Winter 2018

Characterization of Language Cortex Activity During Speech Production and Perception

Hassan Baker
Old Dominion University, hassanbakersaleh@gmail.com

CHARACTERIZATION OF LANGUAGE CORTEX ACTIVITY DURING SPEECH PRODUCTION AND PERCEPTION

by

Hassan Baker
B.S. June 2015, Birzeit University, Palestine

A Thesis Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

ELECTRICAL AND COMPUTER ENGINEERING

OLD DOMINION UNIVERSITY
December 2018

Approved by:

Dean Krusienski (Director)

Christian Zemlin (Member)

Anastasia Raymer (Member)

Jiang Li (Member)

ABSTRACT

CHARACTERIZATION OF LANGUAGE CORTEX ACTIVITY DURING SPEECH PRODUCTION AND PERCEPTION

Hassan Baker
Old Dominion University, 2018
Director: Dr. Dean Krusienski

Millions of people around the world suffer from severe neuromuscular disorders such as spinal cord injury, cerebral palsy, amyotrophic lateral sclerosis (ALS), and others. Many of these individuals cannot perform daily tasks without assistance and depend on caregivers, which adversely impacts their quality of life. A Brain-Computer Interface (BCI) is technology that aims to give these people the ability to interact with their environment and communicate with the outside world. Many recent studies have attempted to decode spoken and imagined speech directly from brain signals toward the development of a natural-speech BCI. However, the current progress has not reached practical application. An approach to improve the performance of this technology is to better understand the underlying speech processes in the brain for further optimization of existing models. In order to extend research in this direction, this thesis aims to characterize and decode the auditory and articulatory features from the motor cortex using the electrocorticogram (ECoG). Consonants were chosen as auditory representations, and both places of articulation and manners of articulation were chosen as articulatory representations. The auditory and articulatory representations were decoded at different time lags with respect to the speech onset to determine optimal temporal decoding parameters. In addition, this work explores the role of the temporal lobe during speech production directly from ECoG signals. A novel decoding model using temporal lobe activity was developed to predict a spectral representation of the speech envelope during speech production. This new knowledge may be used to enhance existing speech-based BCI systems, which will offer a more natural communication modality. In addition, the work contributes to the field of speech neurophysiology by providing a better understanding of speech processes in the brain.

ACKNOWLEDGEMENTS

My deepest appreciation goes to my supervisor Dean Krusienski. You taught me how to express ideas and how to approach research problems. Thank you for your patience and answering my countless questions. In addition to your support, insightful comments and encouragement, I will never forget the opportunity you provided to me to expand my knowledge in the BCI field and build my research skills. Working in the Advanced Signal Processing in Engineering and Neuroscience laboratory (ASPEN lab) was the most educational experience I have ever had. Besides the great knowledge I obtained about speech-based BCI and research in general, I believe that I have acquired many additional skills. I would also like to thank Dr. Oscar Gonzalez for his help through the administration process. He provided a huge support to me through all my struggles during my study at ODU.

I am very grateful to my parents; without their invaluable love, kindness, support, encouragement, and guidance I would not have reach this level. I am also grateful to my brothers; each one of them contributes to my personality and beliefs. Special thanks to my brother Anas who opened my eyes to the beauty of science. Without my family, I would never be able to achieve what I have achieved so far.

I am also very thankful to my friends who always support me and always are ready to listen to my issues and provide help. Deep thanks to my friends in Norfolk: Wael, Mohammad, Belal and many others for all your help and the good company. A great thanks to my elementary and high school teachers. Also, I am very thankful to my people in Palestine who taught me that we exist to resist and never give up. Regardless of the struggles of that land, I am very thankful to have grown up on the land of Palestine, and I am very proud to be one of these great people. Special thanks to my undergrad school, Birzeit University, which provided me with great knowledge and skills.

Finally, all praises to Allah the most gracious and the most merciful, who created all these beauties in life, for all blessings, guidance, and infinite mercy that he has been giving me. “My worship and my sacrifice and my living and my dying are for Allah, Lord of the Worlds.”

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Technology has been evolving in many aspects to help humans to perform tasks more efficiently, yet there are people who are not able to benefit from current technologies: people who suffer from severe neuromuscular disorders such as spinal cord injury, cerebral palsy, amyotrophic lateral sclerosis (ALS), and others. This group cannot perform daily tasks without assistance, and they depend on caregivers which adversely impacts their quality of life. Existing assistive technologies have limitations that do not allow these patients to communicate and perform their tasks in an independent and efficient way. A Brain-Computer Interface (BCI) is a promising tool for these patients. BCI technologies act as the pathway between the brain and other devices, such as wheelchairs, artificial limbs, and speech synthesizers. BCI technologies are also evolving beyond traditional domains to Augment and Virtual Reality [1], improving athletic performance [2], and helping doctors cure diseases like Attention-deficit/hyperactivity disorder (ADHD) [3].

1.1 BRAIN-COMPUTER INTERFACE

A Brain-Computer Interface (BCI) is a way to build a bridge between computers or machines and the brain. Its ultimate goal is to give people the ability to control, modulate the environment, augment, and communicate with the outside world [4]. There are many BCI techniques that have been developed based on different physiological signals like electroencephalogram (EEG), which is a non-invasive way to measure brain signals by placing

electrodes over the scalp, in addition to, Electrocorticography (ECoG) (see 2.1 for further details). Where each one of these has its own advantages and drawbacks.

The general framework for BCI is depicted in Fig. 1, which is constituted of four stages: signal acquisition, feature extraction, translational algorithm, and device commands. These modules will be briefly explained in this section, and more explanation will be provided in 2.3.1. Different neural signals are acquired from the brain; some of them are non-invasive, such as EEG and functional magnetic reasoning imaging (fMRI), and others are invasive like ECoG and stereotactic EEG (sEEG). Brain signals need to be analyzed and converted into better representations, and here comes the role of the signal processing module. Brain signals are commonly characterized in terms of frequency bands or oscillatory activities such as Delta [.5-4]Hz, Theta [4-8]Hz, Alpha [8-15]Hz, sensorimotor rhythms(SMR) [12.5-15.5]Hz, Beta [15-30]Hz and Gamma [30-250]Hz [5]. Unified definitions for these oscillation terms have not been agreed on, and they can vary slightly in other literature. Although it was empirically found that the frequency domain is a better representation for neural signals in general, the time domain still provides unique information. For instance, knowing the temporal propagation for speech-related neural activities is a fundamental key to decoding speech from neural signals.

After converting signals to a more informative representation, the step of constructing features from that representation is taken. For instance, the mean of the gamma band power in specific time duration. Then, translation algorithms map the resulting features on specific commands or other signals like speech or text [7]. The common implementations for these translational algorithms come from machine learning, where the problem turns into a

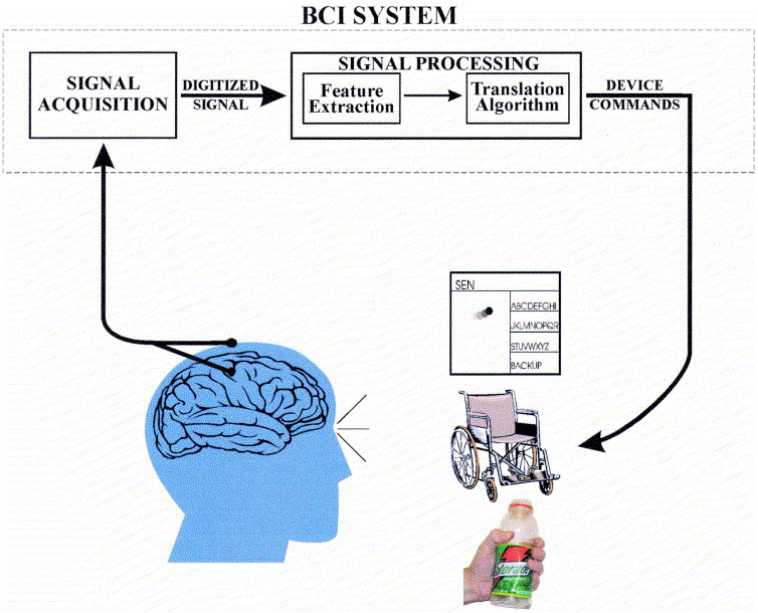


Fig. 1: The general framework for BCI. First of all, the signal acquisition is performed then the signals are processed using various signals processing techniques, yields features. Those features are fed into a translation model such as machine learning models. These models map the features into their corresponding action [6].

classification or regression problem.

1.2 SPEECH-BASED BCI

Various attempts to decode the intent of users have been employed [8]. Techniques based on ERP, like P300 and SSVEP were developed [9, 10, 11, 12, 13, 14, 15]. There is also motor imagery based BCI where the patient imagines moving hands, arms, or feet to issue a certain command [16]. These BCI techniques have many limitations: for example, the number of commands they can provide is generally insufficient or restricted, or they require a distracting focus on the paradigm and not the task. On the other hand, speech is the most natural way for humans to communicate. It can be decomposed into smaller units like auditory (e.g., formants, Mel frequencies) and articulatory features (e.g., place and manner

of articulations), yet composing these units gives a rich and dense way to convey information. The ability to decode such information from the brain will give humans the ability to communicate in a more natural and dynamic way. However, this proves to be a very challenging proposition. First of all, the brain is responsible for many tasks and actions and yields speech-related neural activities that are tangled with other information related to completely different tasks. For instance, the motor cortex plays an important role in speech production [17] and voluntary movements [18]. Therefore, if the patient attempts to move a finger while trying to convey his inner voice, it will cause a change in the information structure embedded in the neural signals from that area. Secondly, the spatial-temporal mapping for speech propagation in the brain is not well understood. This limits our knowledge of what timings and which electrodes should be used. Thirdly, the best representation of the speech-related neural signals is not well understood. Many hypotheses have been proposed to determine whether the best representation is articulatory, auditory, semantic features, or all of these combined. The answer to this question is critical to speech-based BCI since this prior knowledge will dramatically change the design and implementation of such a system. Fourthly, speech-based BCI aims to decode the inner(imagined) speech. Because there are no external cues about the exact onset and offset times for the inner speech, it becomes harder to isolate the inner-speech-related neural activity. Thus, there are many challenges to developing a practical speech-based BCI system; a system that works in a "plug-and-play" fashion like virtual assistants such as Siri, Google assistant, and others.

1.3 MOTIVATION

Current BCI techniques suffer from low information transfer rate (ITR) and very restrictive paradigms. For instance, the highest ITR that is reported to date from a BCI is 5.32 bits/second [19]. The ITR for spoken communication, however, is much higher. Therefore, having a device that is capable of compensating for the absence of the ability to speak due to neuromuscular diseases is the ideal solution.

The development of a BCI that decodes speech directly from the brain will lead to a more natural communication modality as well as higher ITR. In addition, speech-based BCI is closely tied to well-studied fields such as speech science, neuroscience, cognitive science, and speech signals processing. Thus, speech decoding work builds from a strong foundation of a confluence of existing studies.

Several prior studies have attempted to decode speech based on the state-of-the-art modeling techniques from signals processing and machine learning methods [20, 21, 22, 23, 24, 25], but the results were not good enough for practical applications. An alternate approach is to better understand the speech processes in the brain to build more prior knowledge into the models.

