoked auditory brainstem response in persistent developmental stuttering. *Neuroscience Letters*, *696*, 28-32.

Muller, L., Hamilton, L. S., Edwards, E., Bouchard, K. E., & Chang, E. F. (2016). Spatial resolution dependence on spectral frequency in human speech cortex electrocorticography. *Journal of neural engineering*, *13*(5), 056013.

Ortiz-Rosario, A., & Adeli, H. (2013). Brain-computer interface technologies: from signal to action. *Reviews in the Neurosciences*, *24*(5), 537-552.

Pei, X., Leuthardt, E. C., Gaona, C. M., Brunner, P., Wolpaw, J. R., & Schalk, G. (2011). Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition. *Neuroimage*, *54*(4), 2960-2972.

Pereira, D. R., Piteri, M. A., Souza, A. N., Papa, J. P., & Adeli, H. (2020). FEMa: A finite element machine for fast learning. *Neural Computing and Applications*, *32*(10), 6393-6404.

Rabbani, Q., Milsap, G., & Crone, N. E. (2019). The potential for a speech brain–computer interface using chronic electrocorticography. *Neurotherapeutics*, *16*(1), 144-165.

Rafiei, M. H., & Adeli, H. (2017). A new neural dynamic classification algorithm. *IEEE transactions on neural networks and learning systems*, *28*(12), 3074-3083.

Rembado, I., Castagnola, E., Turella, L., Ius, T., Budai, R., Ansaldo, A., ... & Fadiga, L. (2017). Independent component decomposition of human somatosensory evoked potentials recorded by micro-electrocorticography. *International journal of neural systems*, *27*(04), 1650052.

Rogers, N., Hermiz, J., Ganji, M., Kaestner, E., Kılıç, K., Hossain, L., ... & Gilja, V. (2019). Correlation structure in micro-ECoG recordings is described by spatially coherent components. *PLoS computational biology*, *15*(2), e1006769.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, *10*(3), e0118432.

Sereshkeh, A. R., Trott, R., Bricout, A., & Chau, T. (2017). Online EEG classification of covert speech for brain–computer interfacing. *International journal of neural systems*, *27*(08), 1750033.

Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, *23*(04), 687-719.

Szostak, K. M., Grand, L., & Constandinou, T. G. (2017). Neural interfaces for intracortical recording: Requirements, fabrication methods, and characteristics. *Frontiers in Neuroscience*, *11*, 665.

Tate, M. C., Herbet, G., Moritz-Gasser, S., Tate, J. E., & Duffau, H. (2014). Probabilistic map of critical functional regions of the human cerebral cortex: Broca's area revisited. *Brain*, *137*(10), 2773-2782.

Viventi, J., Kim, D. H., Vigeland, L., Frechette, E. S., Blanco, J. A., Kim, Y. S., ... & Litt, B. (2011). Flexible, foldable, actively multiplexed, high-density electrode array for mapping brain activity in vivo. *Nature neuroscience*, *14*(12), 1599-1605.

Vomero, M., Cruz, M. F. P., Zucchini, E., Ciarpella, F., Delfino, E., Carli, S., ... & Stieglitz, T. (2020). Conformable polyimide-based μECoGs: Bringing the electrodes closer to the signal source. *Biomaterials*, *255*, 120178.

Wang, W., Degenhart, A. D., Sudre, G. P., Pomerleau, D. A., & Tyler-Kabara, E. C. (2011, August). Decoding semantic information from human electrocorticographic (ECoG) signals. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 6294-6298). IEEE.

Wang, Z., Gunduz, A., Brunner, P., Ritaccio, A. L., Ji, Q., & Schalk, G. (2012). Decoding onset and direction of movements using electrocorticographic (ECoG) signals in humans. *Frontiers in neuroengineering*, *5*, 15.

Weltman, A., Yoo, J., & Meng, E. (2016). Flexible, penetrating brain probes enabled by advances in polymer microfabrication. *Micromachines*, *7*(10), 180.

Zhu, Y., Liu, J., Ristaniemi, T., & Cong, F. (2020). Distinct patterns of functional connectivity during the comprehension of natural, narrative speech. *International journal of neural systems*, *30*(03), 2050007.

3 Research Project: Speech Perception

3.1 Personal contribution

The text and figures in the following paragraphs represent a draft of a scientific article that will be submitted shortly, of which I will be the only first author. The data recording was conducted by myself together with Dr. Elisa Dolfini and Alice Tomassini Ph.D. I implemented the experiment, performed all the analyses and wrote the correspondent Matlab and Python software. My colleague Alice Tomassini Ph.D. helped and guided me through the analyses with her precious and deep knowledge about EEG analysis.

3.2 Abstract

Speech processing entails a complex interplay between bottom-up entrainment to the quasi-rhythmic properties of speech acoustics and top-down modulations guiding attention in time and aiding selection of the most relevant input subspaces. Top-down signals are believed to originate primarily from motor regions, yet similar activities have been shown to tune attentional cycles also for simpler, non-speech stimuli. Here we examined whether neural signals encode detailed articulatory information, pointing to the involvement of a domain-specific mechanism during speech listening. We measured electroencephalographic (EEG) data while participants listened to sentences for which articulatory kinematics of the lips, jaws and tongue were also available (via Electro-Magnetic Articulography, EMA). We captured the patterns of articulatory coordination through Principal Component Analysis (PCA) and used Partial Information Decomposition (PID) to identify

whether the speech envelope and each of the kinematic components provided unique, synergistic and/or redundant information regarding the EEG signals. Interestingly, tongue movements contain both unique as well as synergistic information with the envelope, that are encoded in brain signals. This demonstrates that during speech listening the brain retrieves highly specific and uniquely motor information that is never accessible through vision, thus leveraging on audio-motor maps arising from the acquisition of speech production during development.

3.3 Introduction

Verbal interaction is an essential part of human behavior and our brains are tuned to decode speech. Neural oscillations in the delta and theta range play a key role in shaping speech perception (Giraud & Poeppel, 2012; Meyer, 2018). Indeed, coupling of brain oscillatory activity to the quasi-rhythmic properties of speech, or speech neural entrainment (Obleser & Kayser, 2019), positively scales with speech intelligibility (Ghitza, 2012; Peelle et. al., 2013; Ding & Simon 2014; Kayser et al., 2015; Riecke et al., 2018) and is tightly related to speech comprehension performance (Ahissar et al., 2001; Luo & Poeppel, 2007; Peelle et. al., 2013; Gross et al., 2013; Ding & Simon, 2014). Importantly, brain stimulation producing entrainment of oscillatory activity causally modulates speech comprehension performance (Zoefel et al., 2018; Riecke et al., 2018; Kösem et al., 2020).

Brain entrainment to speech in the delta and theta band most likely increases comprehension via a facilitation of task-relevant information (Obleser & Kayser, 2019). The cocktail party effect (Cherry, 1953; Ding & Simon, 2012) is a perfect example in this respect, showing that selective attention translates into increased neural entrainment to the attended acoustic stream (Golumbic et al., 2012; Kerlin et al., 2010; Golumbic et al., 2013a; O'Sullivan et al., 2015; Vander Ghinst et al., 2016). Furthermore, if other speech-related cues are available, neural activity can also entrain to these signals. For instance, when acoustic intelligibility is compromised, oscillatory occipital activity couples to the periodicity of lips or face movements (Giordano et al., 2017; O'Sullivan et al., 2021; Park et al., 2016; Peelle & Sommers, 2015; Giordano et al., 2017). Unsurprisingly, speech comprehension

mostly benefits from visual cues in suboptimal listening conditions (Sumby & Pollack, 1954; Schroeder et al., 2008; Golumbic et al., 2013b).

Neural entrainment to speech thus reflects top-down influences (Kösem et al., 2018; Di Liberto et al., 2018; Cope et al., 2017) which are driven by prior knowledge and/or context to predict the temporal structure of the stimuli (Calderone et al., 2014; Poeppel, 2003; Keitel, Gross & Kayser, 2018; Poeppel & Assaneo, 2020). One source of top-down modulations originates from the frontal lobe, specifically oscillatory activity in the left inferior frontal cortex (between 1 and 3 Hz) and motor cortex (between 4 and 8 Hz) modulates the phase of low-frequency activity in auditory areas (Park et al., 2015). This modulation may reflect a domain-general mechanism extending beyond speech processing with the motor system orchestrating sensory processing in time (Morillon & Baillet, 2017). Whether the motor system provides domain-general temporal predictions or richer domain-specific information about articulatory features, is however still unclear. Indeed, top-down motor influences may exploit action circuits to implement an internal 'simulation' of movements (Morillon et al., 2019; Arnal & Giraud 2012; Schubotz 2007).

To investigate this fundamental question, we designed an EEG experiment where participants listened to auditorily presented sentences. The sentences were obtained from a publicly available dataset (Canevari, Badino, & Fadiga, 2015) in which acoustic data is synchronized with articulatory data recorded via electromagnetic articulography (EMA). EMA uses miniaturized sensor coils placed on articulators (lips, jaws, tongue) to measure accurate position data with a high sampling frequency during speech production. Of key relevance to the current research question is that the EMA provides the accurate description of speech articulators that is essential to uncover whether motor information contribute to the representation of speech in the listener's brain. To this end, we used the Partial Information Decomposition (PID) method (Williams & Beer 2010; Ince, 2017), that is designed to separate unique, redundant (shared), or synergistic (complementary) information provided by two source signals (here speech envelope and kinematic data) about a third target signal (here brain activity). We thus tested whether articulatory kinematics is encoded during listening and conveys information about speech that cannot be obtained from the speech

envelope alone, i.e. unique neural information about kinematics or synergistic neural representation of speech envelope and kinematics (a better prediction of the neural response from both modalities simultaneously). Our hypothesis was that, if speech neural entrainment entails also a domain-specific motor process, entrainment to speech kinematics will be observed.

3.4 Methods

3.4.1 Partecipants

A total of 23 healthy naive volunteers were recruited for this study and were paid 30€ for their participation. All participants were native speakers of Italian, right-handed (by self-report) and had a normal or corrected-to-normal vision. One participant was excluded because of technical problems during data acquisition. Analysis was performed on data from the remaining 22 participants (9 males and 13 females). Participants were informed about the experimental procedure and gave their written consent before participation. The experiment was approved by the local ethical committee *"Comitato Etico Unico della Provincia di Ferrara"* (approval N. 170592).

3.4.2 Stimuli

The stimuli were selected from the Multi-SPeaKing-style Articulatory corpus (MSPKA; Canevari, Badino, & Fadiga, 2015) which comprises simultaneous recordings of audio and articulatory (lips, jaws and tongue) data of three mother-tongue speakers pronouncing sentences in Italian. Audio was recorded at a sampling rate of 22.05 kHz. Articulators were tracked at a sampling frequency of 400 Hz by means of an electromagnetic articulography system (EMA; NDI Wave, Northern Digital Instruments, Canada; Berry, 2011). In the present study, we used data corresponding to x, y, and z positions of 7 sensor coils glued on the upper lip (UL), lower lip (LL), upper jaw (UJ), lower jaw (LJ), tongue tip (TT), tongue middle

(TM) and tongue back (TB) (see Figure 12(a) for a schematic illustration). The EMA data provides a very accurate characterization of mouth kinematics and it is commonly used in speech technology research (Savariaux et al., 2017).

For this study, we used 50 sentences (duration ranging from 6.2 to 9.4 s) pronounced by the same female speaker (referred to as *"lls"* within the dataset). The acoustic stimuli were manually checked and processed to remove any silent and/or noisy part at the beginning and end of the sentences. All acoustic stimuli were then normalized to the same average intensity (71 dB). Data corresponding to one sentence (out of 50) were discarded from the analysis because the corresponding EMA data turned out to be corrupted. During the experiment, participants were provided only with the acoustic stimuli. The corresponding EMA data were only used for data analysis (see below).

3.4.3 Experimental setup and procedure

Participants sat at a ~80-cm distance in front of an LCD monitor (VIEWPixx/EEG; 24", 120 Hz) with their right hand resting on a button-box (Cedrus RB-840 response Box). On each trial, participants were presented with a black fixation cross at the center of a uniformly gray screen; after a variable time (ranging between 0.1 and 1.1 s), a randomly selected sentence was presented acoustically via two loudspeakers placed at ~20 cm from both sides of the screen. The fixation cross was removed after a variable time (between 0.1 and 1.1 s) from the end of the acoustic stimulus, and one word appeared at the center of the screen. The presented word rhymed 50% of the times with one of the words contained in the previously heard sentence (excluding the first and last words in the sentence and all monosyllabic words). Participants had to indicate whether the word rhymed or not by pressing one of two buttons, located few centimeters apart, using always the same finger (the right index). The rhyming task was included to encourage participants to listen attentively to the whole sentences. To avoid possible biases in the participants' responses, we ensured that rhyming and non-rhyming words were matched for number of syllables and their frequency of use in the Italian language by means of an online software tool (http://linguistica.sns.it/esploracolfis/home.htm). Different words were presented for

each repetition of the same sentence (amounting to 8 words for each sentence, 200 words in total).

Every trial ended when participants provided their response in the rhyming task; trials were automatically ended if no response was provided within 10 s. Participants were asked to reduce blinks as much as possible and maintain their eyes on the fixation cross for the whole duration of the sentence.

The experiment consisted of four separate blocks of 50 trials each (200 trials in total), with short in-between breaks. The whole experiment lasted about 2hrs, including the EEG cap mounting and preparation. Stimulus presentation and button-press acquisition were controlled via Matlab (The Math Works, Inc.; https://www.mathworks.com; RRID:SCR_001622) and the PsychToolbox-3 extensions (http://psychtoolbox.org; RRID:SCR_002881). All relevant events in the trial (e.g., trial start, stimulus onset, button press) were converted in a TTL by the VIEWPixx/EEG system to accurately synchronize them with the EEG data.

3.4.4 EEG recording and analyses

EEG data were recorded continuously during the experiment with a 64-channel active electrode system (BrainAmp MR Plus, Brain Products GmbH, Gilching, Germany). Electrooculograms (EOGs) were recorded using 4 electrodes from the cap (FT9, FT10, PO9, and PO10) that were removed from their original scalp sites and placed at the bilateral outer canthi and below and above the right eye to record horizontal and vertical eye movements, respectively. All electrodes were online referenced to the left mastoid. The impedance of the electrodes was kept below 10 kΩ. EEG signals were acquired at 1000 Hz.

Analyses were performed within the Matlab and Python computing environments, using open-source toolboxes and libraries such as Fieldtrip (http://www.fieldtriptoolbox.org; RRID:SCR_004849) (Oostenveld et al., 2011), MNE (Gramfort et al., 2013) and PID library (https://github.com/robince/partial-info-decomp) as well as custom-made code. Analyses were performed only on trials in which participants gave correct responses in the rhyming task (76.3±7.5%; MEAN±SD).

3.4.5 Speech Envelope Extraction

The amplitude envelope of the acoustic speech signals was calculated by adapting a previously described method (Smith et. al., 2002, Park et al., 2018). As in the Chimera toolbox (Smith et al., 2002), we defined 6 frequency bands in the range 80-8820 Hz that are equally spaced on the cochlear map. The speech signal was first filtered within those six frequency bands (MNE filter_data function, two-pass Butterworth filter, 4th order). Then, we computed the absolute value of the Hilbert transform for each bandpass-filtered signal. Finally, the speech envelope was obtained by summing up the result across all the frequency bands. The envelope was down sampled to 400 Hz to match the sampling frequency of the EMA data.

3.4.6 Kinematic features extraction

To capture meaningful speech coordination patterns in the high-dimensional EMA data (i.e., 7 sensors X 3 dimensions = 21 time series of position data) we used a dimensionality reduction technique. We applied Principal Component Analysis (PCA) as implemented in the Fieldtrip Toolbox (function: ft_componentanalysis; method: pca). PCA outputs feature activations over time (principal components [PCs], see Figure 13(a)) that explain part of the variance in the EMA measurements and are orthogonal to each other. Furthermore, PCA provides information about the relative contribution of each kinematic feature (PC weights, see Figure 12(c)) to the reconstruction of the EMA recordings in single trials. By visually inspecting (the absolute values of) the PC weights it is thus possible to assess the physiological validity of the articulatory coordination pattern identified by each PC.

3.4.7 EEG pre-processing

The continuous EEG data were band-pass filtered between 0.5 and 100 Hz (two-pass Butterworth filter, 4th order), and down-sampled to 400 Hz to match the sampling frequency of the EMA data. Data were then re-referenced to the common

average and time-aligned to the acoustic stimulus onset (from -1 s to the duration of the longest sentence plus one second). Data were visually inspected, and noisy trials were removed. Independent component analysis (ICA) was then applied to identify and remove artifacts related to eye movements and heartbeat. Noisy channels (T8 for one subject) were excluded from the ICA analysis and substituted by linear interpolation of neighboring channels (after ICA-based artifact rejection). The total amount of trials retained for further analysis was 142 ± 21.7 (MEAN±SD).