The current approach of understanding the speech-correlated neural signals is done by mapping these signals to corresponding articulatory and auditory features [26, 27, 28, 29, 30]. As an extension to this approach, this thesis aims to address the following questions:

1. How is speech-related neural activity in the motor cortex represented in terms of the articulatory and auditory features? The place and manner of articulations were chosen as representations for the articulatory features. Phonemes were chosen as

representations for the auditory features.

2. What is the nature of temporal activations for each representation in the motor cortex?

3. What is the role of the temporal lobe during speech production? And how are the neural activities represented during speech production?

1.4 PRIMARY CONTRIBUTIONS OF THIS THESIS

A majority of earlier studies discussed what feature representation is encoded in the motor cortex. Also, they discussed the role of different speech-related regions in speech production and perception stages. The first contribution of this thesis is that it provides a comprehensive summary and review of earlier studies for decoding and characterizing speech-related activities in the brain in both stages, as described in Chapter 2. An important contribution of this study is that it decodes and characterizes the temporal propagation of the auditory and articulatory features in the motor cortex. Discovering the temporal propagation of these two representations is important to know how the neural activities in the motor cortex will be used in a speech-based BCI system. Another contribution of this thesis is that it defines the role of the temporal lobe during speech production. Although the role of the temporal lobe is known during speech perception (especially, the auditory cortex), its role during speech production is still not well-defined. Discovering the role of the temporal lobe while speech production takes place can increase the amount of information that is obtained during speech production, which will increase the efficiency of a speech-based BCI system. Thus, this thesis contributes to providing new knowledge to address these two issues as follows:

1. Characterization and decoding the auditory and articulatory features of speech in the motor cortex: Auditory representations were chosen to be phonemes and two articulatory features were chosen, which are places and manners of articulation. These representations were decoded using gamma ECoG activity in the motor cortex. Also, their temporal propagation before and after the speech onset was performed using classification and statistical tests.

2. Defining the role of the temporal lobe during speech production: a modified version of speech spectrogram was chosen to be a speech representation and decoded using gamma ECoG activity. Deep learning was utilized in this analysis.

These analyses highlight the temporal propagation of articulatory and auditory features with respect to the onset at a high temporal resolution. Previous researchers have not used ECoG in such a way. Instead, there were attempts to understand the temporal differences between each representation using fMRI which has a poor temporal resolution. Secondly, these analyses highlight the role of the temporal lobe during speech production as well as the activation/engagement of the temporal lobe along different time lags with respect to speech onset. A very recent fMRI study indicated that there is predictive coding in the auditory cortex during speech production. This analysis confirms this conclusion. Lastly, these analyses highlight the usefulness of deep learning as an analysis tool in BCI. These collective findings provide important insights toward developing an efficient speech-based BCI system.

1.5 DISSERTATION OUTLINE

The remainder of this thesis is organized as follows: Chapter 2 discusses the background for this study, highlighting the state of the art in both speech decoding from ECoG signals and characterizing speech-related neural activities studies. Chapter 3 covers the datasets analyzed and their corresponding experimental paradigm in this thesis, in addition to the procedure of data analysis. Chapter 4 details the decoding and temporal propagation characterization of auditory and articulatory features, in addition to the characterization of the temporal lobe role in speech production and perception with respect to different lags. Chapter 5 concludes the thesis with a discussion of the main results and possible future work.

CHAPTER 2

BACKGROUND

This chapter provides a comprehensive overview of the properties, advantages, and drawbacks of ECoG signals and explains why they are considered superior over other types of electrophysiological signals such as EEG. In addition, it gives a comprehensive overview of current knowledge and research that has been conducted on speech production and perception. Lastly, a comprehensive overview of research on spoken and perceived speech decoding from the brain is provided.

2.1 ECOG PROPERTIES

Electrocorticography (ECoG), or intracranial electroencephalography (iEEG), is a type of invasive electrophysiological signal that is the result of voltage differences between electrodes placed directly on the exposed surface of the brain. Fig. 2 shows the anatomical structure for ECoG placement in relation to other electrophysiological signals. ECoG has a better temporal and spatial resolution compared to other techniques like EEG and fMRI. In addition to a high signal-to-noise ratio (SNR), ECoG is less susceptible to artifacts. The temporal and spatial resolutions are less than 1 millisecond and 1 cm respectively [31]. Although fMRI has high spatial resolution and it has been used in the characterization of speech production and perception (see, for example, [32] and [33]), it lacks sufficient temporal resolution for BCI applications. Also, in comparison, EEG has an excellent temporal resolution but poor spatial resolution(about 6–9 cm) [34], which makes it incapable

of capturing the subtle variations of speech-related neural activity. Therefore, ECoG has significant benefits for the development of neuroprosthetic devices. Nevertheless, it is worth mentioning a recent study done by Ibayashi *et al.* (2018), where they tried to classify five vowels using three different invasive techniques: single/multi-unit activity (SUA/MUA), local field potential (LFP), and ECoG. Their aim was to test which signal gave the better representation of speech-related neural activities in the human ventral sensorimotor cortex (vSMC). Their conclusion was that no significant difference existed among the decoding accuracies of the three individual signal modalities when averaged across subjects. On the other hand, when all three signal modalities were combined, it was found that the decoding performance was significantly improved. This suggests multi-scale signals convey complementary information for speech, at least from the vSMC area [35].

Since ECoG is a valuable technique for brain mapping, it is routinely applied for clinical purposes such as epilepsy monitoring. Data collection is done while the electrodes are implanted in the patients, and experiments are conducted while ECoG data are collected. In short, ECoG is a very promising tool to study speech-related neural activities because of its spatial and temporal resolution. All other methods compromise spatial resolution over the temporal or vice versa.

2.2 SPEECH-RELATED NEURAL ACTIVITIES

In this section, the focus will be on ECoG-related studies and non-ECoG studies (i,e. fMRI) with notable results. Our knowledge of spatial and temporal characterization for speech production and comprehension in the brain is still scarce. Many challenges exist for thorough characterization of these phenomena. Firstly, it is vital to understand which brain

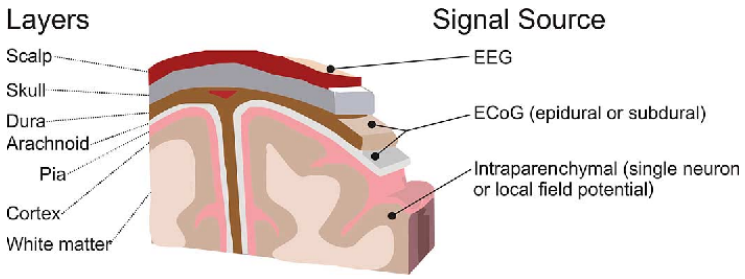


Fig. 2: Anatomical structure for different electro-physiological signals. Adapted from [4].

regions are involved in speech production and comprehension and what specific functionalities these regions are responsible for. Also, it is important to understand the nature of interactions between these brain regions and associated networks. Moreover, it is necessary to discern the temporal propagation of speech activities through the language cortex.

Local field potentials (LFP) recordings have been used to classify words from face motor cortex (FMC) and Wernicke's area [23]. FMC is known to be involved in controlling the articulation of speech [36], so the capability of decoding words from FMC may result from that unique sequence of articulations for each word. The capability to decode words from Werchicke's area is primarily due to two reasons. Firstly, Wernicke's area is identified as an important unit in language processing [37]. Secondly, words have unique semantics in addition to auditory and articulatory features. Brumberg *et al.* (2016) showed that fronto-motor areas become active in the planning and execution of speech, while auditory regions are primarily active for the role of acoustic feedback processing. They also showed that speech intensity is an appropriate representation for overt but not for covert speech production and a new way of alternative features that better represent covert speech is needed. In addition, between the two conditions, covert and overt, the auditory, pre-motor and motor areas were

activated [38]. Another study showed that in speech perception experiments, replacing or masking a phoneme by noise will induce a neural representation correlated with the actual phoneme [39, 40]. This phenomenon is called the phonemic masking effect. This may mean that a phonemic-related neural representation exists during speech imagining [41]. Bouton *et al.* (2018) hypothesized that the current models (especially those which rely on machine learning models) are capturing speech-related information from associative or redundant neural processes. Using a combination of fMRI, ECoG, and EMG they argue that early neural activity in posterior superior temporal gyrus (pSTG) is sufficient to decode syllables by a machine and all more distributed activity patterns, although classifiable by machine, reflect collateral processes of sensory perception and decision [42].

Music and speech processes share common brain networks [43], so any progress in one of these areas may improve the other. Martin *et al.* (2017) designed a novel experimental paradigm that allows a more accurate temporal localization of imagined music. They recorded ECoG signals while a piano player played two piano pieces with and without auditory feedback of the sound produced by the piano. The piano signal was recorded in both conditions in synchrony with the ECoG which allowed a more accurate temporal localization for the covert condition. They found robust similarities between the neural activities of both conditions in both temporal and frequency properties in auditory areas [30, 41]. This supports the possibility of implementing a BCI device based on inner voice, like imagining a musical tone or a speech sound.

To summarize, models of speech processes in the brain include the following cortical

regions: ventral primary motor cortex, ventral premotor cortex, inferior frontal gyrus, somatosensory cortex, primary auditory cortex, and peri-Sylvian auditory regions. Models show that those regions have roles in speech production, perception, or both [38].

2.2.1 SPATIAL AND TEMPORAL CHARACTERIZATION OF COVERT AND OVERT SPEECH PRODUCTION

Kellis *et al.* (2010) showed that face-motor cortex performance in words classification is superior over Wernicke's area. Pulvermller *et al.* (2006) showed in a fMRI study with spoken syllables that the motor cortex maps the articulatory features of speech sounds[44]. In another study, Pei *et al.* (2011) showed in an overt and covert (inner voice) word repetition task that the overt word production was associated with high gamma activities in superior and middle parts of the temporal lobe, Wernicke's area, the supramarginal gyrus, Broca's area, premotor cortex (PMC), and primary motor cortex. Covert word production was associated with high gamma activities in the superior temporal lobe and the supramarginal gyrus. Moreover, this study showed that the articulatory processing for overt and covert speech of the subjects was correlated with changes in high gamma power. This study also showed that in the covert condition, the (inner) speech-related activities are weaker than the overt condition, except at the superior temporal lobe where they were the same. In addition, this study was one of the first studies that successfully classified imagined speech [45]. Bouchard *et al.* (2013) showed that speech-articulatory representations in vSMC reflect a temporal organization during speech production. Each articulatory representation was spatially distributed across vSMC and a discriminability of consonants and vowels through temporal propagation in the same area was also found [46]. Bouchard *et al.* (2014) showed

that the acoustic features (formants) of cardinals vowels can be decoded from the sensorimotor cortex during speech production [47]. Mugler *et al.* (2014) classified all phonemes from the functional speech motor cortex [20]. The interesting observation from this study is that the neural activities in the motor cortex after the phoneme uttering onset were still related, which may give a hint that the motor cortex role goes beyond speech production to monitoring. Furthermore, the misclassified phonemes were correlated with the phoneme that shares similar articulatory features, which points to the role of articulatory features in speech processing. Flinker *et al.* (2015) showed that Broca's area mediates the activations of sensory representations from the temporal lobe to their corresponding articulatory features in the motor cortex. Thus, Broca's area coordinates the information processing on large-scale cortical networks before articulations [48]. Martin *et al.* (2016) showed in an imagined words classification study that speech imagination is represented in the temporal lobe, frontal lobe and sensorimotor cortex, consistent with previous findings in speech perception and production [49].

A recent fMRI study, however, suggests that there is predictive coding in the auditory cortex during speech production [50], where it was activated during silent uttering comparing with an imagined condition. Nonetheless, this suggestion is inconclusive and different interpretations of such results can be proposed. Thus, further investigation must be done.

Conant *et al.* (2018) designed a novel experimental paradigm to detect supralaryngeal articulators (lips, jaw, and tongue) by combining ultrasound and video monitoring with ECoG recording during vowel production, which gives the capability to measure articulatory parameters, for instance, vertical distance between lips, in sync with ECoG recordings.

They found that the high gamma activities in vSMC electrodes strongly encode at least one of the kinematics (position, speech, velocity, and acceleration) of the previously mentioned articulators but less for vowel formants and identity. The best kinematic parameter represented in the high gamma activity was speed. They also found that the best encoding and decoding occur at the onset and offset of the articulators [51].

2.2.2 SPATIAL AND TEMPORAL CHARACTERIZATION OF COVERT AND OVERT SPEECH PERCEPTION

As in the case of speech production, the picture of spatiotemporal propagation for speech perception has not been completed yet, so there are many efforts to extend the current knowledge of speech perception. It is worth mentioning that speech production and perception complement one another, since the process of speech is a closed-loop of production and perception. Crone *et al.* (2001) showed that the neural responses for speech and non-speech (tones) stimuli differ, where speech-related gamma power is higher than the nonspeech-related high gamma power [52]. Kubanek *et al.* (2013) showed that while listening to a stream of verbs associated with hand(e.g., throw) and mouth(e.g., blow) embedded within an unintelligible nonwords sequence, the perception of verbs compared to nonwords activates first the posterior STG, then the middle STG, followed by the superior temporal sulcus (STS) [26]. This supports previous studies that STG plays a major role in speech comprehension [53]. Chang *et al.* (2013) showed that the alteration of the pitch in the subject's recorded voice will activate the auditory cortex more than listening to unaltered speech recordings compared to the auditory cortex activities while speaking. The subjects were found to compensate for this perturbation by changing the pitch of their sound [54]. This

high activation reflects a sensitivity to unexpected feedback. This also implies the complex interaction between speech production and perception. Cheung *et al.* (2016) showed that speech-related activities in the motor cortex while listening, substantially differ during articulation of the same sounds. In addition, they found that the structure of these activities during listening was organized in an auditory-cortex-acoustic-features pattern. This leads us to conclude that while listening to speech, the motor cortex neural activities represent auditory-vocal information [55].

Commonly, speech-perception-related neural activities are studied in terms of speech signal features. Numerous studies showed that these features can be reconstructed from high gamma activities in the auditory cortex. Pasley *et al.* (2010) demonstrated that the high gamma band is correlated with spectro-temporal fluctuations of aurally presented words and sentences, where the spectrogram for single words was constructed from ECoG signals. They found that speech signal is represented by two models in the auditory cortex. The first one is the spectrogram which well-represents low frequency speech features such as syllable rates while fast temporal fluctuations like syllable onset and offset are represented in a nonlinear transformation of the spectrogram model, which is the second model [28]. It also was shown that the auditory cortex tracks the speech envelope and this correlation has a phonological nature [26]. In another study, Chang *et al.* (2010) found that speech-related neural activities in the STG have a representation of phonemes, which is direct evidence for acoustic-to-higher order phonetic level encoding of speech sounds [56]. Mesgarani *et al.* (2012) showed that the auditory cortex gives rise to the perceptual aspects relevant to the listener's intended goal in a mixed speakers experiment. They found that the spectrogram

of mixed speech reconstructed from neural data reveals the salient spectral and temporal features of the attended speaker[57]. Mesgarani *et al.* (2014) studied phoneme category (high-front vowel, nasals, etc.) representations in STG and found that speech-related sites in STG are selective to specific phoneme groups. Furthermore, the manner of articulations plays a major role in this selectivity and secondarily by place of articulation. In addition, an encoding of acoustic formants, voice onset time (VOT), and spectral peak from STG were successfully done. As a result, STG appears to give rise to the phoneme representations [58]. Martin *et al.* (2014) showed that the neural activities for both overt and covert speech have similar encoding for speech features (spectrogram features and non-linear transformation of the spectrogram) where the linear model built based on the overt condition was used to predict speech features for the covert condition [25].

Berezutskaya *et al.* (2017) conducted a study based on ECoG and fMRI to investigate the neural representations during continuous speech comprehension before they are processed in the STG area. Their results showed that low-level speech features propagate throughout the perisylvian cortex [59]. The speech envelope contains key information about speech intensity and phonemic content essential for a complete understanding of speech. The neural response in the auditory cortex is temporally aligned with the speech envelope in the delta/theta band. This phenomenon is called "speech-brain entrainment" [60]. Lower frequencies (<4Hz) are associated with syllable (defined as a cluster of sounds that contains at least one vowel) rate, but higher frequencies (>16Hz) are related to syllable onsets and offsets. Formants (defined as the spectral peaks in the spectrogram space) are relatively unique for each phoneme(they vary for each person) [61, 62, 63]. Riecke *et al.* (2018)

showed that this entrainment has a causal role in speech intelligibility using brain stimulation experiments [60].

Chang *et al.* (2015) were able to decode the intended speech in a multi-speakers environment (cocktail party problem) using neural features from the STG area. Their experimental paradigm was to represent two speakers at the same time and let the subject focus on a specific target. Using neural signals outperforms the acoustic signal when they were used to build a model to detect the target speaker, which suggests that the brain ignores the non-target-related speech signal [64].

2.3 DECODING SPEECH FROM ECOG SIGNALS

Research of characterization of speech-related neural activities in the brain started many decades ago. Decoding speech from neural signals, however, has only recently been attempted while the goal for the former is to better understand speech propagation in the brain in such a way, for instance, to better diagnose speech-related diseases. The goal of the latter is to provide technology for the severely disabled to improve their quality of life. Although decoding speech from the brain is not fully concerned with discovering the nature of speech-related activities, it is often necessary to advance the quality of the decoding schemes since traditional black box approaches have limited efficacy. Thus, prior knowledge is required to enhance these models or to create new approaches based on the special characteristics for speech-related neural signals. Other methods to process a speech signal (before the appearance of deep learning), were influenced by the way the auditory system processes speech (see MFCC method as an example). Based on that, revealing prior knowledge of very complicated and dynamic signals like ECoG is a logical pursuit.

This section provides a detailed explanation of both the general framework of decoding speech from ECoG signals and the literature review for related studies. The section will conclude with a discussion of the limitations of existing techniques.

2.3.1 THE FRAMEWORK OF DECODING SPEECH FROM ECOG

The general framework of extracting speech-related information from brain signals was briefly discussed. In this section, a detailed review will be presented. The proposed speech-based BCI decoding models either perform continuous speech reconstruction, such as the spectrogram of an intended speech, or they perform a discrete output, such as phonemes, words, or articulation places. The discrete output systems have an advantage over the continuous ones as their margin of error is narrower. On the other hand, the discrete output systems are highly restricted and may not be able to be generalized to build an efficient speech-based BCI system. For example, a speech-based BCI which can decode five words that cannot be generalized to be used in complicated tasks, such as generating continuous speech. Therefore, the choice of either a discrete or a continuous output system depends on the application, for instance, a system that reliably decodes ten words might be suitable for some scenarios. Fig. 1 shows the different modules that compose a speech-based BCI system. In the following, each module will be discussed in detail by providing the working principle, considerations, challenges, and conclusions from recent investigations.

Signal Acquisition and Characteristics

Although there are different types of neural signals to capture speech-related activities, the focus will be on ECoG signal acquisition and properties. ECoG is used clinically for

locating epilepsy seizure foci prior to surgical resection. An ECoG electrodes grid is placed on the cortex, usually beneath the dura. The typical number of electrodes in a grid is 8×8 electrodes of 4 mm diameter each and a 1 cm inter-electrode distance. It is typical to require 5-12 days of continuous recording to localize epileptic foci [31]. Additionally, the location of electrodes is solely chosen based on clinical purposes.

Signal Processing and Features Extraction

The raw ECoG signals are pre-processed by first filtering using the common average reference (CAR) filter. This process reduces the common components across channels such as global artifacts [52, 53]. Next, the signals are high-pass-filtered with a cutoff frequency between 0.5-2 Hz to eliminate DC drift. The signal is notched-filtered around the fundamental power line frequency and its harmonics (60 or 50 Hz depending on the electrical system).

The Discrete Fourier Transform (DFT) is commonly used to extract the features of the signals as much of the information in ECoG can be captured by the spectral dynamics. For example, information related to phenomena such as speech, memory, cognitive function, learning and motor tasks has been mostly found in the Gamma band [65, 66, 67, 68, 28]. To take into account the temporal variability of the signal, the short time Fourier transform (STFT) and wavelet transform (WT) are the common tools to perform time-frequency domain analysis [69, 20]. After transferring signals to more informative representations (e.g., time-frequency domain), features can be extracted. For example, in the case of the usage of STFT in decoding speech from neural signals, the average power of the gamma band with a 50 ms time window was taken [28, 20, 21]. Generally speaking, feature extraction

algorithms must represent the information in the neural signals to better facilitate modeling that will be carried out by the feature translation module.

Features Translations

After obtaining features from the extraction module, the translation module tries to translate these features into a certain command or an output, such as a phoneme. Feature translation relies on machine learning algorithms where it learns what feature vector represents the desired output, based on a mathematical model. These models can be linear, for instance, linear discriminant analysis (LDA), or nonlinear like artificial neural networks. Neural signals are very dynamic and they vary according to many factors, such as disease, emotional state, etc., since brain regions that are related to speech are responsible for many other functions. Therefore, there are many other factors and variables that implicitly exist in the neural signals that carry speech-related information. Based on that, and up to now, most feature translation algorithms have been subject-dependent and require training on data acquired from an individual subject.

After constructing the model of feature translation, it needs to be validated. A complicated model can simply over-fit the data which may hinder performance on independent test data, which is a well-known issue in machine learning. The common technique to handle this is to divide the data into training and testing data. The ratio of training to testing data can vary in the BCI literature from 7:3 to 9:1. Moreover, a further step is taken to ensure that over-fitting has not occurred which is to partition the data into n chunks (usually 5 or 10) and build the model using $n-1$ chunks and the remaining chunk will be used as testing data. This process is repeated n times where each chunk is used exactly once as testing

data. Next, the average of the testing performances of each chunk is taken. This is known as cross-validation.

Other fields like image and audio processing started using deep learning to combine this module and the feature extraction module in one algorithm. However, for neural signals, this remains a challenge because of lack of data, data variability, and complexity. Nonetheless, the deep network seems to be a valuable tool in speech-based BCI, which will be explained in section 2.3.2

Output Device

An Output device will convert the command into an actual output. This can vary depending on the application, where it can be a wheelchair, a prosthetic arm, or a speech synthesizer.