3.4.8 Neural coupling to speech envelope and kinematic features

To quantify neural coupling to speech production, we used mutual information (MI), a measure of statistical dependence that captures any type of relationship (even non-linear and non-monotonic) between two signals (Shannon, 1948). Our aim here was to uncover the neural representations of the different kinematic components in the brain of the listener and quantify the contribution of these representations to the neural encoding of speech. To this end, we computed the MI between each recorded EEG signal and a) the speech envelope I(EEG;SE) and b) each one of the i=1,…,4 extracted PCs (kinematic components) I(EEG;PCi). Before computing MI, we first removed 1.5 s after sound onset for each trial to exclude stimulus-locked evoked potentials and then shifted the EEG signals forward in time by 0.2 s relative to the SE and PCs; The time-shifting was based on the assumption that stimulus encoding necessarily follows stimulus presentation. An extensive literature has indeed consistently showed that speech-brain coupling (entrainment) is maximal at about 0.2-s lag (i.e., for brain activities following auditory/visual speech by 0.2s; Keitel et al., 2017). More specifically, we cut the SE and PCs signals from +1.5 s relative to stimulus onset up to stimulus offset (variable length depending on stimulus duration) and the EEG signals from +1.7 s relative to stimulus onset up to +0.2 s after stimulus offset. Finally, all signals (EEG, SE, PCs) were padded (mirror padding) and then band-pass filtered between 0.5 and 10 Hz (two pass Butterworth, 2nd order). This relatively broad frequency range was set based on prior inspection of the power spectra of both the acoustic (SE) and kinematic (PCs) signals, as it encompasses virtually all of their spectral content (see Figure 14 and Figure 13(b)). The choice of the high cut-off

frequency (10 Hz) is also consistent with evidence that coupling between brain activities and speech envelope is mostly confined to frequencies below the alpha range (Bröhl & Kayser, 2021). MI was then calculated using a recent implementation of the Gaussian Copula Mutual Information method which provides a lower bound of the actual MI and is robust to high-dimensional signals (Ince et al., 2017).

3.4.9 Partial Information Decomposition (PID)

We then focused on a) the contribution of each kinematic component and b) the interactions between speech and kinematic components to the neural encoding of speech. We thus employed Partial Information Decomposition (PID), a recent multivariate mathematical framework, originally proposed in Williams and Beer (2010), to quantify and characterize representational interactions in the human brain.

PID decomposes the mutual information between a target variable and a multivariate set of predictor variables, called sources (Timme et al. 2014). Indeed, if the sources are not statistically independent from the target, they will provide non-zero joint mutual information about the target which, in other terms, indexes the degree of dependence. PID allows then to disentangle this information, parcelling it out into information that is uniquely carried by each of the sources (‘unique’), information that is shared by the sources (‘redundant’) and information that is accessible only when considering the two sources together ('synergistic’).

Here, we considered the EEG measurement at each channel as the target signal. We run 4 different PIDs, every time including as sources: 1) the speech envelope (SE; derived directly from the acoustic stimuli), and 2) one of the 4 kinematic features (PC1, 2, 3, 4; obtained from the EMA data through PCA; see above). The decomposition yielded 4 outcome terms:

• U_{SE} (EEG; SE): The unique information that the speech envelope carries about the EEG signal and cannot be obtained from the kinematic PCi.

• U_{PCi} (EEG; PCi): The unique information that the kinematic PCi carries about the EEG signal and cannot be obtained from the speech envelope.

• SYN_i (EEG; SE, PCi): The information that the joint observation of the two predictors {SE; PCi} provides about the EEG signal that cannot be obtained by observing each predictor separately.

• RED_i (EEG; SE, PCi): The information about EEG that is shared by the two sources, SE and PCi, thus reflecting a common neural representation of speech and kinematic component.

If the interaction between SE and PCi is redundant (RED_i), the information (about the EEG) that is carried by PCi can be obtained also from SE and vice versa. In other words, there will be no information loss if either the SE or the PCi is not available. In contrast, if the interaction is synergistic (SYN_i), neural information is encoded by the relationship between SE and PCi. In other words, we would obtain a better estimate of the EEG signal by considering SE and PCi together rather than independently. Finally, unique information (U) is carried by only one of the two predictors. For example, a significant U_{PCi} would suggest that the corresponding brain response can only be predicted by that specific kinematic signal (PCi) and not by the speech envelope.

PID was performed using a recent modification of the original algorithms which is based on common change in surprisal (Ince, 2017). In a first PID analysis, all the signals were pre-processed in the same way as described above for MI (i.e., epoching, relative time-shifting of EEG data), including band-pass filtering between 0.5 and 10 Hz. After these preprocessing steps, signals for all trials were concatenated and copula normalized (Ince et al., 2017).

To further evaluate whether the acoustic and kinematic features carry information at different spectral ranges, we also performed a frequency-resolved PID analysis. To this end, all signals were band-pass filtered by applying a sliding window along the frequency axis in the range between 0.5 and 10 Hz in steps of 0.5 Hz and with a frequency window length of 1 Hz. A separate PID was then applied for each band-pass filtered set of signals.

3.4.10 Statistical analysis

The output values obtained both in the MI and PID analysis were statistically evaluated against surrogate data. The original relationship between the two signals (for the MI) or between the target (the EEG activity) and the sources (the SE and the PCi) (for the PID) was destroyed without affecting the statistical properties of each signal, including its autocorrelation structure (Montemurro et al. 2007). More specifically, the EEG activity at each electrode and for each trial (epoched and bandpass filtered as for the original analysis) was circularly shifted by a number of samples that was randomly selected between N/4 and N-N/4, where N represent the number of samples of the shortest trial (i.e., 1892 samples). The time-shifted data were then submitted to the same processing steps as described above for the original data (i.e., trial concatenation, copula normalization) before applying the MI/PID algorithms. As for the original analysis (see above), MI was computed between the EEG and each SE/PC feature (I(EEG;SE) and I(EEG;PCi) with i=1,…,4). The same applies for the PID analysis whereby 4 separate PIDs were run by including as sources the SE and one of the 4 kinematic PCs. This procedure was iterated 1000 times yielding a surrogate distribution for each participant and each information component (I_{SE}, I_{PCi}; U_{SE}, U_{PCi}, SYN_{SE-PCi}, RED_{SE-PCi}). We performed group-level statistics by applying one-tail paired-samples t-tests on the original information values (at each electrode) against the mean of the surrogate distribution. We corrected for multiple comparisons across electrodes by controlling the False Discovery Rate (FDR; as described in Benjamini & Yekutieli, 2001). PID results were also statistically evaluated at the single-subject level by computing (separately for each subject and each electrode) the probability that the original information values exceeded the 95% of the surrogate distribution. Again, resulting p-values were corrected for multiple comparisons across electrodes with FDR.

3.5 Results

Neural entrainment to the speech envelope (Meyer, 2018; Giraud & Poeppel 2012; Keitel, Gross, & Kayser, 2018) – as well as the lips motion (Park et al., 2016; Giordano et al. 2017; Ozker, Yoshor & Beauchamp, 2018) – are very well-documented phenomena. However, only a fraction of speech articulation is available to vision while most speech-relevant information is in principle contained in hidden articulators (e.g. tongue movement). We here set out to investigate whether articulatory kinematics that is not directly available to the listener still conveys information about the produced speech that goes above and beyond that contained in the speech envelope. We recorded the electroencephalographic (EEG) brain-wide activity while native-language participants were listening to acoustic speech stimuli taken from the Multi-SPeaKing-style Articulatory corpus (MSPKA; Canevari, Badino, & Fadiga, 2015). This corpus contains simultaneous recordings of audio and kinematic data of the articulatory tract (measured via electromagnetic articulography; see methods) while speakers were pronouncing Italian sentences. The dimensionality of the articulatory data was reduced by means of PCA and the first 4 PCs accounting for most of the variance were selected for further analyses to examine the relationship between the kinematics associated to speech production (in the speaker) and the listener's ongoing brain activity.

3.5.1 Kinematic principal components

The first 4 components derived from PCA explained most of the total variance of the kinematic data (85%; Figure 12(b)). Inspection of the PC weights indicates that the first 2 components (PC1 and PC2) represented almost entirely movements of the tongue on the antero-posterior (x-) and vertical (z-) axis, respectively (Figure 12(c)). Two of the movements that contribute significantly to articulation (Perrier et al., 2007), such as that of the tongue towards (and away from) the lips (PC1) or the palate (PC2), were thus automatically identified by PCA. Lower lip and jaw (again along the antero-posterior as well as vertical axes) as well as the tongue mainly

contributed to PC3 which, despite explaining a smaller amount of variance (10%) compared to the tongue movements (PC1: 52%; PC2: 17%), appeared to capture another meaningful articulatory component, reflecting most likely mouth opening/closing, lip protrusion and lip-tongue coordination. Finally, a more composite mixture of articulators moving along multiple directions contribute to PC4 (6%), possibly reflecting complex tongue-lip movement synergies. The remaining components explained negligible amounts of variance (<5% each) and their articulatory interpretation was less straightforward; these components were thus excluded from further analysis.

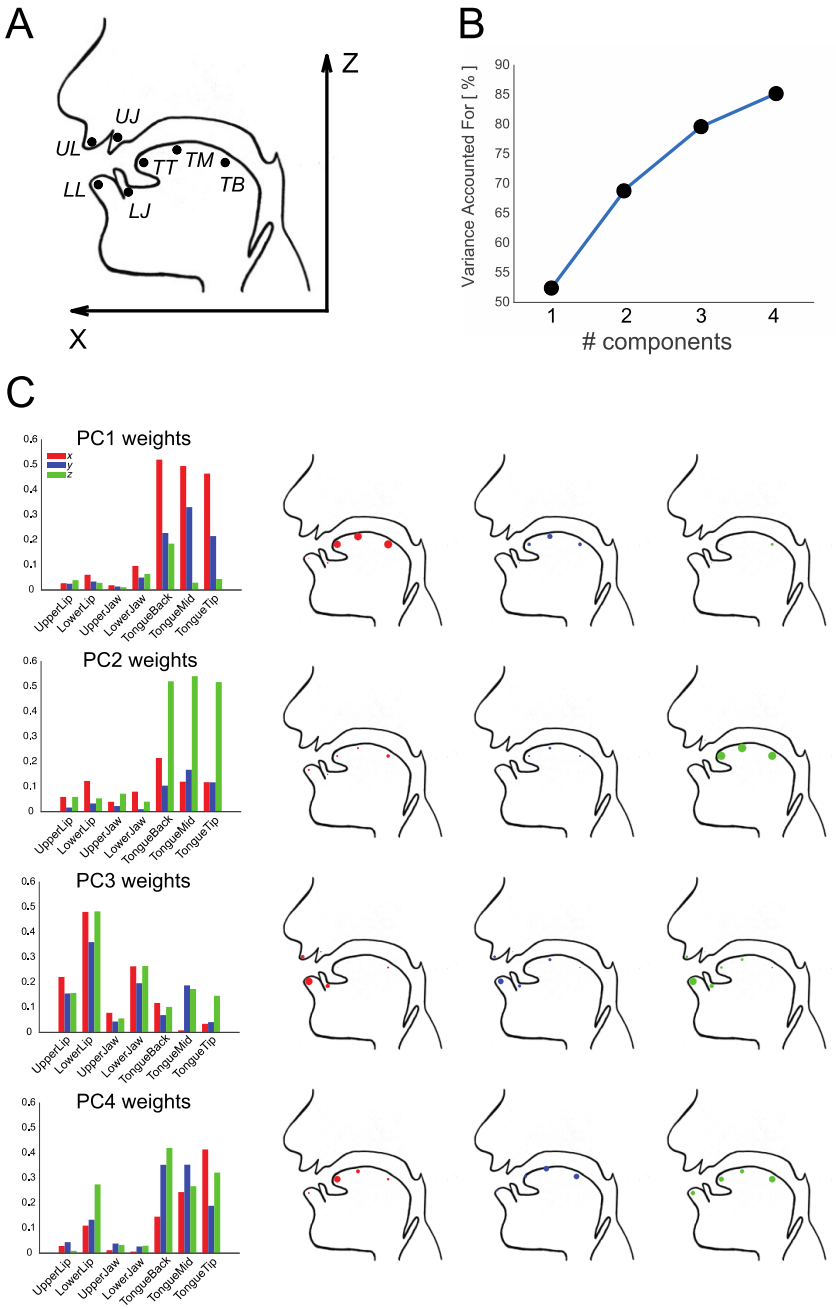

Figure 12. Kinematic principal components. A. Schematic of the positions of the electromagnetic sensors: upper lip (UL), lower lip (LL), upper jaw (UJ), lower jaw (LJ), tongue tip (TT), tongue middle (TM), tongue back (TB). B. Cumulative variance (%) of kinematic data that is explained by the first four principal components (PC1, 2, 3, 4). C. Bar plots represent the weights (absolute values) of each kinematic variable (x-, y- and z-axis for each sensor) for the PC1, 2, 3, 4. Dot size in the three vocal tract schematics

show the relative contribution of each sensor across the movement axis (x, y and z respectively in red, blue and green).

Examples of the reconstructed time series for the 4 retained kinematic components along with the corresponding speech envelope are shown in Figure 13(a). Analysis of their spectral content reveals that all the kinematic components show spectral concentration over a low frequency range between 1 and 4 Hz (delta band); the speech envelope instead, in line with previous evidence (Bröhl & Kayser 2021; Doellin et al., 2014; Gross et al., 2013; Luo & Poeppel, 2007; Peelle & Davis, 2012; Bosker & Ghitza 2018), is marked by relatively higher frequencies, with a broad spectral peak between 4 and 8 Hz (theta band).

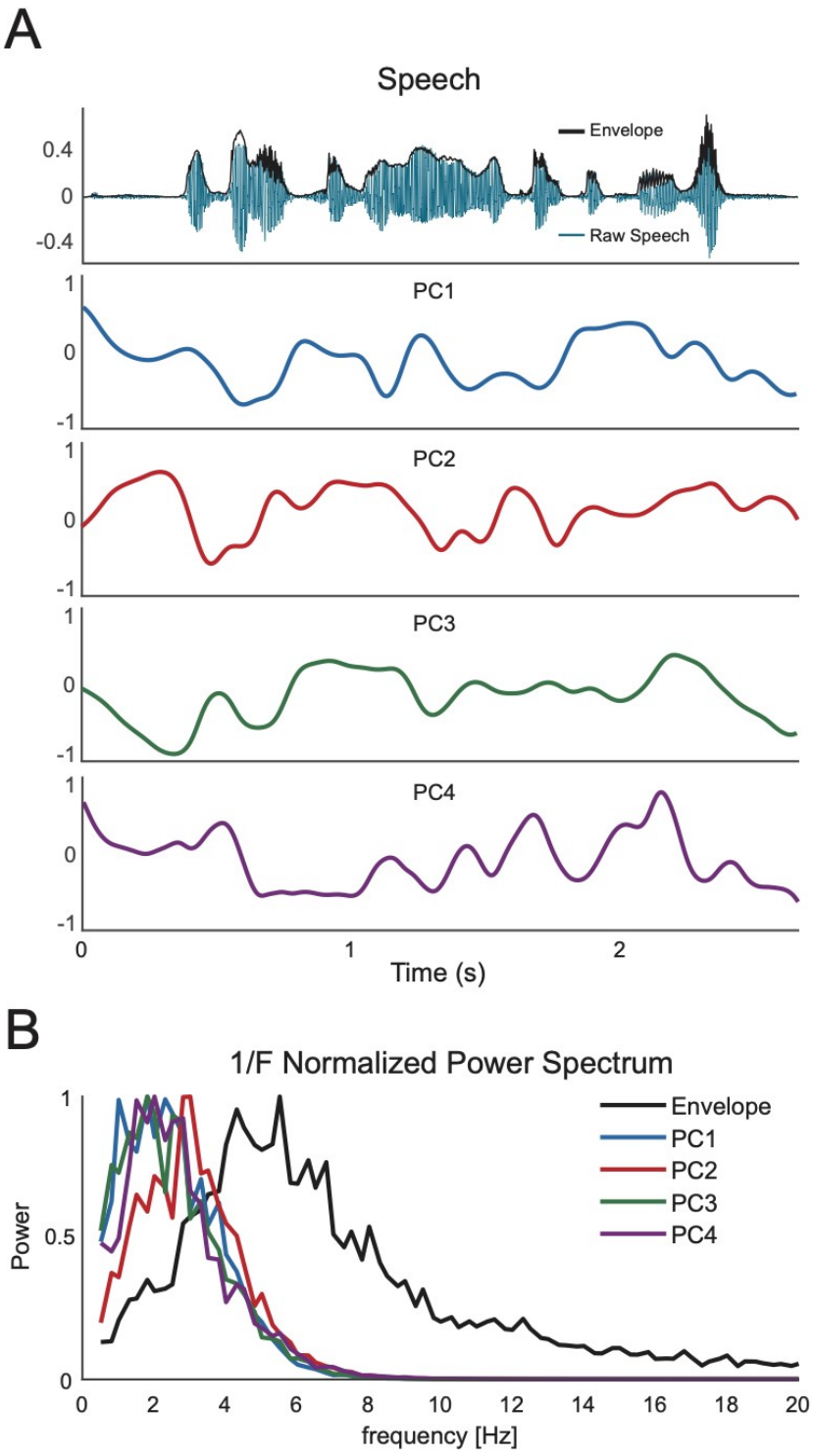

Figure 13. Acoustic and kinematic stimulus features. A. Example time series of the raw speech signal (blue), its envelope (black) and the kinematic PCs corresponding to the same stimulus. B. Normalized power spectra for all features (envelope, PC1, PC2, PC3 and PC4).