2.3.2 DECODING SPEECH FROM BRAIN

Several recent studies have attempted to decode speech from the brain. In this section, the decoding of imagined and actual speech will be discussed. The main focus will be on ECoG related studies. In an earlier work, Kellis *et al.* (2010) classified a set of 10 words from local field potential (LFP) recordings from the motor cortex and Wernicke's area with an above average chance of accuracies. Their feature extraction algorithm relies on the gamma band power and principal component analysis (PCA) for features reduction. To improve their classification, channels' selection were applied and it was found that the best performance was obtained based on motor cortex electrodes. Their conclusion demonstrates the importance of the motor cortex in speech processing, where the motor cortex

is responsible for monitoring speech articulations muscles and joints [23]. Blakely *et al.* (2008) classified four phonemes in a pair-wise fashion. They used the support vector machine (SVM) to classify the features which were the power of six different frequency bands for 0.265 s time window with 37% overlap [70]. Pei *et al.* (2011) decoded four vowels and nine pairs of consonants in a monosyllabic words experiment, where words were composed of a Consonant-Vowel-Consonant (CVC) pair. The experiment was conducted under two conditions: imagined and spoken conditions. For both conditions, the classification results were statistically significant above the level of chance. Their classification was based on naive Bayesian classifier and feature selection was applied using a maximum relevance and minimum redundancy technique [24]. Their study was one of the first efforts to examine the possibility of classifying imagined vowels and consonants [71]. Leuthardt *et al.* (2011) were able to classify two imagined phonemes in an online study on two subjects. Their experimental paradigm was based on two conditions: covert and overt. In their study, the same feature extraction and classification techniques were used for both conditions, which demonstrates the similar characteristics of overt and covert speech [72]. Zhang *et al.* (2012) classified two spoken sentences from the posterior part of the inferior frontal gyrus using high gamma power envelop and Fisher discriminant analysis as a classifier. Dynamical time warping (DTW) was used in their study to find the optimal temporal onset of sentence uttering, where it is used to overcome the variability in time and speed of single-trial speech such that an assumption was made where there is a unique temporal-neural pattern for each class or category follows. They avoided using trials averaging as it determinates finding this pattern. Their results are much higher than the level of chance and the DWT

approach outperforms an approach that was based on SVM without DWT [73]. Extending these ideas, Mugler *et al.* (2014) classified all American phonemes in a words-repetition experiment. They classified all phonemes with up to 36% accuracy. In their procedure, STFT was applied on ECoG signals and feature selection was performed based on analysis of variance (ANOVA) and LDA was used as a classifier [20]. Lotte *et al.* (2015) demonstrated that phonetic features can be decoded from ECoG data. They tried to identify which brain regions and what time in respect of phoneme onset are most activated under three different features representations, i.e phonemes, place of articulations, and manner of articulations. They used LDA to model neural activities based on the three mentioned representations. In their procedures, different models were built using different sets of features that vary in their spatial and temporal characteristics or parameters. The model that performs the best based on the level of chance baseline, its spatiotemporal characteristics were given the highest score. Furthermore, nearly all speech-related brain regions were included. Although their results were statistically significant above the level of chance, they did not report their classification accuracies [74]. Kanas *et al.* (2014) worked on detecting speech activities from different areas of the brain. Their experimental paradigm was based on repeating two syllable nonwords composed from both a consonant among six different consonants and a vowel among two different vowels (e.g. pah, dah). Their methodology was based on joint spatial-frequency clustering of the ECoG feature space where they grouped both spatial (channels) and frequencies features based on their contribution to the target using a k-means algorithm. In particular, the first group contains the top features which make the best contributions to the target, and the second group contains the second-top

features and so on so forth. Then, different models were built based on these groups of features. For instance, the first model was built based on the first group features, and then another one was built based on the first and the second group, etc. The model which shows the best performance was chosen for testing. A 98.8% testing accuracy was achieved and also based on their analysis the 8 Hz frequency was the optimal frequency to detect speech activities in the brain. However, this study was based on one subject [75]. Herff *et al.* (2015) utilized a language model to classify phones and words during a continuous speech experiment. Gaussian models were used to map ECoG to the corresponding phones and a bi-gram language model was used to support or oppose the Gaussian ECoG-based models. In this case, the language model is known as prior knowledge and the Gaussian models as a posterior probability. The final decision is taken for the class which maximizes the product of these two quantities [21].

So far, the aforementioned studies are based on the discrete output (phonemes, consonants, vowels, sentences, etc.). On the other hand, many attempts to decode speech from ECoG signals in a continuous output fashion have been conducted. In a consonant-vowel syllables (CVs) uttering experiment, Bouchard *et al.* (2014) were able to predict the formant frequencies from ECoG data that were recorded from the vSMC area. In their procedure, they used adaptive principal components regression as a feature extraction algorithm [47]. Martin *et al.* (2014) successfully decoded two representations of the speech signal. The first one is the spectrogram of the speech signal, which is defined as the amplitude of speech over time and frequencies. The second one is modulation-based speech signal, which is a nonlinear transformation of the spectrogram. Their work was based on two conditions,

overt and covert, wherein the overt condition, subjects loudly read short paragraphs and in the covert condition, subjects silently read the same paragraph. The experiment was designed in a way to ensure the pace of both silent and loud readings were the same. Since there are no covert speech markers, the authors used DTW to realign the constructed overt speech with the covert speech. A linear decoding model was used to predict both representations of the two conditions based on overt condition data. That is, the prediction of the two representations in the case of the covert condition was based on the overt condition data. The correlation coefficient was the metric to validate the model. More specifically, the realigned (using DTW) constructed overt speech was correlated with the corresponding speech yielded from the overt condition. The reconstructions of both representations in the two conditions were found statistically significant. This study provides a proof of concept that covert speech can contribute to overt speech decoding [25]. Also, Martin *et al.* (2016) conducted a word repetition experiment in both overt and covert conditions. They utilized an SVM model based on DTW distance kernel. That is the time series of ECoG segments was realigned using DTW and then they were inputted into the SVM model. The reasoning behind DTW-based kernel SVM is that imagined speech lacks obvious markers, which means that two trials that belong to the same class may have different time alignments. Therefore, the classifier might not recognize two trials as belonging to the same class. The classification was done in a pair-wise fashion, where both covert and overt conditions were statistically significant above the level of chance. The mean accuracy for the overt condition was 89% and for the covert one was 58%[49].

Decoding speech from the brain while listening is also considered a very important tool in

the development of efficient BCI devices which help people who suffer from hearing impairments. Decoding speech perception can also play an important role in decoding the intended speech since auditory feedback has a role in monitoring speech during speech production [76]. Chang *et al.* (2010) were able to classify three consonant-vowel syllables from the pSTG area in a listening task. They achieved that by characterizing differences between neural representations of each syllable using L1-regularized logistic regression. Then, the output of L1-regularized logistic regression was fed into the k-means algorithm to determine the final classification decision [56]. Pasley *et al.* (2012) examined constructing spectrograms of speech from the auditory cortex in a listening to sentences experiment. Two representations were found in the auditory cortex: spectrogram and modulation-based representation. The accuracy of word identification was 0.89 median percentile rank for 47 words [28]. Moses *et al.* (2016) utilized automatic speech recognition (ASR) techniques in classifying perceptual speech. Their approach was based on using hidden Markov models (HMM) that use probabilities coming from both the language model and ECoG-based models. The features were extracted by applying eight semilogarithmic-bandwidth-increasing bandpass Gaussian filters on the gamma band (70-150). Then, the envelop of each resulted band was taken using Hilbert transform. A feature reduction was applied using the first principal component. To find the better temporal characteristics of feature vectors, a Grid search was performed for each subject. Their results were similar to what Herff *et al.* (2015) showed; the performance was better when using the language model [77]. Moses *et al.* (2018) were able to classify 10 sentences in real time. They used a phoneme-based analysis using HMM, where they broke each sentence down into its phonemes. They also used a direct classification based on

sentence-level neural activities using LDA. They achieved 98% using HMM and 90% using LDA. The training time required to achieve this accuracy was less than seven minutes [78].

Deep Learning Role in Decoding Speech from the Brain

Deep learning recently has been demonstrated as a powerful tool in machine learning for many classification and regression problems. However, the main challenge for deep learning in the BCI field is lack of data, especially for a hard-to-obtain signal such as ECoG. Nevertheless, many studies utilize deep learning in neural data analysis (i.e., EEG and ECoG signals analysis). In this section, the focus will be on deep learning utilizations in speech-based BCI. Chang *et al.* (2015) were able to decode target speech based on STG neural signals during a multi-speakers task, where they asked the subject to pay attention to a specific speaker and to try to ignore the rest. The performance of the model which was built based on neural signals was much better than the model which was built based on acoustic signals. They used a deep neural network (DNN) and HMM, where DNN outputs were used as posteriors to derive emission probabilities for HMM [64]. Their results suggest that the STG area acts as a filter which removes unintended-speech-related information. This study can help people who are suffering from hearing impairments to better distinguish sounds in a multi-speakers environment. Livezey *et al.* (2018) built a DNN to be used as a classification and analysis tool. They exploited neural data which were recorded from the vSMC area during consonant-vowels (CV) syllables uttering. For a consonant-vowel pair classification task (57 classes), the deep network outperformed the logistic regression with a statistically significant difference for two among four subjects. The performance of these two subjects was increased by 50% and 100% respectively over logistic regression using deep

networks. For the consonants classification task (19 classes) the deep network outperformed logistic regression but exhibited similar performance to LDA. They also reported that the deep network's accuracy has been increasing with increasing data size for all subjects, which was not observed when using linear models. This result demonstrates the capability of deep learning to better model speech-related ECoG data than traditional methods. Furthermore, they showed that the confusion of the deep network, which is defined as being where the misclassification of a class is distributed over other classes, follows the articulatory structure of the phonemes [79]. O'Sullivan *et al.* (2017) used a long-short-term-memory-based DNN (LSTM-DNN) to implement a source separation system in a multi-speakers environment based on neural signals from the STG area. Their experimental paradigm was similar to the previously mentioned paradigm of Livezey *et al.* (2018), as they designed it in a way that subjects listened to different speakers reading a story at the same time and they were asked to focus on one of them. Their implementation was based on implementing a DNN for each target speaker to distinguish against the other. That is, each DNN was trained using a spectrogram of the speech signals mixture and with the intended speaker spectrogram as the output. Next, the output of these DNNs was correlated with a spectrogram constructed based on the neural data that were recorded from the STG during listening, where they employed a method known as stimulus-reconstruction to reconstruct the spectrogram from neural data. The final decision of what the subject was attempting to attend to was taken based on the DNN that gave the best correlation. For instance, if the first DNN gave the best correlation, then the first speaker will be considered what the subject was listening to. Electrodes selection was based on statistical significance for each electrode in the task

of distinguishing speech from silent. The validation of this system showed it was able to decode the attention of subjects. They also identified the electrodes that contributed the most to this task within the auditory cortex [80]. In a recent study, Akbari *et al.* (2018) showed that DNN outperforms traditional methods in auditory stimulus reconstruction by 65% over the baseline (traditional classifiers accuracy). Their experimental paradigm was that three subjects listened to isolated digits, where the goal was to reconstruct speech signals from auditory cortex neural activities. They used two different representations for acoustic signals: auditory spectrogram, which is obtained based on auditory perception model, in addition to speech vocoder, which synthesizes speech from four main parameters: 1) spectral envelop, 2) fundamental frequency, 3) band aperiodicity, and 4) a voiced-unvoiced (VUV) excitation label. To evaluate the models, subjective and objective evaluations were performed, where the subjects were asked to listen to a digit pronunciation and then speech signal was reconstructed from neural signals that were recorded during listening. After that, subjects listened to the reconstructed speech and reported what they had heard. The accuracy of this subjective measure was 75%. For the objective evaluation, they used the extended short time objective intelligibility (ESTOI) measure, which is a measure for the intelligibility assessment of speech synthesis technologies. The average ESTOI of all subjects was consistent with what subjects reported. Furthermore, they measured the correlation coefficient for decoding a speech spectrogram from neural signals taken from STG area. DNN outperformed the linear models such as linear regression in all cases. Also, according to the deep network model, the lower frequencies (0-50 Hz) have also contributed to the spectrogram reconstruction [81].

CHAPTER 3

METHODOLOGY

This chapter describes the methodologies for experimental paradigms used in this thesis. Firstly, experimental paradigms and data acquisition procedures will be described. Secondly, the signal processing, as well as the technique to decode articulatory and auditory features will be discussed. Thirdly, the primary technique to characterize the temporal propagation of auditory and articulatory features in the motor cortex will be explained. Since characterizing the temporal propagation of these two features representations needs a more controlled events localization (e.g., speech preparation and perception do not overlap), data obtained from a monosyllabic word repetition experiment was used. The experimental paradigm of this dataset gives the ability to eliminate any neural activities related to uttering other phonemes, which gives the ability to study neural activities that exclusively related to a particular auditory or articulatory feature. Characterization of a certain phenomenon requires eliminating any other factors that can change the measured signal. Lastly, the technique of decoding and characterizing the speech-related activities in the temporal lobe during speech production and comprehension will be discussed. Since this goal is defined on a macro level, where the ultimate goal is to determine the main function or role of the temporal lobe during speech production, ECoG data based on a continuous speech experiment was used. The continuous speech will generate neural activities that combine to perform the main function (e.g., speech monitoring). Therefore, the usage of a continuous speech

dataset is appropriate for defining the role of the temporal lobe during speech production and comprehension without considering the micro details.

3.1 DATA ACQUISITION AND EXPERIMENTAL PARADIGM

Two different datasets were used in this thesis. The first one, called the monosyllabic word dataset, was used to decode and characterize the articulatory and auditory features in the motor cortex. The second, called the continuous speech dataset, was used for decoding and characterizing speech-related activities in the temporal lobe during speech production and comprehension.

3.1.1 MONOSYLLABIC WORDS DATASET

Data Acquisition

Data were collected from three subjects who required intraoperative ECoG monitoring during awake craniotomies for glioma removal. All three subjects were native English speakers with no tumor-related symptoms affecting speech production. All of them gave informed consent to participate in the experimental paradigm, which was approved by the Institutional Review Board at Northwestern University. Both anatomical landmarks and functional responses to direct cortical stimulation were used to determine electrodes grid placements. Areas producing reading arrest when they were stimulated were designated as being associated with language. Areas producing movements of the articulators when they were stimulated were designated as being associated with speech prodution. ECoG grid placement covered targeted areas of ventral motor cortex (M1v) and premotor cortex

(PMv).

A 64-channel, 8×8 ECoG grid (Integra, 4 mm spacing) was placed over the speech motor cortex connected to a Neuroport data acquisition system (Blackrock Microsystems, Inc.). A customized version of the BCI2000 software was used to control stimulus presentation and to collect data [82]. A unidirectional lapel microphone (Sennheiser) was placed near the patient's mouth to measure the acoustic energy. The microphone signal was wirelessly transmitted directly to the recording computer (Califone), sampled at 48 kHz, and synchronized to the neural signal recording. All ECoG signals were bandpass-filtered from 0.5-300 Hz and sampled at 2 kHz.

Experimental Paradigm

Words were presented in randomized order on a screen at a rate of a single word every 2 seconds, in rounds of 4.5 in minutes length. Subjects were instructed to read each word aloud as soon as it appeared. Each subject completed 2 or 3 rounds. Stimulus words were chosen based on their simplicity, and phoneme frequency and variety. The words were monosyllabic words with consonant-vowel-consonant (CVC) structure and were selected from the Modified Rhyme Test [83], where the frequency of phonemes is approximately like the American English phoneme frequency [84]. Additional CVC words were added to the set to include all American English phonemes in a uniform frequency.

Events Labeling

Since the recorded speech signal was synchronized with ECoG signals, ECoG data were temporally aligned with phonemes based on the recorded speech signal. Phoneme labeling

based on speech signal was done using "Penn Phonetics Lab Forced Aligner for English" software [1], which is built based on hidden Markov model kit (HKT) [2]. In short, HKT is based on HMM using both continuous density mixture Gaussians and discrete distributions. The labels were validated visually and aurally using Praat software [3]. For articulatory features representations, the phonemes were grouped according to their articulation properties. Places of articulation and manners of articulation were chosen as two different grouping techniques.

The places of articulation are defined as where the vocal tract is either constricted or closed by one of the articulators. In this thesis, places of articulation were divided into four different groups: labial, coronal, dorsal, and laryngeal. Labial consonants occur when one or both lips participate in producing a consonant (e.g., /m/). Coronal consonants use the front part of the tongue (e.g., /n/). Dorsal consonants occur when the back of the tongue (the dorsum) has a role in the articulation (e.g., /k/). Laryngeal consonants have the primary articulation in the larynx (e.g., /h/).

The manners of articulation are defined as the configuration and interaction between articulators when producing a sound. The manners of articulation that were considered are: nasal, plosive, fricative, approximant, flap, and lateral approximant. Nasal sounds are nasal occlusive (i.e., closed vocal tract) sounds that are produced while the velum is in the lower position, and the air flows freely from the nose(e.g., /m/). Plosive sounds are sounds that are oral occlusive sounds produced by completely restricting the vocal tract with explosive release (e.g., /p/). Fricative sounds are produced by partially restricting the

[1] `https://web.sas.upenn.edu/phonetics-lab/facilities/`
[2] `http://htk.eng.cam.ac.uk/`
[3] `http://www.fon.hum.uva.nl/praat/`

vocal tract (a narrow channel), by placing two articulators close together and passing air through the channel (e.g., /s/). Approximant sounds are less extreme than Fricative, since they place two articulators together but not to the point that a turbulent airflow occurs (e.g., /r/). Flap sounds are produced when a single articulator is thrown against another for a short time such that no burst can happen, as in the medial sound in "ready". Lateral approximant sounds are produced only when the center of the tongue touches the roof of the mouth (e.g., /l/).

3.1.2 CONTINUOUS SPEECH DATA ACQUISITION

Data Acquisition

Data were collected from four subjects who suffer from intractable epilepsy and were undergoing treatment at Albany Medical Center. All of them gave informed consent to participate in the experimental paradigm, which was approved by the institutional review board of the hospital. The subjects were physically and visually able to perform the task and all of them had an IQ of at least 85.

The implanted electrode grids (Ad-Tech Medical Corp., Racine, WI) consisted of platinum-iridium electrodes (4 mm in diameter, 2.3 mm exposed) that were embedded in silicone and spaced at an inter-electrode distance of 1 cm. One subject was implanted with an electrode grid (PMT Corp., Chanhassen, MN) with 6 mm inter-electrode spacing. Grid placement and duration of ECoG monitoring were based on clinical purposes without considerations of this study's requirement. Fig 3 shows the electrode location of each subject.

Data Collection and Experimental paradigm

ECoG signals were recorded using eight 16-channel g.USBamp biosignal acquisition devices (g.tec, Graz, Austria). ECoG signals were recorded simultaneously along with speech signal at a sampling rate of 9600 Hz. Electrodes distant from epileptic foci and areas of interest were used for reference and ground. After removing channels which did not contain ECoG activity, subjects had 56-120 channels. The subjects' eye gaze was also recorded using a monitor with a built-in eye tracking system (Tobii Tech., Stockholm, Sweden). BCI2000 software was used to collect the ECoG, microphone, and eye tracker, in addition, to control the experimental paradigm [82].

Each subject was seated in a semi-recumbent position in a hospital bed facing a video screen with 1m distance. The text of a famous passage (e.g., Gettysburg Address or Humpty Dumpty nursery rhyme) ranging from 109 to 411 words, scrolled from right to left across the video screen at a constant rate between 20% and 35% per second. The pace was chosen based on the preference and cognitive capabilities of each subject. This resulted in run durations between 129.9 and 590.1 seconds. Each subject spoke the scrolling text aloud and the speech was recorded using the microphone.

3.2 DATA ANALYSIS

Data analysis is divided into two different parts, the first part regarding the monosyllabic words dataset, was chosen because it gives the ability to ensure the speech-perception and speech-preparation neural activities do not overlap; the second part regarding the continuous speech dataset, was used because it gives the ability to define the main function of

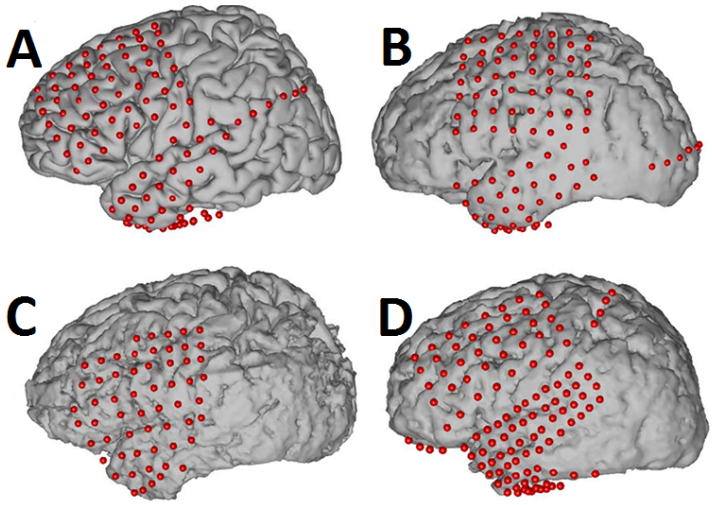


Fig. 3: The electrodes location of each subjects

the temporal lobe during speech production and perception. The pre-processing for both datasets is the same, but the data analysis is specific to each part.

3.2.1 PRE-PROCESSING THE DATA

ECoG channels were visually inspected for abnormal activities which were excluded from the analysis. Then, the data were high-pass-filtered with a cutoff frequency of .01 Hz to remove low-frequency components that do not contribute with information. After that, the common average reference (CAR) was applied. CAR removes the average of ECoG channels from each channel. Following this, a notch filter at the multiples of 60 Hz that lie within the range [70,290] Hz was applied to remove the power line noise. The resulting signals were band-passed in the range [70, 290] Hz. All applied filters were zero-phase filters since this guarantees that signals original components were not shifted or changed.