3.5.2 Neural entrainment to speech envelope and tongue kinematics

Firstly, we evaluated whether the brain encodes the information contained in the speech envelope as well as in the hidden speech kinematics by computing the mutual information (MI; Shannon, 1948, Ince et al., 2017) between the respective signal pairs (i.e., I(EEG;SE); I(EEG;PCi) with i=1,…,4). As expected, MI for the speech envelope was significant and maximal in two distinct foci, one overlaying the central electrodes and the other distributed more laterally over the right temporo-parietal electrodes (see Figure 14; paired-samples one-tail t-tests against surrogate data with FDR-correction for multiple comparisons across channels; see Methods for details). Such a topography is indeed very similar to what reported in previous works when quantifying neural speech entrainment with linear (Molinaro & Lizarazu 2018; Boucher, Gilbert & Jemel 2019) as well as non-linear (Kayser et al., 2015) coupling metrics. Remarkably, we found significant MI also for one of the analyzed kinematic features, i.e., PC1, which mainly relates to the antero-posterior movement of the tongue (see Figure 14 and Figure 12(c)). MI for PC1 largely increases over central electrodes – similar to what obtained for the speech envelope – but also symmetrically over right- and left-sided temporo-parietal electrodes (see Figure 14).

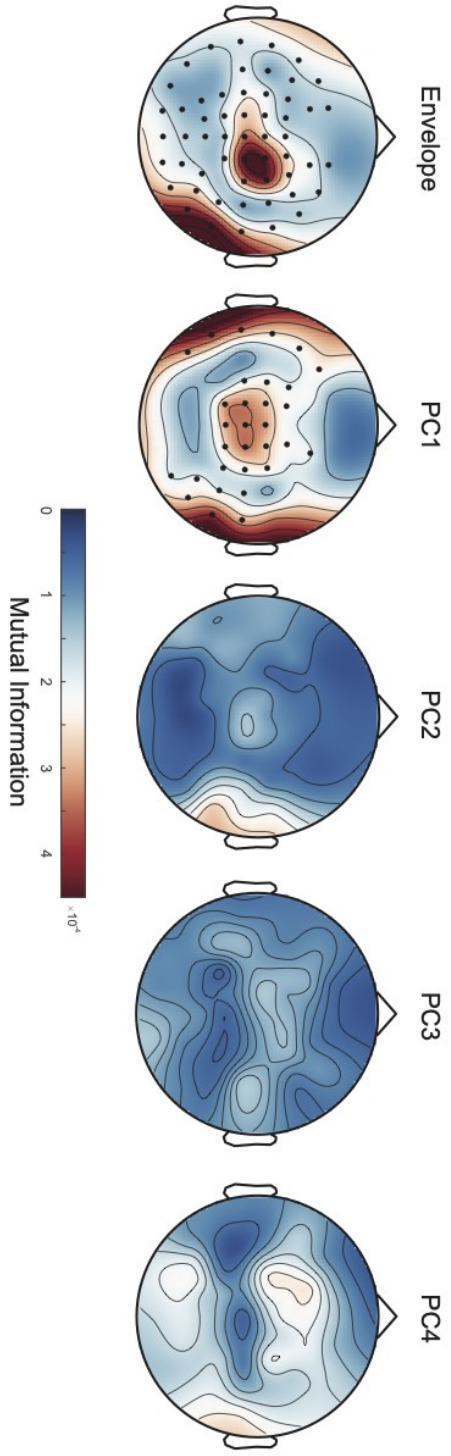

Figure 14. MI results. Topographical distribution of across-subjects average information values computed via Gaussian Copula Mutual Information performed on the broad-band filtered data (0.5-10Hz). Black dots highlight significant channels (after FDR-correction for multiple comparisons).

3.5.3 Partial Information Decomposition

The speech envelope and the tongue movements (PC1) could, however, provide (brain-relevant) information that is either exclusive to each feature, fundamentally shared, or complementary between the two features. To disentangle among these different possibilities, we employed a computational approach known as Partial Information Decomposition (PID) (Williams & Beer, 2010) that decomposes the total information which the sources – here the speech envelope (SE) and the kinematic features (PCi) – carry about the target – here the EEG activity – into four distinct terms that we denote as follows: 1) U_{SE}, the Unique information provided by SE; 2) U_{PCi}, the Unique information provided by the kinematic feature (PCi); 3) SYN_i, the synergistic information which can only be obtained when SE and PCi are considered together; 4) RED_i, the redundant information which is common to SE and PCi.

Not surprisingly, the speech envelope carries unique information in the 4 different PID models (Figure 15, first column; paired-samples one-tail t-tests against surrogate data with FDR-correction for multiple comparisons across channels; see Methods for details). The topographic distribution of such activity is very similar across all PID models and closely resembles that already observed for the MI.

Most remarkably, passive listening also entails neural encoding of kinematic information that is not accounted for by the speech amplitude fluctuations, i.e., the SE. Specifically, the PC1 provides unique information (not carried by the SE; U_{PC1}) that are consistently represented in the listener's brain (see Figure 15, first row). PC1 not only carries unique informational content but also interacts significantly with the acoustic information in a synergistic fashion; in other words, its combination with the SE leads to a net increase in information encoded in the listener's brain ($SYN_{SE\text{-}PC1}$; Figure 15, third column). Although far less strongly and with a sparser spatial distribution, we also observe significant synergic information between the SE and PC3 (mainly representing the coordination between mouth opening/closing and tongue motion). Indeed, PID enables to uncover representational interactions in the listener's brain that cannot be directly observable in pairwise measures of dependence, such as MI here. No redundancy

is instead observed between the PC1 (as well as any of the analyzed kinematic features) and the acoustic speech envelope, suggesting that the listener's brain mainly exploits them as independent sources of information and/or integrates them in a super-additive way.

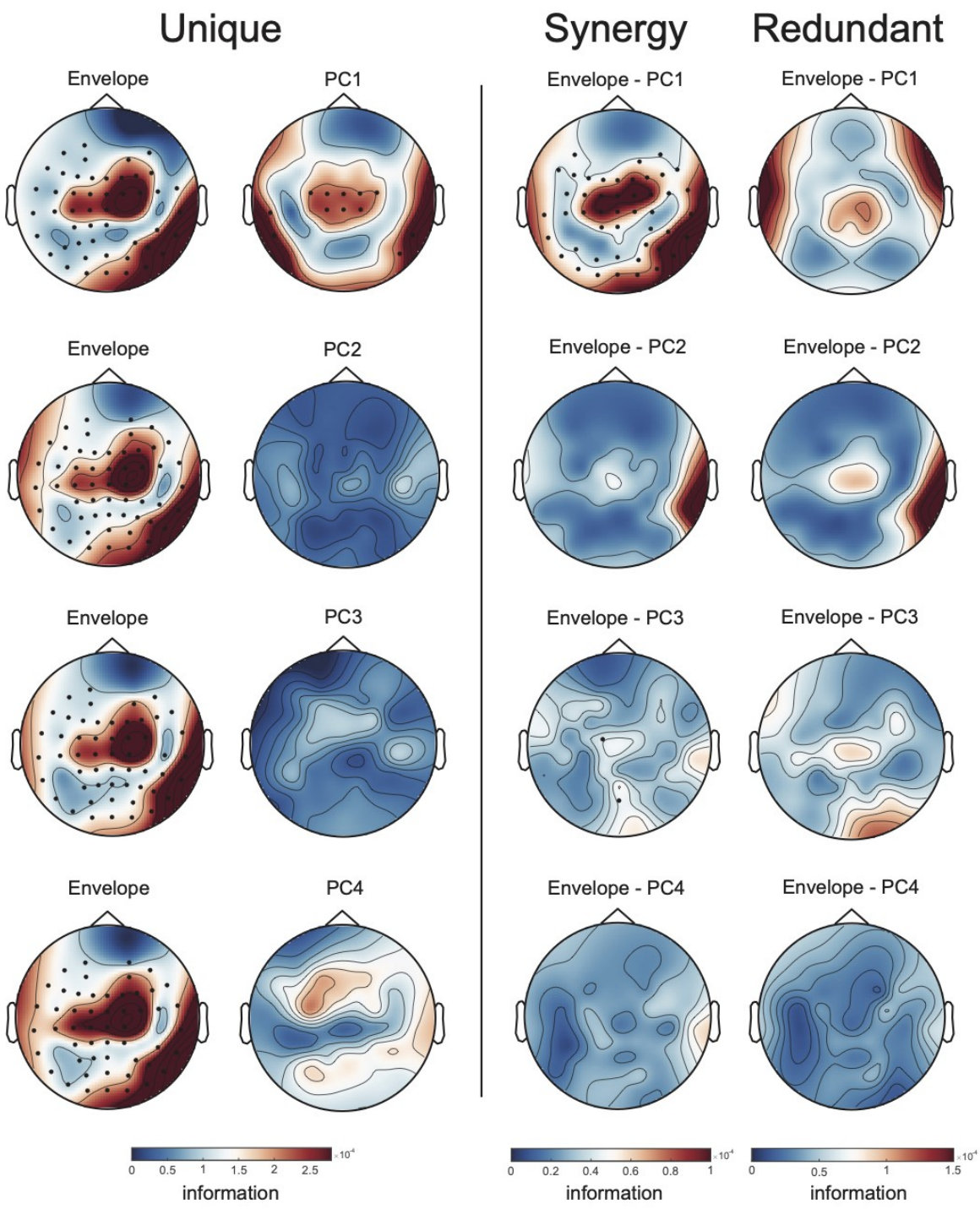

Figure 15. PID results. Topographical distribution of across-subjects average information values obtained by PID analyses performed on the broad-band filtered data (0.5-10Hz). Black dots highlight significant channels (after FDR-correction for multiple comparisons).

Overall, the same pattern of results is observed also when statistical evaluation is performed at the single-subject level, with a large proportion of participants showing significant U_{SE} and U_{PC1}. All PID systems resulted in 14 subjects having at least one significant channel for the SE unique information (U_{SE}) while PC1 unique information (U_{PC1}) was significant in 11 participants (Figure 16). U_{PCi} was significant for 0, 1, 3 subjects for PC2, PC3 and PC4, respectively. Synergistic information between SE and PC1 was significant in 14 subjects while far less for the other components (8, 4 and 1 subjects for PC2, PC3 and PC4, respectively). A non-negligible number of participants also report significant redundant information between SE and the first three PCs (8, 7, 6 subjects, respectively), suggesting that substantial inter-individual variability in the spatial distribution of these effects may have negatively impacted corresponding group-level statistics. Single subjects results are summarised in Table 2.

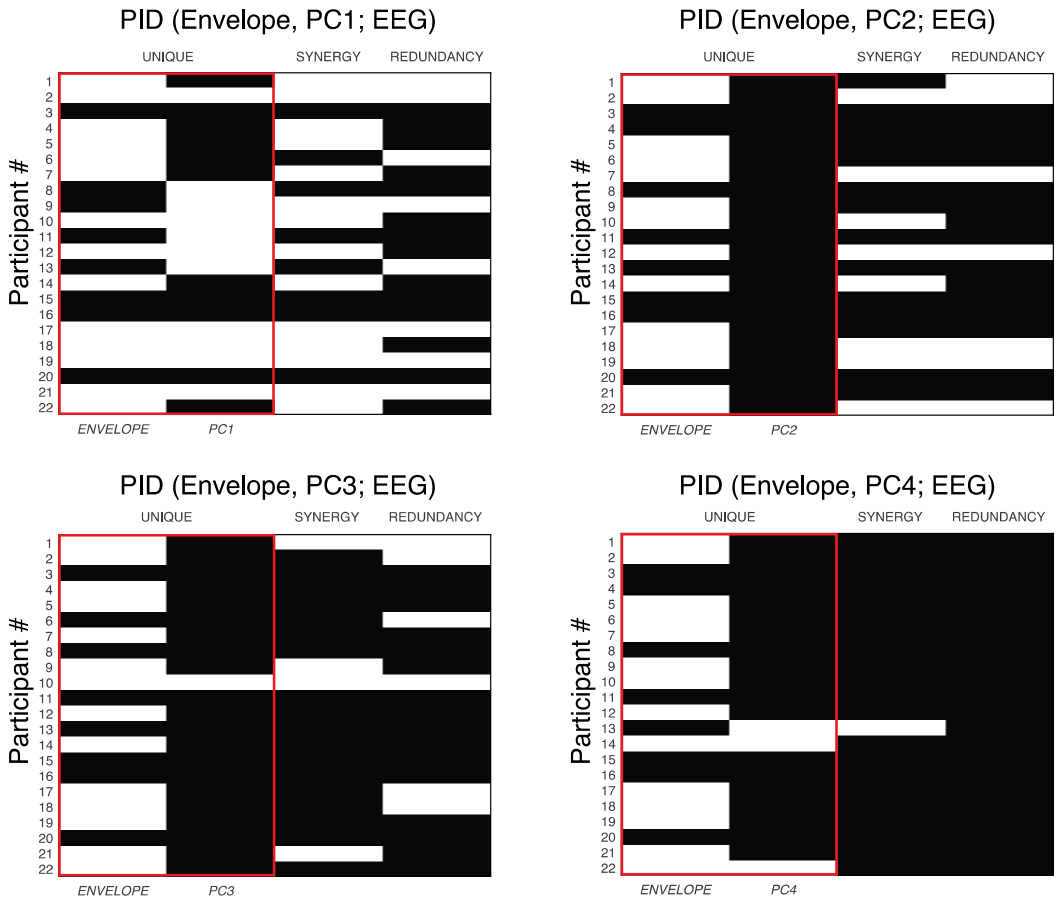

Figure 166. *Single subject results. Statistical tests run on individual subjects are shown for each PID component. White squares indicate subjects where at least one significant channel was obtained (after FDR-correction for multiple comparisons).*

	Unique SE	Unique PC_i	Synergy SE+PC_i	Redundancy SE+PC_i
PC_1	14	11	14	8
PC_2	14	0	8	7
PC_3	14	1	4	6
PC_4	14	3	1	0

Table 2. Single subject results. Statistical tests run on individual subjects are shown for each PID component. The amount of subjects showing at least one significant channel is reported (after FDR-correction for multiple comparisons).

3.5.4 The kinematic information encoded in the brain

The above-reported results indicate that the brain encodes at least a certain amount of information carried by articulatory kinematics, in particular that captured by the PC1, that cannot be equivalently extracted from the speech envelope. To gain insight into whether information conveyed by the kinematic and acoustic signals is band-limited and possibly concentrated within distinct frequency ranges, we repeated the PID analysis in a frequency-resolved way. Figure 17 shows the outcome of PID as a function of frequency for those components of information for which we obtained significant results in the previous (broadband) analysis (see above). The information that is uniquely carried by the SE is clearly spectrally selective with maximal values being observed at ~5-6 Hz, i.e., in the theta band. A very similar spectral selectivity characterizes also the information that is conveyed by the synergistic interaction between SE and PC1 or PC3. A different spectral fingerprint, however, marks the information that is uniquely contained in the kinematics (PC1) which is enhanced within a lower frequency range between 1 and 4 Hz, i.e., in the delta band.

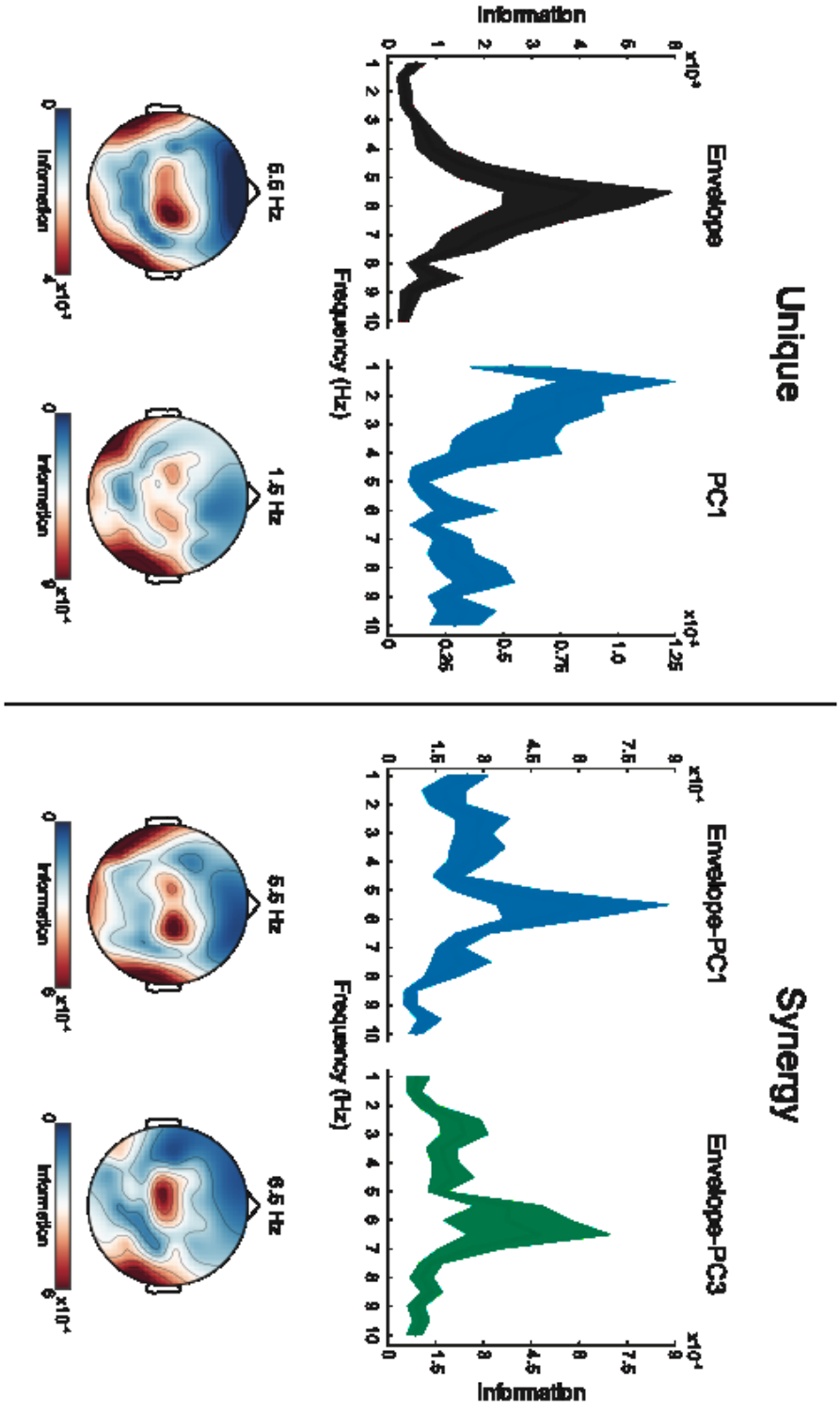

Figure 17. Frequency-resolved PID analyses. Results as a function of frequency for the significant information obtained on broad-band filtered data (see Figure 15): Unique information (Envelope, PC1) and Synergistic information (Envelope-PC1 and Envelope-PC3). Topographies are shown for the frequency where information is maximal.

3.6 Discussion

Neural entrainment to speech originates from the integration of bottom-up processing and top-down projections from higher-order functional nodes in the brain to select and isolate particular signals of interest (Rimmele et al., 2018). Top-down projections, based on context and prior learning, bias bottom-up sensory sampling via predictive models operating at multiple levels (e.g. Meyer, 2018; Keitel, Gross & Kayser, 2018). For instance, the neural computations run in the motor system may provide key top-down signals to isolate segmental or suprasegmental cues (Giraud & Poeppel, 2012). In fact, the sensory ambiguity characterizing the acoustic stream may partially be solved by unambiguous (or less ambiguous) endogenous signals (Meyer, Sun & Martin, 2020) arising from the inherent rhythmicity in speech articulation (Poeppel & Assaneo, 2020). Yet, for such a claim to be tenable one should be able to find traces of articulatory signals in brain activities of speech-listening participants. Most importantly, neural activities should encode articulatory information in a way that is not trivially explained by the encoding of other tightly coupled speech acoustic features (i.e. mouth opening and speech envelope; Chandrasekaran et al., 2009). Here we show that vocal tract configurations are encoded in the EEG signal and that they contain information that goes above and beyond that carried by the speech envelope.

This result was obtained through the combination of a series of targeted novel approaches. First and foremost, participants listened to a set of acoustic speech stimuli for which synchronized articulatory data was available (Canevari, Badino & Fadiga, 2015). EMA data is characterized by a spatial and temporal resolution (Rebernik et al. 2021) that is not otherwise achievable through other technologies for speech kinematic analysis (i.e. Ultrasound or MRI). Moreover, as it is customary in the field of motor control (Ting, 2007), we reduced the dimensionality of the data (i.e. PCA) to derive a tractable number of signals explaining most of the variance (Lambert-Shirzad & Van del Loos, 2017). Yet, this choice has also key theoretical implications since data reduction techniques capture physiologically meaningful data-driven patterns of coordination across articulators (i.e. vocal tract synergies), which have far greater functional relevance than the raw time-varying spatial

positions of isolated articulators (Story, 2005; Sorensen et al., 2019). Secondly, we used a mathematical framework (Partial Information Decomposition – PID; Williams & Beer 2010; Ince 2017) that captures complex nonlinear relations between variables and decompose these relations into atoms of information between the target (i.e. EEG data) and the sources (i.e. speech envelope and speech kinematics principal components). The PID can indeed extract unique, synergistic or redundant information contained in the sources (Park et al., 2018; Daube et al., 2019; Delis et al., 2022).

As expected from the abundant literature on neural entrainment to speech, the PID analysis highlights that the speech envelope contains unique information encoded in the EEG data. The topographical distribution of this effect matches the one normally observed with other analytical methods, with the involvements of central and right temporo-parietal electrodes (Ding et al., 2017). At the same time, we found that the first kinematic principal component, reflecting the antero-posterior movement of the tongue, also carries unique information about the neural signals. The topography of this effect is confined to central and bilateral temporal electrodes. Interestingly, the synergistic interaction between the speech envelope and the first or third kinematic components (PC3 describes the coordination between lips, jaw and tongue movements) convey additional information about the EEG data. In both cases, the topographies show a distribution covering both central and right temporo-parietal electrodes.

Importantly, the PID analyses reported no redundancy between any of the kinematic components and speech envelope. This result is further supported by the frequency dissociation emerged in the subsequent frequency resolved PID analysis. Lower frequencies (from ~0.5 to 4 Hz) appear relevant for kinematics-specific information (unique information provided by the PC1), in agreement with prior evidence that entrainment in the delta-range originates from higher-level processes in frontal (Molinaro et al., 2016) and motor cortices in particular (Park et al., 2015; Morillon et al., 2019; Biau et al., 2022). Instead, higher frequencies (between 4 and 8 Hz) contain unique information carried by the speech envelope and, most importantly, by the synergistic interaction between kinematics (PC1 or PC3) and speech envelope. Such a delta/theta dissociation is compatible with the idea that neural entrainment in the theta-band is associated to the phonetic

features that are critical for speech recognition, while the delta range entrainment is more closely related to the perceived acoustic rhythm of speech (Ding & Simon, 2014; Meyer, Sun & Martin, 2020). All in all, our results offer new insight regarding the functional origin of the delta/theta dissociation observed in speech neural entrainment, especially regarding the contribution provided by domain-specific motor processes.

In fact, the goal of our study was to investigate if speech listening does entail neural coupling to highly granular speech kinematic information that is not readily available to the participants. In our study, participants listened to auditory speech signals and were never – explicitly nor implicitly – exposed to the articulatory side of speech production. Recent studies showed that brain signals encode missing information such as acoustic features when only silent lip-reading is allowed (Hauswald et al., 2018; Bourguignon et al., 2020). In this case, the tight audio-visual contingencies experienced during early childhood (as well as throughout life; Chandrasekaran et al., 2009) offer a solid ground to explain these phenomena according to a Bayesian perspective and as the result of multimodal associative learning (van Wassenhove, 2013). In our case, kinematic data contain information that is not available during the experiment nor visually accessible throughout life (i.e. tongue kinematics). It follows that neural coupling to unavailable information cannot be explained by the life-long learning of audio-visual associations (i.e. as is the case for lip motion).

Neural coupling to tongue kinematics would still require that (at least part of the) articulatory information is retrieved from speech acoustics. Yet, the mapping between acoustic speech targets and articulatory configurations is tremendously complex (Atal et al., 1978). Known as the *"one to many"* mapping problem, it forces the brain to solve an ill-posed inverse problem. An answer to this conundrum is that, during development, the brain approximates a solution for this inverse problem - mapping intended acoustic targets back to vocal tract articulatory parameters to allow intelligible speech productions (Guenther, 1995; Tourville & Guenther, 2011). Indeed, infants explore how sounds are produced by experimenting the full range of their vocal tract configurations (Bruderer et al., 2015; Kuhl et al., 2014). In support of this idea, automatic speech recognition models trained with both acoustic and articulatory data achieve better classification

performance with far fewer examples than acoustic-only training regimes (King et al., 2007; Gosh & Narayanan, 2011). These models recapitulate some key properties of speech production development (Canevari et al., 2013; Badino et al., 2014) and demonstrate that learning auditory-motor mappings grants more compact and efficient representations of speech acoustics (Badino et al., 2016). As a consequence of learning audio-motor contingencies, speech auditory processing should in principle be tuned to capture those cues that allow triggering of endogenously guided reconstruction of missing articulatory signals (Meyer, Sun & Martin, 2020). To date, however, there was no evidence that speech neural entrainment encodes motor signals whose relevance is functionally dependent on the acquisition of speech production - and thus reflecting an intrinsically domain specific process.

Yet, the idea of the motor system involved in speech perception is not a new concept (Pulvermuller &Fadiga, 2010; Fadiga et al., 2002; Watkins et al., 2003, Wilson et al., 2004, Pulvermüller et al., 2006) and transcranial magnetic stimulation of the motor cortex has been shown to produce specific (Meister et al., 2007; Möttönen et al., 2009; Sato et al., 2009) and somatotopic effects on speech discrimination performance (D'Ausilio et al., 2009; Bartoli et al., 2015). A recent series of studies proposed a more detailed oscillation-based mechanism through which the motor system could impact on speech perception. Assaneo and Poeppel (2018) found synchronized brain activity between motor and auditory areas during a syllable listening task and successfully modelled the speech motor cortex as an oscillator coupled to the auditory system. According to this model, neuronal oscillations observed in auditory and motor cortices indeed synchronize in a frequency range corresponding to the mean syllable rate across languages (~4.5 Hz). Endogenous signals from the motor system would phase-reset neuronal oscillations in the auditory cortex to align the most excitable states to the occurrence of an expected event (Rimmele et al., 2018) with benefits on perceptual performance (Assaneo et al., 2021).

Concerning the functional relevance of entraining auditory cortices to endogenous motor rhythms, the most likely explanation is that perception is an inherently noisy process that, in order to cope with speech-intrinsic (talker-specific) and speech-extrinsic (environment-specific) noise (Ru, Chi & Shamma, 2003), integrates and

weighs multiple sources of information depending on their reliability (Golumbic et al., 2013b; Schroeder et al., 2008). In this regard, when the acoustic signal is corrupted the increased importance of visual cues is evident in stronger entrainment to lip movements (Giordano et al., 2017; O'Sullivan et al., 2021; Park et al., 2016; Peelle & Sommers, 2015; Park et al., 2018). Here, we provide a demonstration that neural speech processing can draw inferences based on highly granular endogenous domain-specific motor signals whose relevance for perception necessarily derives from the acquisition of speech production.

3.7 References

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

Arnal, L. H., & Giraud, A. L. (2012). Cortical oscillations and sensory predictions. *Trends in cognitive sciences*, *16*(7), 390-398.

Assaneo, M. F., & Poeppel, D. (2018). The coupling between auditory and motor cortices is rate-restricted: Evidence for an intrinsic speech-motor rhythm. *Science advances*, *4*(2), eaao3842.

Assaneo, M. F., Rimmele, J. M., Sanz Perl, Y., & Poeppel, D. (2021). Speaking rhythmically can shape hearing. *Nature human behaviour*, *5*(1), 71-82.

Atal, B. S., Chang, J. J., Mathews, M. V., & Tukey, J. W. (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *The Journal of the Acoustical Society of America*, *63*(5), 1535-1555.

Badino, L., Canevari, C., Fadiga, L., & Metta, G. (2016). Integrating articulatory data in deep neural network-based acoustic modeling. *Computer Speech & Language*, *36*, 173-195.

Badino, L., D'Ausilio, A., Fadiga, L., & Metta, G. (2014). Computational validation of the motor contribution to speech perception. *Topics in cognitive science*, *6*(3), 461-475.

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

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165-1188.

Biau, E., Schultz, B. G., Gunter, T. C., & Kotz, S. A. (2022). Left motor delta oscillations reflect asynchrony detection in multisensory speech perception. *Journal of Neuroscience*.

Bosker, H. R., & Ghitza, O. (2018). Entrained theta oscillations guide perception of subsequent speech: behavioural evidence from rate normalisation. *Language, Cognition and Neuroscience*, *33*(8), 955-967.

Boucher, V. J., Gilbert, A. C., & Jemel, B. (2019). The role of low-frequency neural oscillations in speech processing: revisiting delta entrainment. *Journal of cognitive neuroscience*, *31*(8), 1205-1215.

Bourguignon, M., Baart, M., Kapnoula, E. C., & Molinaro, N. (2020). Lip-reading enables the brain to synthesize auditory features of unknown silent speech. *Journal of Neuroscience*, *40*(5), 1053-1065.

Bröhl, F., & Kayser, C. (2021). Delta/theta band EEG differentially tracks low and high frequency speech-derived envelopes. *Neuroimage*, *233*, 117958.

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

Calderone, D. J., Lakatos, P., Butler, P. D., & Castellanos, F. X. (2014). Entrainment of neural oscillations as a modifiable substrate of attention. *Trends in cognitive sciences*, *18*(6), 300-309.

Canevari, C., Badino, L., & Fadiga, L. (2015). A new Italian dataset of parallel acoustic and articulatory data. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Canevari, C., Badino, L., D'Ausilio, A., Fadiga, L., & Metta, G. (2013). Modeling speech imitation and ecological learning of auditory-motor maps. *Frontiers in Psychology*, *4*, 364.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & 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.

Cope, T. E., Sohoglu, E., Sedley, W., Patterson, K., Jones, P. S., Wiggins, J., ... & Rowe, J. B. (2017). Evidence for causal top-down frontal contributions to predictive processes in speech perception. *Nature Communications*, *8*(1), 1-16.

Daube, C., Ince, R. A., & Gross, J. (2019). Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. *Current Biology*, *29*(12), 1924-1937.

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

Delis, I., Ince, R. A., Sajda, P., & Wang, Q. (2022). Neural encoding of active multi-sensing enhances perceptual decision-making via a synergistic cross-modal interaction. *Journal of Neuroscience*.

Di Liberto, G. M., Lalor, E. C., & Millman, R. E. (2018). Causal cortical dynamics of a predictive enhancement of speech intelligibility. *Neuroimage*, *166*, 247-258.

Ding, N., & Simon, J. Z. (2012). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of neurophysiology*, *107*(1), 78-89.

Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: functional roles and interpretations. *Frontiers in human neuroscience*, *8*, 311.

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

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage*, *85*, 761-768.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *European journal of Neuroscience*, *15*(2), 399-402.

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.

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

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

Golumbic, E. M. Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., ... & Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”. *Neuron*, *77*(5), 980-991.

Golumbic, E. M. Z., Poeppel, D., & 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., & Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a “cocktail party”. *Journal of Neuroscience*, *33*(4), 1417-1426.

Gramfort, Alexandre, et al. "MEG and EEG data analysis with MNE-Python." *Frontiers in neuroscience* (2013): 267.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS biology*, *11*(12), e1001752.

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological review*, *102*(3), 594.

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

Ince, R. A. (2017). The Partial Entropy Decomposition: Decomposing multivariate entropy and mutual information via pointwise common surprisal. *arXiv preprint arXiv:1702.01591*.

Ince, R. A., Giordano, B. L., Kayser, C., Rousselet, G. A., Gross, J., & Schyns, P. G. (2017). A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. *Human brain mapping*, *38*(3), 1541-1573.

Kayser, S. J., Ince, R. A., Gross, J., & Kayser, C. (2015). Irregular speech rate dissociates auditory cortical entrainment, evoked responses, and frontal alpha. *Journal of Neuroscience*, *35*(44), 14691-14701.

Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS biology*, *16*(3), e2004473.

Keitel, A., Ince, R. A., Gross, J., & Kayser, C. (2017). Auditory cortical delta-entrainment interacts with oscillatory power in multiple fronto-parietal networks. *Neuroimage*, *147*, 32-42.

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

King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., & Wester, M. (2007). Speech production knowledge in automatic speech recognition. *The Journal of the Acoustical Society of America*, *121*(2), 723-742.

Kösem, A., Bosker, H. R., Jensen, O., Hagoort, P., & Riecke, L. (2020). Biasing the perception of spoken words with transcranial alternating current stimulation. *Journal of cognitive neuroscience*, *32*(8), 1428-1437.

Kösem, A., Bosker, H. R., Takashima, A., Meyer, A., Jensen, O., & Hagoort, P. (2018). Neural entrainment determines the words we hear. *Current Biology*, *28*(18), 2867-2875.

Kuhl, P. K., Ramírez, R. R., Bosseler, A., Lin, J. F. L., & Imada, T. (2014). Infants' brain responses to speech suggest analysis by synthesis. *Proceedings of the National Academy of Sciences*, *111*(31), 11238-11245.

Lambert-Shirzad, N., & Van der Loos, H. M. (2017). On identifying kinematic and muscle synergies: a comparison of matrix factorization methods using experimental data from the healthy population. *Journal of neurophysiology*, *117*(1), 290-302.

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

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

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., Sun, Y., & Martin, A. E. (2020). Synchronous, but not entrained: exogenous and endogenous cortical rhythms of speech and language processing. *Language, Cognition and Neuroscience*, *35*(9), 1089-1099.

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

Molinaro, N., Lizarazu, M., Lallier, M., Bourguignon, M., & Carreiras, M. (2016). Out-of-synchrony speech entrainment in developmental dyslexia. *Human brain mapping*, *37*(8), 2767-2783.

Montemurro, M. A., Senatore, R., & Panzeri, S. (2007). Tight data-robust bounds to mutual information combining shuffling and model selection techniques. *Neural Computation*, *19*(11), 2913-2957.

Morillon, B., & Baillet, S. (2017). Motor origin of temporal predictions in auditory attention. *Proceedings of the National Academy of Sciences*, *114*(42), E8913-E8921.

Morillon, B., Arnal, L. H., Schroeder, C. E., & Keitel, A. (2019). Prominence of delta oscillatory rhythms in the motor cortex and their relevance for auditory and speech perception. *Neuroscience & Biobehavioral Reviews*, *107*, 136-142.

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

Obleser, J., & Kayser, C. (2019). Neural entrainment and attentional selection in the listening brain. *Trends in cognitive sciences*, *23*(11), 913-926.

Oostenveld R, Fries P, Maris E, Schoffelen JM. FieldTrip: Open source software for advanced analysisof MEG, EEG, and invasive electrophysiological data. Computational intelligence and neuroscience.2011; 2011:156869. Epub 2011/01/22. https://doi.org/10.1155/2011/156869 PMID: 21253357

O'Sullivan, A. E., Crosse, M. J., Di Liberto, G. M., de Cheveigné, A., & Lalor, E. C. (2021). Neurophysiological indices of audiovisual speech processing reveal a hierarchy of multisensory integration effects. *Journal of Neuroscience*, *41*(23), 4991-5003.

O'sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., ... & Lalor, E. C. (2015). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral cortex*, *25*(7), 1697-1706.

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

Park, H., Ince, R. A., Schyns, P. G., Thut, G., & 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., & Gross, J. (2018). 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., & Gross, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *Elife*, *5*, e14521.

Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in psychology*, *3*, 320.

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169-181.

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral cortex*, *23*(6), 1378-1387.

Perrier, P., Perkell, J., Payan, Y., Zandipour, M., Guenther, F., & Khalighi, A. (2007). Degrees of freedom of tongue movements in speech may be constrained by biomechanics. *arXiv preprint arXiv:0709.1405*.

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.

Poeppel, D., & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature reviews neuroscience*, *21*(6), 322-334.

Pulvermüller, F., & Fadiga, L. (2010). Active perception: sensorimotor circuits as a cortical basis for language. *Nature reviews neuroscience*, *11*(5), 351-360.

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

Rebernik, T., Jacobi, J., Jonkers, R., Noiray, A., & Wieling, M. (2021). A review of data collection practices using electromagnetic articulography. *Laboratory Phonology*, *12*(1).

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

Rimmele, J. M., Morillon, B., Poeppel, D., & Arnal, L. H. (2018). Proactive sensing of periodic and aperiodic auditory patterns. *Trends in cognitive sciences*, *22*(10), 870-882.

Ru, P., Chi, T., & Shamma, S. (2003). The synergy between speech production and perception. *The Journal of the Acoustical Society of America*, *113*(1), 498-515.

Sato, M., Tremblay, P., & 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., & 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.

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

Schubotz, R. I. (2007). Prediction of external events with our motor system: towards a new framework. *Trends in cognitive sciences*, *11*(5), 211-218.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379-423.

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

Sorensen, T., Toutios, A., Goldstein, L., & Narayanan, S. (2019). Task-dependence of articulator synergies. *The Journal of the Acoustical Society of America*, *145*(3), 1504-1520.

Story, B. H. (2005). Synergistic modes of vocal tract articulation for American English vowels. *The Journal of the Acoustical Society of America*, *118*(6), 3834-3859.

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

Timme, N., Alford, W., Flecker, B., & Beggs, J. M. (2014). Synergy, redundancy, and multivariate information measures: an experimentalist's perspective. *Journal of computational neuroscience*, *36*(2), 119-140.

Ting, L. H. (2007). Dimensional reduction in sensorimotor systems: a framework for understanding muscle coordination of posture. *Progress in brain research*, *165*, 299-321.

Tourville, J. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. *Language and cognitive processes*, *26*(7), 952-981.

van Wassenhove, V. (2013). Speech through ears and eyes: interfacing the senses with the supramodal brain. *Frontiers in psychology*, *4*, 388.

Vander Ghinst, M., Bourguignon, M., de Beeck, M. O., Wens, V., Marty, B., Hassid, S., ... & De Tiege, X. (2016). Left superior temporal gyrus is coupled to attended speech in a cocktail-party auditory scene. *Journal of Neuroscience*, *36*(5), 1596-1606.

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

Williams, P. L., & Beer, R. D. (2010). Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature neuroscience*, *7*(7), 701-702.

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

4 Research Project: Speech Interaction

4.1 Personal contribution

The text and figures in the following paragraphs are the summary of my work and results on measuring acoustic speech convergence. The project is still ongoing because few criticalities have to be addressed. I designed the model architecture from scratch and the deep learning paradigm, from datasets selection and processing to model training, validation and testing.

4.2 Introduction

Measuring acoustic speech convergence is a very complex problem as proved by a long history of debates on how to define it. Recently, researchers used machine learning (ML) models to estimate convergence (Mukherjee, 2017; Ostrand & Chodroff, 2021). If properly selected and trained, indeed, ML models can, in theory, learn every kind of relation between variables, predicting very complex phenomena.

The paradigm that we followed to address the problem of measuring acoustic convergence consisted in creating an effective speaker verification model. The speaker verification is a system that, given two phrases or words is capable to discriminate if the two acoustic signals are pronounced by the same speaker or not, producing a similarity score between them.

The similarity score is basically a continuous value ranging between two extrema, usually 0 and 1, representing how far the two voices are. Thresholding the

similarity allows to obtain a binary label, representing the decision of the model about considering or not the two voices belonging to same speaker.

Nonetheless, the similarity score is a continuous value as mentioned. Hence, the idea behind the use of a speaker verification system consisted in using the output of the model to track the variation of similarity between the speakers' acoustic signatures.

In principle, the distance between the voices of two specific subjects that have never interacted could be hypothesized to be fixed under steady conditions. In contrast, after (or during) a verbal interaction, the acoustic features of each of the two speakers could be influenced by the partner, shifting towards a common point or diverging. Hence, in this situation the similarity output should consequently increase or decrease.

4.2.1 Obtaining a measure of voices similarity

Nowadays, the speaker verification is a relatively old and well addressed problem. A recent work (Mukherjee et al., 2017) tried to measure the acoustic convergence using this kind of speech technology. The proposed model was based on Gaussian Mixture Model-Universal Background Model, considered a default model in the field of speaker verification during the past decades (Saeidi, Sadegh Mohammadi & Khalaj-Amirhosseini, 2005; Reynolds, Quatieri & Dunn, 2004). In practice, after having trained a global model (UBM), fitted on all the subjects in the dataset, an adaptation is performed onto a specific subject.

This characteristic represents a crucial weakness. This model can indeed only discriminate if the analysed speech belongs to the subject used for the model adaptation. This peculiarity implies that the model must be readapted every time the speaker verification has to be performed onto the voice of a different speaker.

The aim of our work is hence to create an algorithm, capable to output a distance metric independently of who is speaking, hence, a measure of similarity between each couple of voices.

4.3 Materials and methods

4.3.1 The model architecture: Siamese neural networks

With the aim of obtaining a distance informing how similar two voices are, we designed a speaker verification model based on Siamese neural network architecture (Chicco, 2021). Siamese neural network are a Deep Learning model particularly suited for learning similarity measures (Mueller & Thyagarajan, 2016; Neculoiu, 2016). Practically, a Siamese neural network consists in two twin deep models, that process simultaneously and in the same way two streams of data. The final output of the model is a similarity index between the two inputs.

4.3.2 Recurrent Neural networks to deal with sentence data

As introduced in the previous paragraph, the model we designed is based on the Siamese Neural Network architecture. This specific architecture is capable to deal contemporary with two inputs extrapolating an index of similarity. However, the model has to successfully deal with a very complex, high dimensional, time varying data, such as the temporal features extracted frame-by-frame from the raw speech. Hence, considering the time-varying nature of the inputs, we selected as the basic blocks of the full model architecture the famous Recurrent Neural Networks (RNNs) (Sherstinsky, 2020; Graves, 2013; Tealab, 2018; Rumelhart, Hinton, & Williams 1986). RNNs are a particular kind of network specifically designed to deal with time series. The name of this neural networks comes from the fact that they operate in a recurrent way. This means that the same operation is performed for every element of a sequence, with the current output depending both on the current input and on the previous operations (see Figure 18 for a schematic).

Usually in neural networks all the inputs are independent of each other. In RNN instead, all the inputs are related to each other. For instance, in the situation where

the next word in a given sentence has to be predicted, RNNs exploit the relation among all the previous words to better predict the final output (i.e. the word.

This is achieved by looping the output of the network at time t with the input of the network at time t+1. These loops allow persistence of information from one time step to the next one.

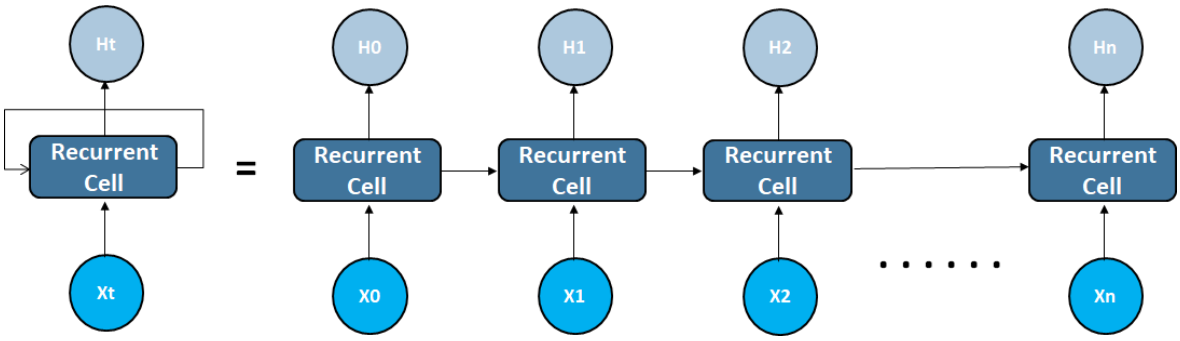

Figure 18. Schematic of the Recurrent Neural Network (RNN). On left is showed the compact representation, on the right the unroll equivalent. The output of the recurrent net at current time H_t depends on the correspondent input X_t as well as the output of the cell itself at the previous time H_{t-1}. Through this mechanism the information extracted by input at previous time persists in the network over time.

RNNs are mainly characterized by the operation performed on the inputs by the base cell. Different kind of RNNs can be listed (Simple Recurrent Units, Gated Recurrent Units, Long Short-Term memory units and others) in respect to this computation.

We performed several experiment using different RNN units, such as the LSTM. In the end, in order to reduce the model complexity the Simple Recurrent Unit has been selected (LSTMs for instance are way more complex than SRUs and usually require an enormously big dataset to be trained properly).

Finally, RNNs can process the inputs both forward and backward over time. This means that in one case (forward) the data sequence is passed to the model starting from the first and ending with the last temporal point, in the other case (backward) instead the computation starts from the end of the sequence and ends

with its first temporal point. This property of the RNNs makes them even more efficient in extracting temporal relations between samples.

4.3.3 Technical description of the Model

After having introduced RNNs and the Siamese architecture in previous paragraphs, I now describe the precise model design and parameters. The Siamese model was built using Tensorflow v2.4.2 library **(**TensorFlow Developers, 2021).

Each of the two twin networks consists in multiple layers. The first layer is a masking layer, implemented using the object Masking contained into the keras.layers library (https://www.tensorflow.org/api_docs/python/tf/keras/layers/Masking). This layer is very important because allows the algorithm to not consider the padded values in the input sequence.

After the masking layer, data are passed to the core of the model. i.e. the recurrent neural networks. The RNN consists of a simple RNN cell processing the features bidirectionally. Specifically, the recurrent neural network has 1 bidirectional layer with 50 hidden neurons included into the recurrent cell. The hyperbolic tangent is used as activation function in order to guarantee a smooth training and force the features to range between -1 to +1. Additionally, on this layer a Lasso regularization (Tibshirani, 1996) is applied to let the network select the most relevant features.

After being processed by the recurrent layer the features were normalized by means of batch normalization (Ioffe & Szegedy, 2015) through the BatchNormalization object of tensorflow (https://www.tensorflow.org/api_docs/python/tf/keras/layers/BatchNormalization).

Subsequently, the output of the last time step of the recurrent layer is fed into a feedforward layer (https://www.tensorflow.org/api_docs/python/tf/keras/layers/Dense) composed by 50 neurons with a sigmoid activation function that forces the output between 0 and

1. Lasso regularization was applied also to the neurons in this layer. At this stage, hence, each of the two twin models outputted a vector of 50 features ranging between 0 and 1.

Finally, in the last layer a simple mathematical operation is computed: the cosine similarity. This function takes two vectors and returns a value equal to the cosine of the angle between them. The value is obtained computing the dot product of the same vectors normalized to both have length equal to 1. Because of the normalization, cosine similarity does not take into account the magnitude of the values considered but only the angle between the two arrays. Results of this operation vary between -1 and 1, with 1 indicating that the two vectors are identical, -1 that they are identical too but with different sign and 0 indicating orthogonality between the two of them. For our model the values of similarity range between 0 and 1. This is due to the fact that the process performed by the sigmoid activation function in the last layer forces the outputs to be positive.

The result of this operation is the final output of the Siamese model and represents the similarity score between the couple of speech streams: the closer to 1 the output the more similar the two voices are. A schematic representation of the model is showed in Figure 19.