3.2.2 MONOSYLLABIC WORDS DATASET

Features Extraction

Pre-processed ECoG signals were segmented based on the onset of the phonemes. Each segment starts 300 ms before the phoneme onset and ends 300 ms after onset. Each segment was divided into 50 ms chunks, and each chunk was processed by applying equally-spaced bandpass filters with 20 Hz bandwidth. Although the signals were notch-filtered at 60 Hz and its harmonics, the bandwidths containing the 60 Hz harmonics were excluded from analysis to further eliminate the possibility of effects caused by the filtering. This yields 8 different band-passed signals ranging from [70-90] to [270-290]. Lastly, the *log* of the average of the absolute value for each band-passed segment was taken. The final number of features, in the case of 300 ms before and after the onset, is $NumberOfChannels \times 12 \times 8$.

Class labels and the corresponding feature vectors that had fewer than 15 instances were deemed insufficient for model training and excluded from the analysis. Thus, there are a different number of classes for each subject based on the total number of words presented during the experiment, which varied based on experimental time and subject compliance.

Features Selection

Since the number of dimensions for the feature vector is too high, a subset of features based on analysis of variance (ANOVA) was selected based on the K lowest p-values. The K was chosen empirically, based on each subject's data. Table 1 shows the number of features selected for each subject.

TABLE 1: Number of features selected for each subject

Subject	Number of features
A	140
B	140
C	160

3.2.3 CLASSIFICATION INDEX PROCEDURE

In this thesis, the classification index is used to indicate the statistically-significant brain activity. This procedure compares the output of a model built using actual data with the output of a model built using randomized data having similar spectral characteristics as the actual data. This type of test is well-established in BCI literature [38, 74]. This procedure will be also applied in Continuous Speech Dataset 3.2.5. This procedure is performed as follows:

1. The dataset is divided into a training and test set. The model created using the training data and the performance is evaluated using the testing data. This represents the performance based on the actual data.

2. For the randomization test, the data is again divided into training and test sets as in step one. The data in each set is then shuffled by randomly shuffling the output cross trials.

3. A model is built using the shuffled training data and the performance is evaluated using the shuffled testing data.

4. Steps 2 and 3 are repeated n times to fit a statistical distribution.

5. The actual performance is compared to the statistical distribution by computing the p-value.

6. the p-value that is larger than .05 (not statistically significant) is set to 1.

7. $-log$ is applied to the p-value.

3.2.4 ESTIMATING THE CRITICAL VALUE OF THE LEVEL OF CHANCE

In statistics, the critical value of .05, $\alpha_{.05}$, is a point on an statistical distribution with a probability of 0.05 to occur. In order to estimate the $\alpha_{.05}$ of the level of chance performance the continuous density function (CDF) of level of chance performance is obtained by applying the second, third and fourth steps of the classification index procedure 3.2.3. Next, $\alpha_{.05}$ is estimated using the percent point function (PPF), which is the inverse of the obtained CDF. Using this procedure, the estimated $\alpha_{.05}$ is considered one of the extreme values (e.g., one of the highest accuracies) that a model based on shuffled data may have. Using the $\alpha_{.05}$ of level of chance will make the interpretation of the results easier for the reader. For instance, when averaging Pearson correlation coefficients cross subjects, it is hard to tell when the mean is statistically significant. Estimating the critical value of the level of chance will determine whether the mean is statistically significant, where if the mean is well-above the critical value then it is statistically significant. Therefore, the reader will have an idea when the mean is significant by comparing with this baseline. In addition, it will save many runs of randomization tests. For instance, in case of a regression problem, estimating $\alpha_{.05}$ of Pearson correlation coefficients for a shuffled-data-based model can avoid

running the classification index procedure many times to judge whether a result is statistically significant above the level of chance. As will be illustrated in section 4.3, it is shown that $\alpha_{.05}$ of the possible Pearson correlation coefficients of an LSTM-RNN model that was built using shuffled output, is the same for all subjects (follows the same distribution with the same parameters).

Segmenting and Classification

The classification was done on the features that were selected using linear discriminant analysis (LDA). In order to determine the testing accuracy, the data was divided randomly into training and testing data with 8:2 proportion, and this was repeated ten times. The final testing accuracy was calculated by averaging the resulting ten values. To test the statistical significance of the model performance, the classification index procedure 3.2.3 was applied. A classification index of zero means no statistical significance for the classification results (i,e. the classification performance is at chance level).

The existence of representations for both auditory and articulatory features in the motor cortex has been discussed in the literature [55, 85, 51]. In this analysis, the nature of temporal propagation for each feature representation in the motor cortex is investigated. Examining the temporal characteristics of each representation will help reveal how and when each representation appears in the motor cortex. Due to the inconsistent electrode locations across subjects, the analysis was presented for each subject individually.

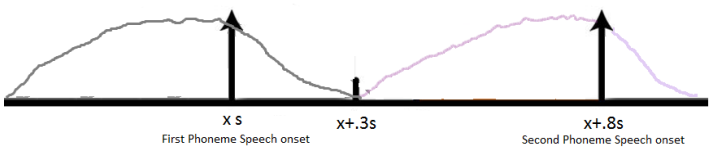


Fig. 4: An exemplary illustration of choosing the stimuli procedure for articulatory and auditory features characterization in the motor cortex. The leftmost arrow represents the onset time x of the first phoneme. A vertical line indicates 300 ms after the onset of the first phoneme. The second arrow represents the onset of the next phoneme. The gray curve represents the neural activities of the first phoneme. The pink curve represents the neural activities of the second phoneme. It is illustrated that the neural activities of speech perception for the first phoneme is effectively diminished after 300 ms and the neural activities of speech production for the next phoneme begins.

Characterization

ECoG data used in characterization of the temporal propagation were chosen to exclude the influence of auditory feedback. If two consecutive phonemes have a less than 800 ms temporal difference based on their onset, then the data corresponding to the second label in the sequence were excluded from the analysis. The 800 ms interval was chosen based on the work of Brumberg *et al.* (2016) [38], where they showed that after 300 ms from the start of the uttering, the perception speech-related activities begin to diminish. Furthermore, they showed that neural activity for speech preparation starts approximately 500 ms before speech onset. Based on their analysis, an 800 ms temporal difference between two consecutive phonemes will ensure that the auditory feedback from previous phoneme utterance has concluded before speech production of the next phoneme. Fig 4 provides an illustration of this concept.

Windows of length 300 ms and 600 ms were used for features extraction as described in Section 3.2.2. The 300 ms window's length was chosen to be half of the window's length

used in classification and approximately half of the window's length used in some of the most related work in the literature [85, 74]. Choosing a shorter window would presumably limit the amount of the information obtained about classes, but it would also increase the temporal resolution. The 600 ms window's length based analysis is necessary to compare the 300 ms based analysis. In other words, the 300 ms based analysis provides a more temporally-localized characterization and the 600 ms analysis provides richer information (technically speaking, higher classification indexes) about the classes. The 300 ms and 600 ms windows started at 500 ms prior to the phoneme onset and ended 200 ms prior to the onset and 100 ms after onset, respectively. Both windows were shifted by 50 ms and for each shift the classification index procedure 3.2.3 was applied. Higher activation indices indicate model outputs that are significantly different from random chance. That is, the model is able to capture meaningful structure related to the speech activity.

Because the shuffling procedure may affect the quality of the trained models, to ensure that the activation indices are valid, the classification index procedure 3.2.3 was repeated m times. The approximated classification indexes were calculated by averaging the resulting classification indexes of the m trials. The 95% confidence interval for the estimated mean was computed using student's distribution since the true standard deviation is unknown. The confidence interval tells the range of possible classification indexes generated by the estimation process, that would contain the true value of these classification indexes with a probability of 95%. This can be considered as the error bar of the estimation. The confidence interval of true mean μ is given by

$$\bar{\mu} \pm t_{\alpha/2}\sigma/\sqrt{n}$$

where $\bar{\mu}$ is the estimated mean of the population, $t_{\alpha/2}$ is obtained from student's distribution, where it was chosen to be at .05 confidence interval, n is the number of samples, σ is the standard deviation of the samples.

3.2.5 CONTINUOUS SPEECH DATA ANALYSIS

Feature Extraction and Characterization

After pre-processing signals as described in Section 3.2.1, electrodes were inspected and those outside of the temporal lobe were excluded from the analysis. This was done by visual inspection. Then, the Hilbert transform of the gamma-band filtered signal was computed. Also, the Hilbert transform was computed for speech signals. The Hilbert transform is used to derive the analytical signal, which is expressed in terms of real and imaginary parts. Taking the absolute value of the analytical signal gives the instantaneous envelope of the signal. As a final step, the gamma-band filtered signal was low-pass filtered with a cutoff frequency of 8 Hz. Finally, ECoG and speech envelopes were decimated to 100 Hz.

To decode and characterize speech information from ECoG data, the gamma ECoG envelope was used as an input to the characterization/decoding model. However, the output of the model was represented by a modified version of the speech envelope's spectrogram. The short time Fourier transform with 300 ms length Hanning window with 50% overlap was applied on the speech envelope. Then, the absolute value was taken to give the spectrum magnitude. After that, the averages of seven spectral bands were calculated, where each band is defined by spectrum values of each three consecutive integer frequencies starting from 1 Hz to 21 Hz resulting in seven values. The spectral values larger than 21 Hz were

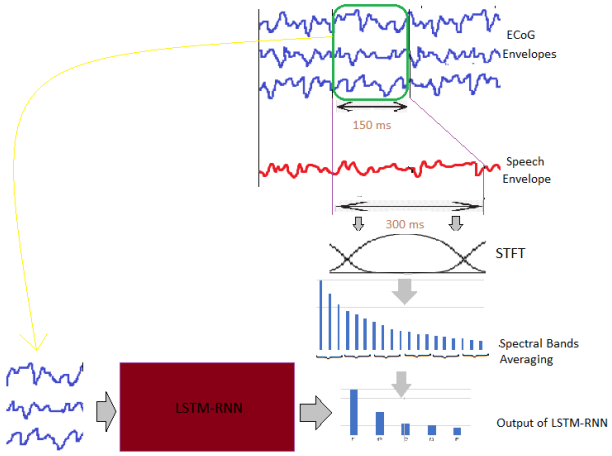


Fig. 5: Characterization procedure of Continuous speech dataset: Each 150 ms of ECoG envelopes is associated with an output, that corresponds to 300 ms speech envelope. LSTM-RNN model was used to map the 150 ms ECoG envelopes to the seven spectral bands within the range [1,21] Hz.

very low in power and do not contribute much in composing the original signal, so they were neglected. The envelope of the gamma-band activity over 150 ms was used to predict this speech signal representation (i.e., seven bands from speech envelope). Fig. 5 illustrates this procedure.

The goal of this analysis is to predict the speech signal representation from the neural activities in the temporal lobe in speech production and comprehension stages. In the production stage, the neural activities in the temporal lobe must lead the speech signal. Thus, both the speech signal and ECoG signals were shifted by the same amount in a way that ECoG signals lead the speech signal. Starting in a position where ECoG signals lead speech signal by 500 ms, this temporal difference was decreased by 100 ms until ECoG and speech signals were synchronized (no shift). In the speech comprehension stage, the same procedure was applied but in a way that ECoG signals lag the speech signal. For each shift,

a decoding/characterization model was built and it was evaluated as will be described in 3.2.5. The best temporal point where the neural activities correlate or represent speech activities is when the model performance reaches its peak.