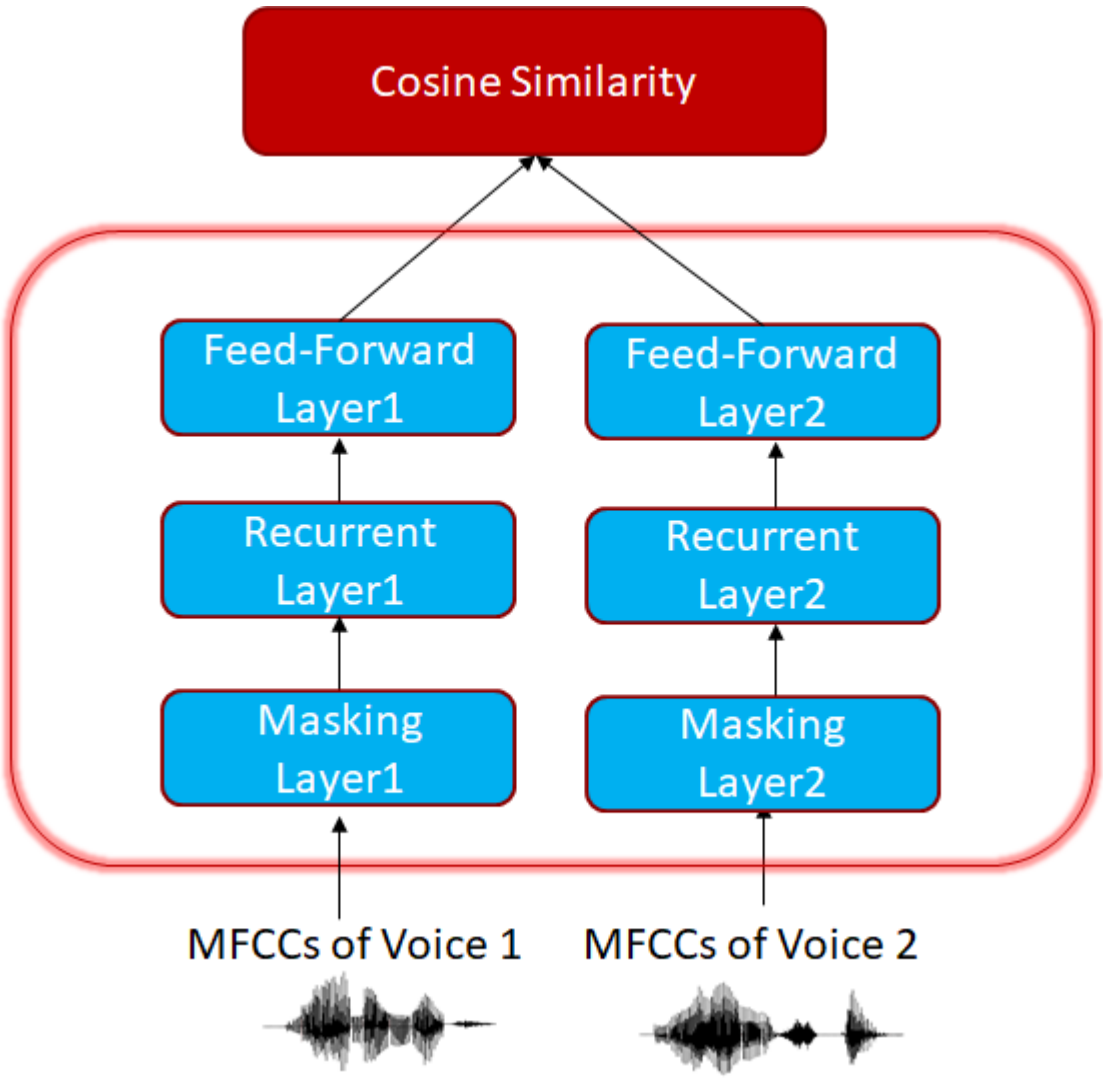

Figure 19. The Siamese Neural Network model. The figure shows a representation of the speaker verification model. It consists in two cascade of layers that equally process the MFCCs of two speech streams. In the last layer the cosine similarity function between the two features' vectors is computed. The result of this final computation represents the measure of distance between the inputted voices.

4.3.4 The VCTK dataset

Performances of deep learning models are known to be highly dependent on the amount of data available for the training. In general, the more the model is complex the more data are needed for a reasonably good training. Additionally, the scientific question that has to be addressed, i.e. learning a similarity measure between voices independently of the sentences pronounced (text-independence)

and of the couple of speaker considered (speaker independence), is an extremely hard to solve problem. The intrinsic variability between the characteristics of the voices, the length of the sentences pronounced and accents is, indeed, brutally high.

For the above mentioned reasons, the dataset used to train the Siamese model required to be sufficiently large in terms of number of sentences, subjects and to cover as many as possible of different accents. To train the model, hence we used an English multi-speaker freely available dataset that seemed the ideal candidate to fit the problem's criticalities: the CSTR's VCTK Corpus (Centre for Speech Technology Voice Cloning Toolkit) (Yamagishi, Veaux & MacDonald, 2012).

Here we report the precise description of the dataset offered by its authors: "CSTR's VCTK Corpus (Centre for Speech Technology Voice Cloning Toolkit) includes speech data uttered by 109 native speakers of English with various accents. Each speaker reads out about 400 sentences, most of which were selected from a newspaper plus the Rainbow Passage and an elicitation paragraph intended to identify the speaker's accent. The newspaper texts were taken from The Herald (Glasgow), with permission from Herald & Times Group. Each speaker reads a different set of the newspaper sentences, where each set was selected using a greedy algorithm designed to maximise the contextual and phonetic coverage."

4.3.5 Domino Task dataset

The VCTK dataset is a big and acoustically rich dataset. It was used to let the model learn a general measure of distance between voices of speakers. The use of the VCTK dataset, indeed, was motivated by its high phonetic and acoustic coverage that allows to include in the Siamese model a wide knowledge about different voices signatures. However, this dataset does not include any verbal interaction between subjects.

For this reason, a second dataset, previously recorded by colleagues at the Italian Institute of Technology and Università degli Studi di Ferrara, was subsequently used. It was recorded in a simplified verbal interaction setting, specifically

designed to force acoustic convergence to be more easily measurable. A total of 8 couples of speakers were involved in a linguistic game called Domino Task, where they had to read aloud words both alone and with a speaking partner. The dataser therefore included two main different situations, one in which subjects interacted with a partner and one where they were reading words without being influenced by the voice of another person.

Here we report the complete description of the dataset reported by authors in their work (Mukherjee et al., 2017**): "** 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.

A verbal domino chain (Bailly & Amélie, 2014) was constructed with English words. To do this, we used the WebCelex (http://celex.mpi.nl/) English lexical database. Disyllabic words were first extracted from the database and then rearranged depending on spoken frequency (Collins Birmingham University International Language Database - COBUILD). A custom algorithm using R (https://github.com/sankar-mukherjee/SPIC-dommino) was then used to build the dominos. The algorithm starts from the highest lexical frequency word and then looks for the next highest frequency word, fulfilling the rhyming criteria and no repetitions. From the list generated, 200 unique bi-syllabic words were selected for the Verbal Domino task.

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.“

4.3.6 Data processing

The words and phrases present into the datasets were processed using Tensorflow 2.42 library and librosa, a python package for audio and music processing (McFee et al., 2015, https://librosa.org/doc/latest/index.html,).

No assumptions have been made about the characteristics of the features to use in order to measure the acoustic speech convergence. Hence, aiming to exploit the full richness of the acoustic spectrum we used Mel-frequency cepstral coefficients (MFCCs) (Davis & Mermelstein, 1980; Aubanel & Nguyen, 2010). MFCCs are standard features used in speech and sound technologies that describe the power spectral envelope of each single frame of an acoustic signal.

The MFCCs were extracted for each frame, with a time window of 25 ms and a time step of 10 ms. Subsequently, the first and second order derivative of the MFCCs were computed and added as features together with standard MFCCs. Computing the first and second order derivatives is also common procedure used to enhance the performances of the systems dealing with acoustic data (Zheng, Zhang & Song, 2001).

As final processing step, data were padded to the length of the sentence with the maximum number of frames. This computation was needed due to the fact that all the phrases and words had different lengths (i.e. different number of frames). Finally, the padded MFCCs and their derivatives were feed into the Siamese neural network, representing the features exploited by the model to output the similarity score.

4.3.7 Adapting the VCTK dataset to the Siamese architecture

As first step, we trained the Siamese neural network onto the above described VCTK dataset.

The whole dataset were split into train and validation set, respectively containing 15% and 85% of the sentences of each subject.

Subsequently, within each of the two sets, the examples used to train the model were created matching two sentences, i.e. creating a couple. The couples were then labelled as 1 if they belonged to the same subject and 0 otherwise. This labelling procedure guided the model to output higher values of similarity (close to 1) when the couple considered contained sentences pronounced by the same speaker, lower values (close to 0) otherwise.

Specifically, the examples for the class labelled as 1 were obtained by taking all the possible combinations between each sentence and all the other pronounced by the same speaker. In the same way examples labelled as 0 were obtained by taking the combinations between each sentence and all the other pronounced by every other speakers into the considered dataset.

The combinations between the two classes were taken without repetition. For instance, considered the couple obtained combining the first and the second sentence of speaker 1 (S_{11} and S_{12}), only the combination { S_{11}, S_{12} } was included, thus the couple { S_{12}, S_{11} } resulted to be excluded.

Due to the presence of an high number of speakers, the number of possible combinations increased exponentially for the class 0. Hence, the total number of examples labelled as 1 resulted to be lower than the 0 class. Hence, in order to avoid to train the model using an highly unbalanced dataset, likely to result in an inefficient training (Masko & Hensman, 2015; Mirus, Stewart & Conradt, 2020), the examples of the training set were down-sampled to be roughly 1.000.000 for each of the two classes.

4.3.8 Training the Siamese model onto the VCTK dataset

After being processed, the VCTK dataset was ready to be used for training the Siamese model. Due to the high number of examples, and the consequent high memory space required to load the whole dataset, the training procedure was performed splitting it in batch of 2048 examples. The model was trained using the binary cross-entropy loss function (https://www.tensorflow.org/api_docs/python/tf/keras/losses/BinaryCrossentropy). This loss is commonly used when dealing with a binary classification problem. This is the case of scientific question where it is needed to determine if two speech streams belong to same speaker or not. The optimization of the training was managed by the Adam optimizer (Adam: A Method for Stochastic Optimization, Kingma & Ba, 2014, https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/Adam) together with the Tensorflow's heuristic (https://www.tensorflow.org/api_docs/python/tf/keras/Model#fit) capable to automatically determine the appropriate learning rate.

The trained model was then tested onto the validation dataset, and the accuracy in correctly classifying the two classes was measured.

Several runs were executed, to find optimal values for the Lasso regularization parameters of the recurrent layer and of the feed-forward layer. This operation required several months due to the high computational cost of each training with a dataset of the above-mentioned dimension.

Finally the model that achieved the best performances was stored, representing the base model to subsequently adapt using the Domino Task dataset, with final aim of measuring the acoustic convergence.

4.3.9 Adapting the Domino dataset to the Siamese architecture

The Domino dataset involves two main different experimental conditions: *"duet"*, i.e. when two speakers are interacting, pronouncing a word after having listened for the partner's voice, and *"solo"*, i.e. when each subject is reading aloud without interacting with the other person.

The *"duet"* condition was identified as the part of the Domino dataset to subsequently use to measure the variations of the similarity between speakers' voices. It was therefore not used to train the Siamese model. Consequently, only the data contained in the *"solo"* part of the dataset were used for training.

As for the VCTK dataset, the processed acoustic data of each subject, i.e. the extracted MFCCs, were divided in training and test. Respectively the 80% of the data were assigned to the training set and 20% to the validation set. After this data fractioning, the words inside the same set were paired and the resulting couples were labelled with 0 and 1 with same criteria adopted for the bigger dataset. Although the split and labelling performed onto the two datasets were the same, there was a main difference in the processing of the two dataset.

When creating the train and validation set for the Domino dataset, indeed, the words belonging to 1 of the 8 couples was completely excluded. The newly generated sets therefore contained the sentences of only 7 of the 8 couples. This procedure was then repeated 8 times, each time leaving a couple out of both the train and validation set. The examples of the left out couple were created as in the other cases. i.e. pairing the words pronounced by the same speaker for class 1 and the ones pronounced by different speakers for class 0.

Finally, the obtained couples were considered an additional set of examples, named left out couple set (LOC set). Repeating the splitting operation as many times as the number of couples produced therefore the same number of train, validation and LOC set.

The rationale behind this particular splitting procedure was to have two possible levels of complexity to test the Siamese model. The first level is the sentence independence, achieved when the model performs well on new sentences

pronounced by couples of subjects already *"known"* by the model. The second level, and the hardest to achieve for a speech technology (Furui, 1997) is the speaker independence, achieved when the model performs well even if the inputs came from subjects whose voices were never included in the training data.

4.3.10 Training the Siamese model onto the Domino dataset

The best model trained using the VCTK dataset was now retrieved and retrained. The training at this stage was needed to adapt the parameters of the pre-trained model to the new data. The Domino dataset, indeed, involved different subjects (therefore different voices) and consisted in words rather than sentences (therefore smaller data lengths).

As previously explained, the pre-training with the VCTK dataset was important to provide the model with a basic knowledge about the distance between several different voices. This step was crucial because of the relatively small dimension of the Domino dataset. Indeed, training from scratch a deep learning model, as the Siamese network, using only a small amount of data, as the ones included in this latter dataset, could in principle be very difficult and ineffective.

Each of the 8 training dataset obtained by the left one couple out splitting procedure was used to train a Siamese model. In this way, as many models as the number of couples were produced. The loss function, the optimizer and the heuristics to determine the learning rate were the same of that selected for pre-training.

4.4 Results

4.4.1 Performances

Each of the 8 models was first tested using its correspondent validation set in order to assess the performance in predicting the voices' similarity when dealing with unseen sentences.

The average accuracy for this task resulted to be 81.4 +- 0.8%. Considering only the negative examples (i.e. the ones obtained matching words pronounced by different speakers) and positive examples (i.e. the ones obtained matching words pronounced by the same speaker) the average accuracies resulted in 84.8 +- 1.8% and 81.4 +- 0.8%.

Subsequently, each model was tested onto the correspondent LOC set to assess its performances when dealing with completely unseen subjects. The average accuracy for the speaker independence test resulted to be 61.1 +-2.2%. Considering only the negative and positive examples the average accuracies resulted 76.9.4 +- 4.2% and 45.4 +- 3.0%. The accuracy for each of the 8 validation and LOC set is reported in Table 3 for both positive and negative examples.

The histograms of the distribution of the similarity values, computed over the 8 possible dataset splits for both the 0 and 1 classes, are showed for the validation set in Figure 20 and for the LOC set in Figure 21.

	Couple1	Couple2	Couple3	Couple4	Couple5	Couple6	Couple7	Couple8
Validation	80.5	80.7	80.2	79.5	83.0	85.1	84.0	78.3
Validation negative	77.1	85.3	88.6	87.2	86.0	87.5	91.7	75.3
Validation positive	83.8	76.0	71.9	71.9	80.1	82.7	76.1	81.4
LOC	52.3	65.1	57.0	56.3	60.5	57.1	70.0	70.9
LOC negative	52.5	85.1	76.2	71.4	83.0	74.3	96.2	76.8
LOC positive	52.0	45.0	37.7	41.2	38.0	40.0	43.9	65.0

Table 3. Classification performance achieved by the Siamese model on each of the eight couples of speakers. Results are showed for validation set and left-on- couple-out set, testing all the examples, or testing only the negative or positive examples.

4.4.2 Sentence independence

From the histograms it is possible to appreciate the capability of the Siamese model to output optimal values of similarity, for both classes, when the validation set is considered. The model, hence, even in presence of *"never heard"* words, successfully learned how to assign correct values of similarity to the voices' couples. This result implies that the model reached the sentence independence, consisting in the first of the two goals to achieve before testing the model onto the *"duet"* dataset to measure the acoustic speech convergence.

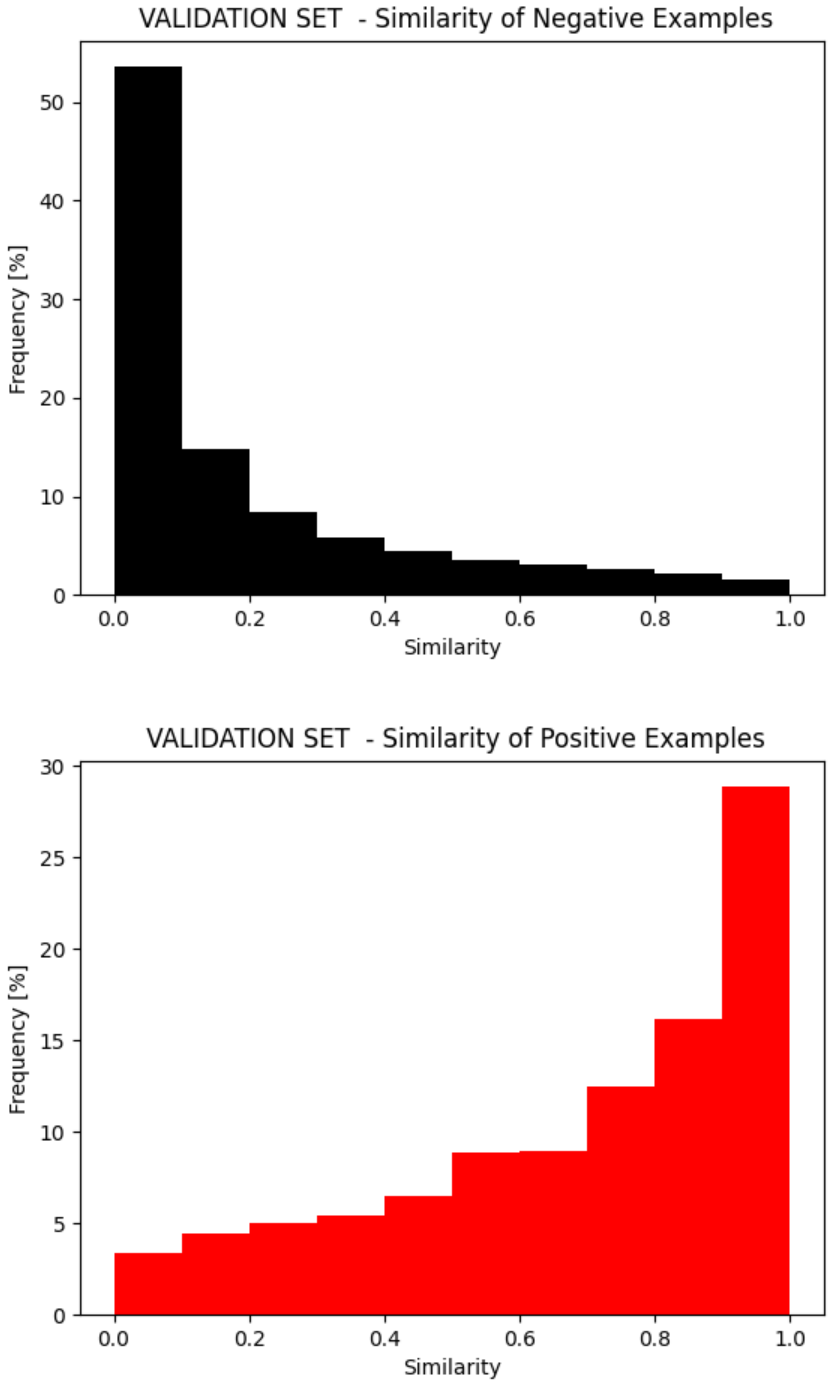

Figure 20. Histograms of similarity values outputted by the Siamese network when testing it on validation set couples. The black histogram is for negative examples (i.e. couples of voices obtained from different speakers), the red histogram is for negative examples (i.e. couples of voices obtained from the same speaker).

4.4.3 Speaker independence

As regard the left out couple set, it is possible to derive from the histograms (see Figure 21) and from the classification performances (see Table 3) that the model is capable to assign correct values of similarity and perform correct predictions when the words are pronounced by different speakers but not when the speaker is the same. In the latter case, indeed, we should have values close to one for most of words' couples, implying a strongly *"right-skewed"* histogram. The behaviour instead is the opposite one with a pronounced left skew similar to the one of negative class. The model, therefore, is not fully capable to generalize to unseen speakers' voices, especially when processing acoustic signals belonging to the same person.

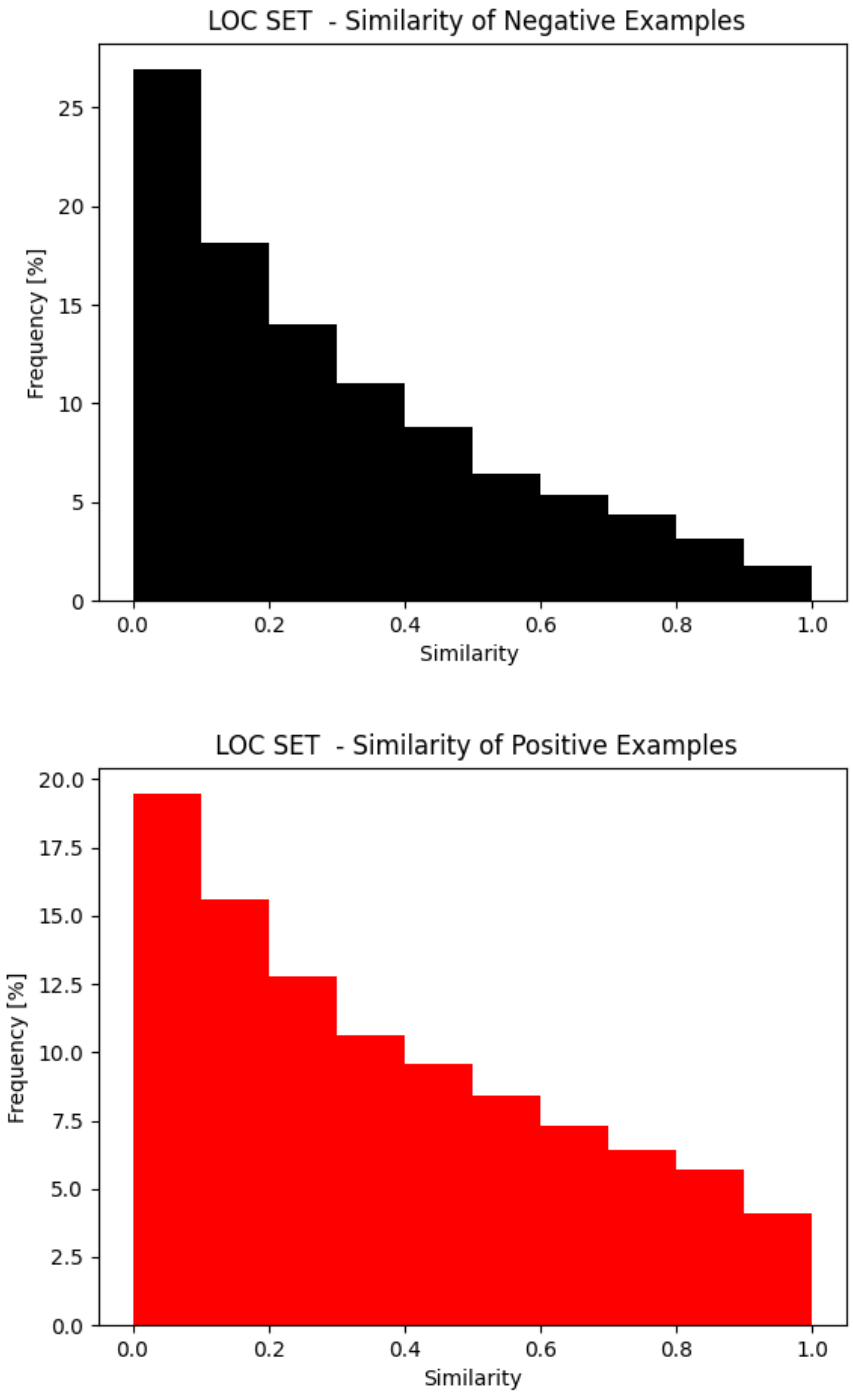

Figure 21. Histograms of similarity values outputted by the Siamese network when testing it on LOC set couples. The black histogram is for negative examples (i.e. couples of voices obtained from different speakers), the red histogram is for negative examples (i.e. couples of voices obtained from the same speaker).

4.5 Conclusions and further steps

We designed a Deep Learning model based on Siamese Neural Network architecture and Recurrent Neural Networks in order to address the scientific problem of measuring the acoustic speech convergence. The rationale behind was to merge the capability of the Siamese network to process couples of speech stream obtaining a measure of similarity and the computational power provided by the RNNs when dealing with time series.

The final goal to achieve through the proposed model is to obtain a mathematical powerful tool capable to compute the similarity between speakers' voices independently of the speakers that compose the interacting couple and of speech pronounced during the verbal interaction. These goals can be translated as speaker independence and sentence independence. The first one in particular is an hard to solve problem because of the tremendous amount of inter-speakers variability (Kenny et al., 2007). Nonetheless, it is fundamental that both of them are achieved before considering the model completely reliable in measuring the acoustic speech convergence in unknown speaker and unknown sentence scenario.

The model that we designed was pre-trained using first a very large dataset, the VCTK's Corpus, used to provide the model with knowledge about a wide variety of inter-speakers variability: it involves indeed more than one hundred English speakers with different accents. Subsequently the model was retrained on the Domino task dataset, a smaller dataset specifically designed to create a constraint verbal interaction where measuring the acoustic convergence is easier. This dataset contains words instead of sentence and Italian speakers speaking in English.

The Siamese model pretty well performed in assigning similarity values to the couples of voices when tested in a sentence independent scenario. In the other hand, when facing with the speaker independence problem the model showed difficulties in correctly computing the distances between voices.

This difficulty of the model has to be fixed before testing it on the *"duet"* part of the domino task dataset. The final aim is indeed obtain values of similarity that are reliable in a verbal interaction that involve every generic couple of speakers pronouncing every generic set of sentences, thus characterizing the Siamese model as a fully deliverable mathematical tool.

The problem faced by the model could be due to the fact that the speakers involved into the Domino dataset are native Italian speakers speaking in English, hence the pre-training with the VCTK Corpus did not provide the model with knowledge about this specific English accent. Additionally the different length of the time series processed by the model when moving from sentences (the VCTK corpus) to words (the Domino task dataset) could have represented an hard-to-overcome obstacle.

Finally, we plan to improve the performances of the Siamese model by adding the Italian accent speakers into the pre-training and randomly segment the long sentences, creating a mixed dataset composed of both sentences and words.

After having tuned the model to deal with every couple of speakers and every couple of sentences we will test the dynamics of the acoustic the acoustic speech convergence during the verbal interaction contained in the Domino dataset to gain insights about the acoustic level of this complex multimodal phenomenon.

4.6 References

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

Bailly, G., & Martin, A. (2014, September). Assessing objective characterizations of phonetic convergence. In *Interspeech 2014-15th Annual Conference of the International Speech Communication Association* (pp. P-19).

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, *28*(4), 357-366.

Furui, S. (1997). Recent advances in speaker recognition. *Pattern recognition letters*, *18*(9), 859-872.

Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 6645-6649). Ieee.

Ioffe, S., & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448-456). PMLR.

Jansson, P. A. (1991). Neural networks: An overview. *Analytical chemistry*, *63*(6), 357A-362A.

Kenny, P., Boulianne, G., Ouellet, P., & Dumouchel, P. (2007). Speaker and session variability in GMM-based speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(4), 1448-1460.

Masko, D., & Hensman, P. (2015). The impact of imbalanced training data for convolutional neural networks.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015, July). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference* (Vol. 8, pp. 18-25).

Mirus, F., Stewart, T. C., & Conradt, J. (2020, July). The importance of balanced data sets: Analyzing a vehicle trajectory prediction model based on neural networks and distributed representations. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

Mueller, J., & Thyagarajan, A. (2016, March). Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 30, No. 1).

Mukherjee, S., d'Ausilio, A., Nguyen, N., Fadiga, L., & Badino, L. (2017, August). The relationship between F0 synchrony and speech convergence in dyadic interaction. In *Interspeech 2017* (pp. 2341-2345).

Neculoiu, P., Versteegh, M., & Rotaru, M. (2016, August). Learning text similarity with siamese recurrent networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP* (pp. 148-157).

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital signal processing*, *10*(1-3), 19-41.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533-536.

Saeidi, R., Mohammadi, H. S., & Amirhosseini, M. K. (2005, September). Efficient GMM-UBM system in text independent speaker verification using structural Gaussian mixture models,". In *Proc. International Symp. of Telecommunications, IST* (Vol. 1, pp. 39-44).

Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*, 132306.

Tealab, A. (2018). Time series forecasting using artificial neural networks methodologies: A systematic review. *Future Computing and Informatics Journal*, *3*(2), 334-340.

TensorFlow Developers. (2021). TensorFlow (v2.4.2). Zenodo. https://doi.org/10.5281/zenodo.4960227

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.

Yamagishi, J., Veaux, C., & MacDonald, K. (2019). Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92).

Zheng, F., Zhang, G., & Song, Z. (2001). Comparison of different implementations of MFCC. *Journal of Computer science and Technology*, *16*(6), 582-589.

5 THESIS DISCUSSION

Human speech is a tremendously sophisticated physiological phenomenon, through which we are capable of delivering complex contents in a very effective way. The neural basis of speech still remains not entirely explained. It consists of cascades of neural mechanisms which give rise to the articulation of words and phrases, as well as to the ability of perceiving and understanding the listened sounds with a final goal: communication. During my work I touched upon different speech research areas, in the attempt to shed light both on the neural processes upon which speech production and perception are grounded, and on the dynamics underlying verbal interaction.

Firstly, I focused on the speech production research field. In this context, the investigation was conducted exploiting a rare speech-related neurophysiological signal, by means of electrocorticography (ECoG) recorded from a salient speech region – known as speech arrest area – located in Broca's area. Electrocorticographic recordings were obtained using dense micro (μ)ECoG arrays placed over the speech arrest area of patients undergoing awake brain surgery.

Thanks to the functional relevance of the recording locus, neural signals synchronised with the patients' voice were used to investigate speech production mechanisms. Interestingly, a burst of high-gamma power activity was present few milliseconds before patients' speech onset. Being this anticipatory neural trigger highly consistent across trials, it was selected as input to train a Support Vector Machine (SVM) based machine learning model, in the attempt to predict speech initiation.

The model was trained using the neural activity recorded from two patients, employing three very diverse recording matrices. This difference between devices, as well as patients, introduced a not negligible level of data variability. However, the high temporal and spatial reliability of this neural correlate was of fundamental importance, leading to a successful prediction of patients' speech onset.

Notably, speech onset prediction was possible even when the model ran in a cross-subjects and cross-devices modality, i.e. when it was tested on the neural

activity recorded with a specific device from a specific patient, different from the one used to train the model.

Speech onset detection systems might represent an important building block in future speech neuro-prostheses (Martin et al. 2018) for patients with severe communications deficits, as happens for example in the Locked-in syndrome. In this clinical population, language areas are often intact (Smith & Delargy, 2005) but patients lose the capability to verbally communicate. Hence, in such a scenario, speech-related neural activity must be decoded by speech-brain computer interfaces (speech BCI) (Anumanchipalli, Chartier, & Chang, 2019; Angrick et al., Moses et al., 2016; Martin et al. 2014) in order to translate brain activity into words and phrases.

For efficiency optimization, a speech decoding procedure should nonetheless start from an explicit language preparation cue. Here is exactly where speech onset detection systems come into play, by identifying a neural trigger whenever the patient intends to speak, thus initiating speech decoding.

However, reported results were so far obtained by exploiting neural activity recorded during overt speech tasks. In contrast, a real life clinical situation implies the need of decoding covert – i.e. imagined – speech (Martin et al. 2016), because of the patients' communication impairment. This could potentially represent a high-impact limitation for the proposed algorithm: Indeed, this operation is not guaranteed for the model when facing such a different condition and, as a consequence, further investigation has to be performed. Nevertheless, the good performances achieved by the model when predicting speech onset in the cross-patient modality suggests the possibility for the model correct functionality even in a very different situation, such as decoding covert speech onset.

After performance assessment in real clinical scenarios, the proposed model could finally be included in a speech-BCI to detect patients' intention to speak, restoring their connection towards the outside world.

Nevertheless, the connection in the opposite direction is fully functional: patients can indeed continue to perceive the incoming speech as every other healthy subject does.

How the human brain perceives and understands speech, however, is still an extensively debated argument. Without the aid of very complex mechanisms to process and interpret it, words and meanings that people intend to communicate to other, , would in fact only be unintelligible sounds.

It was long believed that only the auditory cortices – especially Wernicke's area – participated in speech perception, excluding any involvement of the motor regions, whose role was relegated to sending motor commands to mouth articulators during speech production (Wernicke, Cohen & Wartofsky, 1874).

Anyhow, the idea that the motor system plays an active role when perceiving speech gained increasing attention in the last decades, with several works providing new evidences in support of this hypothesis (Watkins, Strafella & Paus, 2003; Wilson et al., 2004, Iacoboni, 2008, Pulvermüller et al., 2006, Fadiga L. 2002, D'Ausilio et al., 2009 , Meister et al., 2007, Möttönen, Dutton & Watkins 2013; Assaneo & Poeppel, 2018). On the other hand, the precise way in which the motor system could act in order to improve speech comprehension is unclear.

In this context, a major perspective is that the motor system of a listening brain simulates the speaker's mouth articulators kinematics (Morillon et al., 2019; Arnal & Giraud 2012; Schubotz, 2007), hence the movements needed to produce the listened words and sentences.

During my research in the field of speech perception, I designed an experiment to investigate this hypothesis, in the attempt of providing new hints about the reconstructive nature of the mechanisms played by the motor system in speech perception. The idea consisted in providing listeners with acoustic information without any speech-related visual or kinematic cues, to subsequently address if kinematics was anyhow encoded in the listening brain.

To this end, a public dataset (Canevari, Badino & Fadiga, 2015), containing speech sentences synchronized with the correspondent mouth kinematics, was selected. However, the only stimuli administered during the experiment were the acoustic ones. 23 healthy subjects listened to the sentences while their brain activity was recorded by means of electroencephalography (EEG).

The presence in the brain of a speech tracking mechanism during listening is well documented and known as speech neural entrainment. This phenomenon is traditionally described as a sort of phase alignment between the auditory cortices neural signals and the low frequency components of the attended speech stream. Entrainment is triggered by speech landmarks – such as phonemes, words and sentences – and has been proved to correlate with higher speech comprehension performances (Ahissar et al., 2001; Luo & Poeppel, 2007; Peelle et. al. 2013).