Decoding Model and Evaluation

The Long-short-term-memory-based recurrent neural network (LSTM-RNN) was used as a model to characterize the relationship between the envelope of ECoG and speech features described above. Since the temporal propagation of neural signals is very important for monitoring and controlling speech (e.g., controlling articulators), it is necessary to model such propagation. LSTM-RNN is an efficient tool to account for the temporal propagation of the input in modeling the input-output relationship. RNNs are well-known in their capability to model time series. However, one of the main drawbacks of traditional RNNs is that they cannot resolve the long-term dependencies in the data. They are not capable of catching subtle information in the short term that would inform the next prediction, in addition to the optimization problem of vanishing and exploding gradients. An LSTM implementation of RNN was capable of compensating for these two issues [86]. The Pearson correlation coefficient between the actual and the predicted output for each frequency band was considered as the evaluation metric. This yields seven different correlation coefficients. The data were randomly divided into training, validation and testing data. The portion of the testing data was 20%, validation data was 10%, and the training data was 70%. This was repeated five times and the final testing correlation was the average of these five times.

In order to measure the amount of information that was learned by this model and to evaluate the statistical significance of the resulting correlation coefficients, the classification

index procedure 3.2.3 was applied such that the output was shuffled across the input trials. The number of iterations n in the classification index procedure was set to 100 because of the time complexity of building an LSTM-RNN model. Furthermore, the procedure of estimating the critical value of the level of chance which was explained in 3.2.4 was applied.

CHAPTER 4

RESULTS

4.1 DECODING AUDITORY AND ARTICULATORY FEATURES

The classification results were normalized to the level of chance, which was approximated by taking the mean of the randomization test results. Since different sets of words were presented for each subject, and any class that has less than 15 instances is excluded to avoid bias, the number of classes for each subject varies. After excluding rare classes, the number of classes for each subject is shown in Table 3. The number of electrodes for each subject is shown in Table 2.

Fig 6 shows the classification results for the three subjects. The consonants classification had the better performance, while places and manners of articulation results were merely the same with respect to each other. All results were statistically significant above the level of chance (p-value $\leq .05$). It is worth mentioning that excluding the 300 ms segment after the speech onset degraded the performance significantly such that all of classification accuracies were no longer statistically significant. This result confirms that neural activity in the motor cortex in the speech perception stage is still strongly related to speech.

TABLE 2: The number of electrodes that are located in the motor cortex for each subject.

Subject	Electrodes number
A	30
B	8
C	20

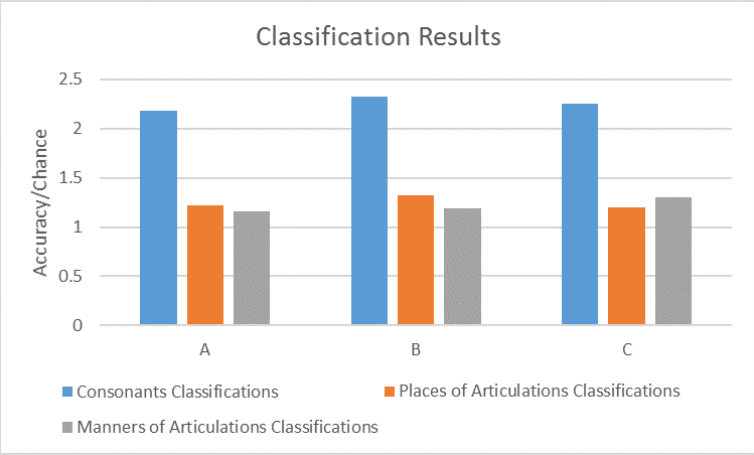


Fig. 6: Subjects A, B, and C Classification Results, the y-axis shows the accuracy per chance.

TABLE 3: Subjects' classes: Flap class was excluded from the manners of articulation classes due to an insufficient number of samples. Also, Laryngeal class was excluded from the places of articulation classes for the same reason. M.of A stands for manners of articulation and P.of A stands for places of articulation.

Subject	Consonants classes number	M.of A classes	P.of A classes
Subject A	10	5	3
Subject B	12	5	3
Subject C	12	5	3

4.2 CHARACTERIZING THE TEMPORAL PROPAGATION OF THE AUDITORY AND ARTICULATORY FEATURES

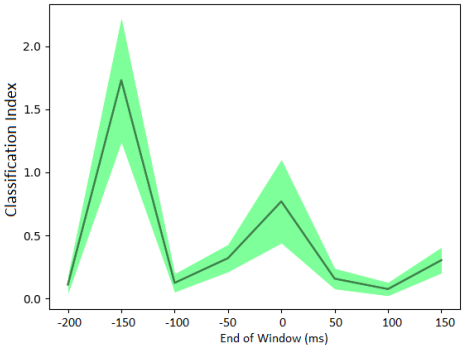
After applying the characterization procedure described in Section 3.2.4, where two consecutive phonemes were excluded if the time difference was less than 800 ms, in addition to applying the classification index procedure on each different temporal parameters, the curves for each subject were plotted. The classification indexes curves are based on two different window lengths: 300 ms and 600 ms.

The analysis started with a time interval, starting from 500 ms prior to the onset for both 300 and 600 ms windows, and ending at 100 ms or 450 ms after the onset in case of 300 ms and 600 ms windows, respectively. Herein, timings before speech onset will be indicated as negative. For all the following characterization figures, the x-axis represents the end of window for each shift in ms.

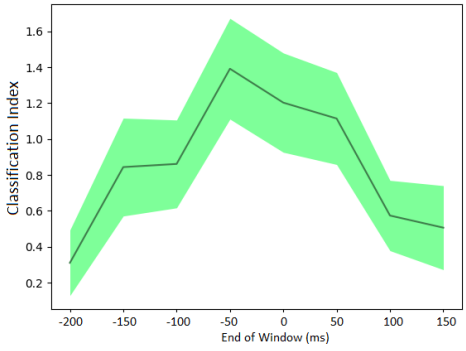
Starting with subject A, the places of articulation classification indexes curve based on a 300 ms window length is shown in Fig. 7b, which indicates a significant positive gain in the performance in the interval [-.35,.05]s which can be described as the interval where the speech production stage is active. The gradual increase and decrease of the classification indexes curve show the consistency of the temporal propagation for this representation. Speech-related information gradually starts to sum up in the speech production stage and then gradually starts to diminish. However, for both consonants and manners of articulation classification indexes in Fig. 7a and Fig. 7c respectively, there are two peaks. The first and the second one occur in the intervals [-450,-150] and [-300,0] respectively for both representations. The curve also shows no consistent decrease or increase through time. It shows that speech-related information in the motor cortex occurs as a discrete burst. Discrete

bursts are not supported by the literature and many studies indicate that speech-related neural activity is continuously modulated over time [38, 74, 79]. This can be interpreted by the window length that was chosen (300 ms), which is too short to capture the information related to these two representations. In other words, the similar curves for both consonants and manners of articulation classification may reflect time-broad representations that need a longer time window to be captured. Moreover, the wide confidence interval around the second peak of manners of articulation characterization compared to the other two curves may indicate that the second peak is superficial.

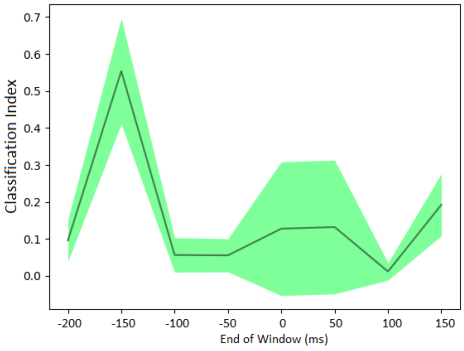
In order to validate the 300 ms-based analysis as well as to test the effect of the window's length, the analysis was repeated with the 600 ms window's length, which is the same window length of the classification analysis. Since the classification results were statistically significant using this window's length, it is reasonable to speculate a higher classification index, at least in the [-300,300] interval. Fig 8 shows that classification indexes were improved as it was speculated. This points to the prolonged-time window of neural activity which is needed to capture the features. The consonants representation's maximum classification index was found in the interval [-200,400] which is shown in Fig. 8a. For the places of articulation which is shown in Fig 8b, the classification indexes reached their maximum in the time intervals [-300,300]. In case of the manners of articulation, the peak of classification index is around [-150,450], where an increase occurs as the window goes to the speech perception stage. Nevertheless, the values of the classification indexes for manners of articulation are very small compared to the other two representations.



(a) Consonants Classification Indexes

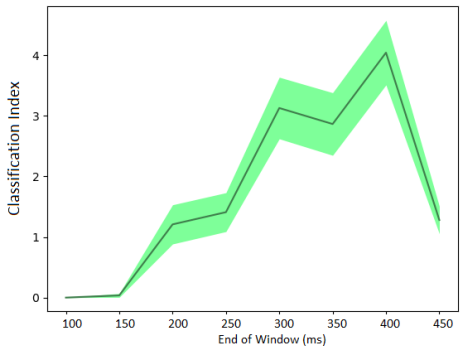


(b) Places of articulation Classification Indexes

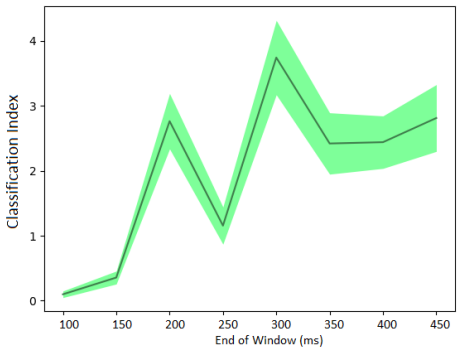


(c) Manners of Articulation Classification Indexes

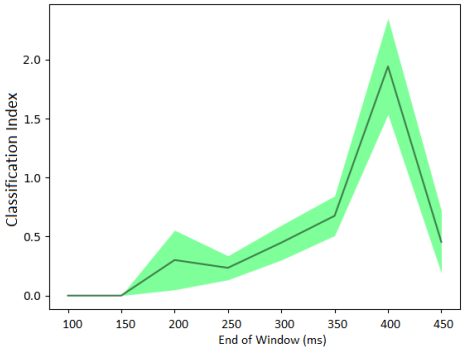
Fig. 7: The estimated mean of the classification indexes for all articulatory and auditory features of subject A using 300 ms window's length. The shaded area represents the 95% confidence interval. The X-axis represents the shift by 50 ms of the features window. The Y-axis represents the classification index.