This neural mechanism has been traditionally measured using phase coherence or correlation methods (Luo & Poeppel 2007; Morillon et al. 2010, Peelle, Gross, & Davis, 2013; Bourguignon et al., 2013), capable to reveal linear trends of phase alignment between time series. Nevertheless, in order to reveal more complex and non-linear relations between speech and brain recordings, mutual information (MI; Shannon, 1948) has recently increasingly replaced previous techniques (Gross et al. 2013, Kayser et al. 2015; Giordano et al. 2017). For what specifically concerns a listening scenario, MI between neural signals and the attended stimulus returns a measure of the amount of information encoded by the brain or, in the other way around, how more predictable neural signals become when knowing the administered input.

When multiple stimuli are present, however, neither MI nor coherence methods can disentangle how different input modalities contribute to the observed entrainment phenomena. In this context, the ideal candidate is represented by the Partial Information Decomposition (PID; Williams & Beer 2010; Ince, 2017), a recent MI-based mathematical framework. Indeed, PID allows to deal with multiple inputs simultaneously and disentangle their informative contribute encoded by the brain in “atoms” of Unique, Redundant and Synergistic information.

Hence, the use of this powerful mathematical tool together with the speaker’s mouth kinematics synchronized with the speech offered a great possibility to investigate the active listening theory. Indeed, finding brain encodings of the speaker’s speech articulators movements would suggest the presence of the neural tracking of the simulated mouth kinematics. Importantly, employing PID in this context could inform about the non-redundancy of these neural mechanisms with respect to entrainment to the acoustic input modality.

However, how the brain controls mouth articulators – i.e. which of them are moved together while speaking – is unknown. For this reason, Principal Component Analysis (PCA) was applied transforming the raw time-varying articulators positions in new physiologically meaningful time series. By means of this data-driven approach, four main principal components were obtained. Specifically, the first two represented the antero-posterior and the vertical movement of the whole tongue, whereas the other two consisted in mouth articulators synergies.

Finally, by coupling such principal components with the speech envelope, the Partial Information Decomposition was applied to find atoms of unique kinematic information encoded in the EEG signals.

PID analysis revealed the presence of “unique” information about the tongue antero-posterior movement encoded in temporal and motor areas. This result is of crucial importance, because it reflects the fact that kinematic information is encoded in motor areas and that this phenomenon is completely different when compared to the speech envelope tracking.

Evidence that the kinematics of the mouth articulators is tracked by the brain has already been observed (Giordano et al., 2017; O’Sullivan et al., 2021; Park et al., 2016; Peelle & Sommers, 2015; Giordano et al., 2017), as well as the fact that motor areas are active during speech perception (Ding et al.,2017; Keitel , Gross & Kayser, 2018). To my knowledge, it is however the first time that motor areas are shown to perform neural tracking of a not readily available speech related signal. Indeed, the fact that subjects’ neural activity resulted to track kinematic stimuli even if not administered suggests the presence of an ongoing simulation of the speaker’s mouth kinematics.

Interestingly, PID analysis unveiled also a synergistic interaction between the speech envelope and two kinematic principal components: this suggests that kinematics signals not only are independently exploited by the listening brain, but could also be integrated with the acoustic information in a different processing stream.

Notably, the two different kind of encoded kinematic information, i.e. unique and synergistic with the speech envelope, showed a frequency dissociation highlighted

by the subsequent frequency-resolved PID analysis. Indeed, unique information resulted to be maximally encoded over low frequencies (1-4 Hz, corresponding to the delta frequency band), in contrast to the synergistic modality showing a clear peak over higher frequency (4-8 Hz, i.e. in the theta band). Moreover, the maximal encoding range for the synergistic information matched that of the speech envelope unique information, by means of the neural entrainment to the pure acoustic input modality.

The link between speech entrainment and comprehension performances is well documented. However only in a recent work (Keitel, Gross & Kayser, 2018) comprehension performances have been related to entrainment arising over motor areas. This represents a fundamental result, because it reveals the importance of the motor system for speech comprehension.

In my research, I provided new evidence for the activation of motor areas during speech perception, highlighting the neural tracking of unavailable – and thus likely being simulated – speech-related kinematics. However, the relation between comprehension performances and the two (unique and synergistic) kinematic atoms of information described has not been analysed.

This investigation represents the natural prosecution of the findings previously described in the speech perception field. Indeed, comparing the strength of the observed phenomena would provide an important contribution to unveil the mechanisms underlying the activation of the motor system during listening, supporting, or eventually rejecting, the thesis that frames neural entrainment as a motor-driven top-down mechanism.

To close the circle of my personal investigation concerning speech, I finally focused on the field that merges together speech production and perception, i.e. verbal interaction between speakers. This fundamental kind of human interaction is used to effectively communicate very complex contents.

During a conversation, speakers share the goals of conveying and understanding concepts and ideas. Indeed, dialog is considered a joint action that exploits intentional and unintentional mechanisms arising at different levels (lexical, kinematic, acoustic). Such phenomenon is known as alignment or convergence,

and is implemented during a conversation in order to enhance the capability of conveying messages (Pickering & Garrod, 2004, 2006; Shockley, Santana, & Fowler, 2003; Hatfield, Cacioppo, & Rapson, 1993; Brennan & Clark, 1996; Holler & Wilkin, 2011).

During my work I focused on the acoustic speech convergence, i.e. on the specific convergence level characterized by the modification of speakers' speech features. During verbal interaction these features (such as speech rate, volume, pitch, etc.) are shifted towards a common acoustic point, making the conversation more effective.

Nevertheless, acoustic convergence lacks of a precise mathematical quantification and is still widely discussed in several aspects, ranging from its origin to its temporal dynamics. Indeed, the previous idea that this mechanism was static during the conversation has been overcome by several studies capturing its time-varying nature (De Looze et al., 2014; Levitan & Hirschberg, 2011; Vaughan, 2011).

Recently researchers started to address acoustic speech convergence measurement via machine learning models (Mukherjee et al., 2017; Ostrand & Chodroff, 2021), trying more comprehensive, data-driven approaches.

Following this intuition, a deep learning-based speaker verification system is here proposed. The model was built combining recurrent neural networks (RNNs) with Siamese architecture. This specific design is thought to merge the RNNs high performances in processing time series (as the speech stream is) with the capability of Siamese architecture to learn a distance measure between couples of inputs (i.e. voices). A model producing such a measure would indeed represent a ready-to-use mathematical tool, capable to inform about the shift of characteristics in the speakers' voice.

The Siamese model was pre-trained using a very large dataset, the VCTK (Yamagishi Veaux & MacDonald, 2019), containing sentences pronounced by hundreds of speakers with different English accents. This step was implemented in order to provide the model with knowledge about very different voices, a key point for each speaker verification system.

The model was subsequently retrained on the *"solo"* block of the Domino task dataset (Mukherjee et al.,2017), with a leave-one(couple of speakers)-out approach. In this way, it was possible to measure the model performances in both a sentence- (using unseen couples of sentences pronounced by known-by-the-model subjects) and a speaker- (testing the left out speakers) independent scenario.

On the one hand, the Siamese model performed successfully when sentence independence was tested, correctly outputting similarity values when couples of voices of both different speakers as well as the same speaker were considered.

On the other hand, the model did not reach the full speaker independence. Performances computed on unseen speakers were indeed reasonably good for couples of sentences pronounced by different speakers, but the model misclassified examples when sentences were pronounced by the same speaker.

The issue faced by the model is likely dependent on the difference in both accents and speech lengths between the two datasets. The domino dataset, indeed, is composed of English words pronounced by non-native (Italian) speakers. However, this particular accent is not included in the variety of English accents contained in the VCTK dataset. Additionally, the latter dataset contains long sentences, as opposite to the simple words present in the domino one. Both these two differences in data characteristics likely compromised the model performances when dealing with the hardest situation, i.e. couples of voices of unseen speakers.

Nonetheless, the domino dataset was selected for its very important feature of containing recording blocks of constrained verbal interaction between speakers. These recordings, indeed, represent a perfect context to measure acoustic speech convergence easily. On the other hand, the importance of the VCTK dataset consists in its huge dimensions and its variety of English accents. Hence, both datasets represent important blocks to build the entire sentence- and speaker-independent model and subsequently use it to measure convergence in a verbal interaction.

In order to solve the speaker independence problem faced by the Siamese model, two approaches can be followed. Either new speakers with Italian accent together

with shorter sentences (or even words) could be added during pre-training or, in the other way around, a different dataset substituting the domino one might be selected. The new dataset should include speakers with English accents similar to those present in the VCTK dataset, as well as speech streams representing sentences instead of words.

Finally, the proposed Siamese model could be employed as a ready-to-use tool in several different verbal interaction contexts which require monitoring of the distance between voices, by means of acoustic speech convergence. Importantly, the possibility to exploit the model in a variety of situations pave the way for its application beyond research, i.e. for commercial purposes.

To cite a practical model application, a second language learning scenario can be considered. Indeed, the acoustic convergence is documented in situations where a non-native speaker is learning a second language. In addition, it is reported that the strength of the phenomenon is higher in talented speaker (Lewandowski & Jilka, 2019). Hence, the system for measuring the speech convergence would represent an effective tool to constantly monitor students' improvement and provide a reliable feedback to refine the learning process.

Another practical scenario where measuring the acoustic convergence can be fundamental is represented by the patient-doctor interaction. Verbal communication skills are, indeed, of crucial importance in this context, and poor doctors' communications capabilities could result in patients' unhappiness and complaints (Kee et al., 2018) and even affect health outcomes (Lou et al. 2022). Therefore, disposing of such a tool for measuring the modification of the patient's voice during a medical consult appears of high relevance for clinicians. The doctor could indeed obtain an on-line feedback, and in case adjust the communication approach with patients, in order to make them feel more comfortable, hence improving their mental and physical wellness.

Beyond these few examples of concrete applications of the proposed system, the list of possibilities can however be enormously enlarged. In addition to its scientific relevance in the speech research field focused on unveiling mechanisms underlying verbal interaction, this gives to the model a significant commercial appeal.

As final remark on the complexity of a verbal interaction, it is worth noticing that even if we perceive it just as the simple, smooth concatenation of production and perception, the dynamics behind speech interaction are much more sophisticated. A proof of this complexity is represented by the rapid turn-taking during interactive talking. Indeed, the sum of production's and comprehension's lags is roughly 600 ms, a very long time compared to the 200ms required by speakers at conversational turn-taking (Lenvinson & Torreira, 2015). Hence, predictive motor-based phenomena are likely shaping verbal interactions in order to anticipate what the interlocutor is going to say, thus lowering the required time for mutual understanding. We indeed neither learn how to produce speech to speak by ourselves nor we learn to perceive it to simply listen. We rather learn how to produce and process human speech to mutually exchange information, thus communicate.

5.1 References

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

Angrick, M., Herff, C., Mugler, E., Tate, M. C., Slutzky, M. W., Krusienski, D. J., & Schultz, T. (2019). Speech synthesis from ECoG using densely connected 3D convolutional neural networks. *Journal of neural engineering*, *16*(3), 036019.

Anumanchipalli, G. K., Chartier, J., & Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. *Nature*, *568*(7753), 493-498.

Arnal, L. H., & Giraud, A. L. (2012). Cortical oscillations and sensory predictions. *Trends in cognitive sciences*, *16*(7), 390-398.

Assaneo, M. F., & Poeppel, D. (2018). The coupling between auditory and motor cortices is rate-restricted: Evidence for an intrinsic speech-motor rhythm. Science advances, 4(2), eaao3842.

Bourguignon, M., De Tiege, X., De Beeck, M. O., Ligot, N., Paquier, P., Van Bogaert, P., ... & 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.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, *22*(6), 1482.

Canevari, C., Badino, L., & Fadiga, L. (2015). A new Italian dataset of parallel acoustic and articulatory data. In *Sixteenth Annual Conference of the International Speech Communication Association*.

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

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

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *European journal of Neuroscience*, *15*(2), 399-402.

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

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biology*, *11*(12), e1001752. doi: 10.1371/journal.pbio.1001752

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1993). Emotional contagion. *Current directions in psychological science*, *2*(3), 96-100.

Holler, J., & Wilkin, K. (2011). Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of Nonverbal Behavior*, *35*(2), 133-153.

Iacoboni, M. (2008). The role of premotor cortex in speech perception: evidence from fMRI and rTMS. *Journal of Physiology-Paris*, *102*(1-3), 31-34.

Ince, Robin AA. "The Partial Entropy Decomposition: Decomposing multivariate entropy and mutual information via pointwise common surprisal." *arXiv preprint arXiv:1702.01591* (2017).

Kayser, S. J., Ince, R. A., Gross, J., & Kayser, C. (2015). Irregular speech rate dissociates auditory cortical entrainment, evoked responses, and frontal alpha. *Journal of Neuroscience*, *35*(44), 14691-14701.

Kee, J. W., Khoo, H. S., Lim, I., & Koh, M. Y. (2018). Communication skills in patient-doctor interactions: learning from patient complaints. *Health Professions Education*, *4*(2), 97-106.

Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLoS biology*, *16*(3), e2004473.

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, *6*, 731.

Levitan, R., & Hirschberg, J. B. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions.

Lewandowski, N., & Jilka, M. (2019). Phonetic convergence, language talent, personality and attention. *Frontiers in Communication*, *4*, 18.

Lou, Z., Vivas-Valencia, C., Shields, C. G., & Kong, N. (2022). Examining how physician factors influence patient satisfaction during clinical consultations about cancer prognosis and pain. *PEC Innovation*, 100017.

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

Martin, S., Brunner, P., Holdgraf, C., Heinze, H. J., Crone, N. E., Rieger, J., ... & Pasley, B. N. (2014). Decoding spectrotemporal features of overt and covert speech from the human cortex. *Frontiers in neuroengineering*, *7*, 14.

Martin, S., Brunner, P., Iturrate, I., Millán, J. D. R., Schalk, G., Knight, R. T., & Pasley, B. N. (2016). Word pair classification during imagined speech using direct brain recordings. *Scientific reports*, *6*(1), 1-12.

Martin, S., Iturrate, I., Millán, J. D. R., Knight, R. T., & Pasley, B. N. (2018). Decoding inner speech using electrocorticography: Progress and challenges toward a speech prosthesis. *Frontiers in neuroscience*, *12*, 422.

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

Morillon, B., Lehongre, K., Frackowiak, R. S., Ducorps, A., Kleinschmidt, A., Poeppel, D., & Giraud, A. L. (2010). Neurophysiological origin of human brain asymmetry for speech and language. *Proceedings of the National Academy of Sciences*, *107*(43), 18688-18693.

Moses, D. A., Mesgarani, N., Leonard, M. K., & Chang, E. F. (2016). Neural speech recognition: continuous phoneme decoding using spatiotemporal representations of human cortical activity. *Journal of neural engineering*, *13*(5), 056004.

Möttönen, R., Dutton, R., & Watkins, K. E. (2013). Auditory-motor processing of speech sounds. *Cerebral Cortex*, *23*(5), 1190-1197.

Mukherjee, S., d'Ausilio, A., Nguyen, N., Fadiga, L., & Badino, L. (2017, August). The relationship between F0 synchrony and speech convergence in dyadic interaction. In *Interspeech 2017* (pp. 2341-2345).

Ostrand, R., & Chodroff, E. (2021). It's alignment all the way down, but not all the way up: Speakers align on some features but not others within a dialogue. *Journal of phonetics*, *88*, 101074.

O'Sullivan, A. E., Crosse, M. J., Di Liberto, G. M., de Cheveigné, A., & Lalor, E. C. (2021). Neurophysiological indices of audiovisual speech processing reveal a hierarchy of multisensory integration effects. *Journal of Neuroscience*, *41*(23), 4991-5003.

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

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169-181.

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral cortex*, *23*(6), 1378-1387.

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

Pickering, M. J., & Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation*, *4*(2), 203-228.

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

Schubotz, R. I. (2007). Prediction of external events with our motor system: towards a new framework. *Trends in cognitive sciences*, *11*(5), 211-218.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379-423.

Shockley, K., Santana, M. V., & Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(2), 326.

Smith, E., & Delargy, M. (2005). Locked-in syndrome. *Bmj*, *330*(7488), 406-409.

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

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

Wernicke, C., Cohen, R. S., & Wartofsky, M. W. (1874). Boston studies in the philosophy of science. *The symptom complex of aphasia: A psychological study on an anatomical basis, D. Reichel, Dordrecht*, *1969*, 34-97.

Williams, P. L., & Beer, R. D. (2010). Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature neuroscience*, *7*(7), 701-702.

Yamagishi, J., Veaux, C., & MacDonald, K. (2019). Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92).