(a) Consonants Classification Indexes

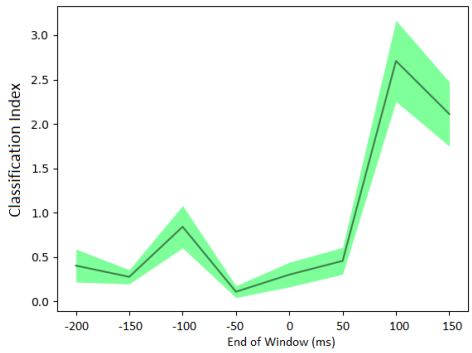


(b) Places of articulation Classification Indexes

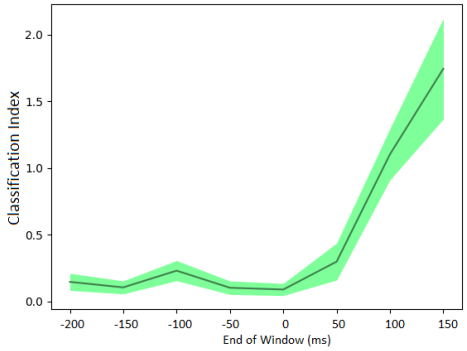


(c) Manners of Articulation Classification Indexes

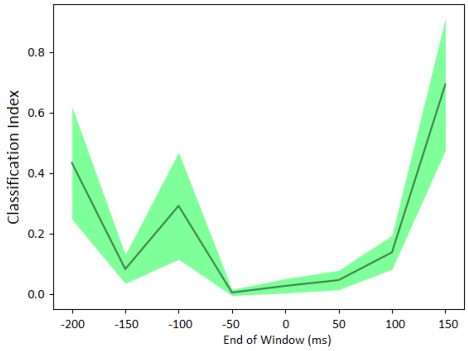
Fig. 8: The estimated mean of the classification indexes for all articulatory and auditory features of subject A using 600 ms window. The shaded area represents the 95% confidence interval. The X-axis represents the shift by 50 ms of the features window. The Y-axis represents the classification index.



(a) Consonants Classification Indexes



(b) Places of articulation Classification Indexes



(c) Manners of Articulation Classification Indexes

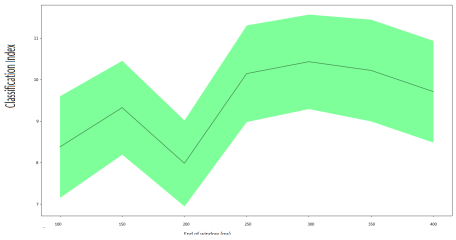
Fig. 9: The estimated mean of the classification Indexes for all articulatory and auditory features of subject C using 300 ms window. The shaded area represents the 95% confidence interval. The X-axis represents the shift by 50 ms of the features window. The Y-axis represents the classification index.

The time-propagation of the classification indexes for subject C based on the 300 ms window is shown in Fig. 9. The manners of articulation curve, in Fig. 9c, shows a very small peak with a wide confidence interval for classification index in the early stage of speech production, then the classification index reached a maximum in the interval [-150,150]. However, the values of the classification indexes are not high compared with the other two plots.

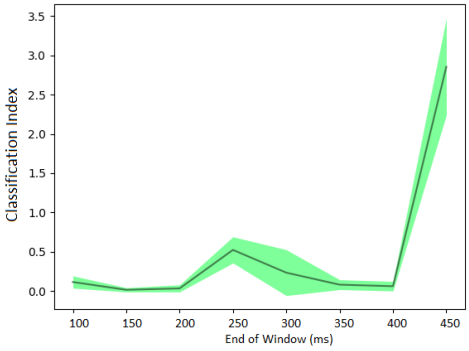
On the other hand, when increasing the window's length to 600 ms, shown in Fig. 10c,

the performance was enhanced and reached its maximum in the interval [-400,200], then decreased and remained constant with a smooth oscillation. The steady state of the curve (smooth oscillations) indicates that no information is gained or lost by shifting the window farther. The places of articulation classification indexes based on the 300 ms window's length are shown in Fig. 9b, which shows an increase in the classification indexes all along the interval of [-300,1500] and the maximum classification index occurred in the interval [-150,150]. When increasing the window's length to 600 ms, the places of articulation classification indexes curve did not change as is shown in Fig. 10b. The consonants classification indexes based on 300 ms window's length is shown in Fig.9a. This curve shows no significant classification index except in the interval [-150,150]. Nonetheless, after increasing the window's length to 600 ms, the consonants classification indexes curve, which is shown in Fig.10a, shows a high classification index in the interval [-300,300] and the performance remains merely constant up to the interval [-100,400], which indicates that no information is gained or lost when shifting the window from [-300,300] to [-100,400].

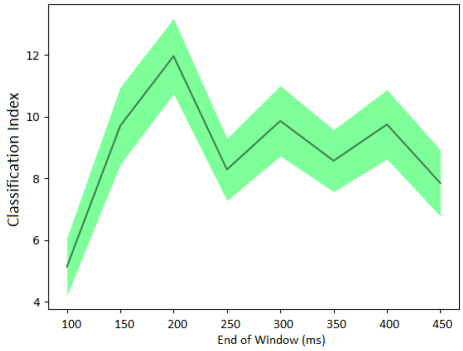
The analysis of the subject B based on the 300 ms window's length is shown in Fig. 11. Consonants classification index based on the 300 ms window's length curve in Fig. 11a is maximized during the interval [-150,150] which is the closest interval to speech perception. The manners of articulation classification indexes based on the 300 ms window's length curve is shown in Fig. 11c, which has a behavior similar to the consonants curve, where it is maximized in the interval [-150,150]. However, the values of manners of articulation classification indexes are much less than consonants. Places of articulation classification indexes based on the 300 ms window's length curve is shown in Fig 11b, where it starts to increase



(a) Consonants Classification Indexes



(b) Places of articulation Classification Indexes

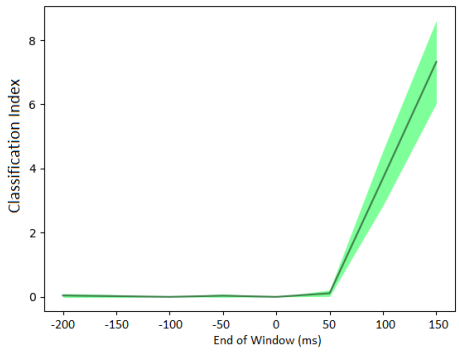


(c) Manners of Articulation Classification Indexes

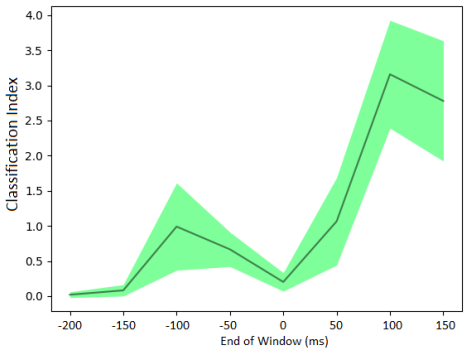
Fig. 10: The estimated mean of the classification indexes for all articulatory and auditory features of subject C using 600 ms window. The shaded area represents the 95% confidence interval. The X-axis represents the shift by 50 ms of the features window. The Y-axis represents the classification index.

in the interval [-400,-100] up to the interval [-200,100] which includes a speech-perception related activity. To further investigate the effect of the speech-perception stage, the analysis was extended to the 600 ms window's length which is shown in Fig. 12. The manners of articulation classification indexes curve is shown in Fig. 11c, where it is maximized in the interval [-250,350]. The places of articulation classification indexes curve, which is shown in Fig. 12b, has a consistent increase starting from the interval of [-500, 100] and then it is maximized in the interval [-.3,.3], and after that, the curve starts to decrease. The consonants classification indexes curve is shown in Fig. 12a, which has a consistent increase starting from [-500,100] and then it is maximized in the interval [-250,350], then it is followed by a decrease. The decrease of both consonants and places of articulation curves tells that shifting the window farther after the classification index-maximized interval causes loss of information.

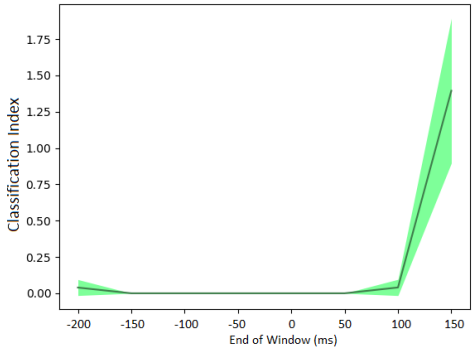
Based on the extended analysis of each subject, the temporal characterization of the auditory features, which are represented by consonants, based on the 300 ms window length shows a higher classification index as temporal-parameters of the features go toward the speech perception stage. Furthermore, the temporal characterization based on the 600 ms window length supports this conclusion, that the speech-perception stage is more related to auditory features than speech-production. Neither of the articulatory features representations, however, show consistent temporal classification indexes across subjects for both representations. For instance, subject B showed high manners of articulation temporal classification indexes but also showed weak temporal classification indexes for the places of



(a) Consonants Classification Indexes

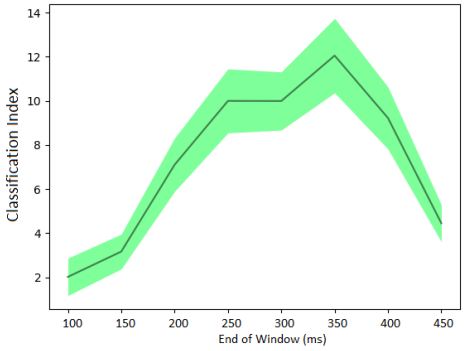


(b) Places of articulation Classification Indexes

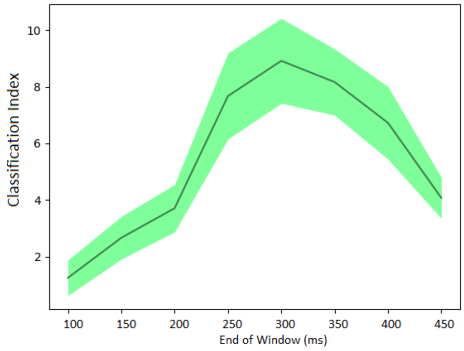


(c) Manners of Articulation Classification Indexes

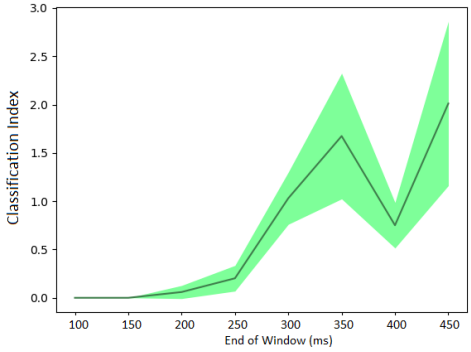
Fig. 11: The estimated mean of the activation Indexes for all articulatory and auditory features of subject B using 300 ms window. The shaded area represents the 95% confidence interval. The X-axis represents the shift by 50 ms of the features window. The Y-axis represents the activation index.



(a) Consonants Classification Indexes



(b) Places of articulation Classification Indexes

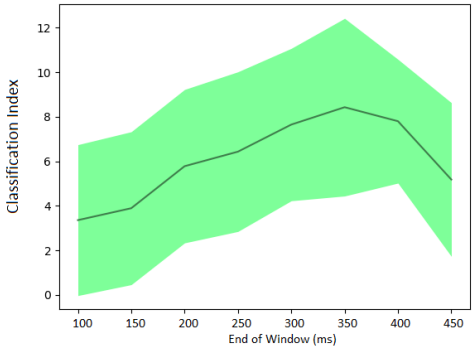


(c) Manners of Articulation Classification Indexes

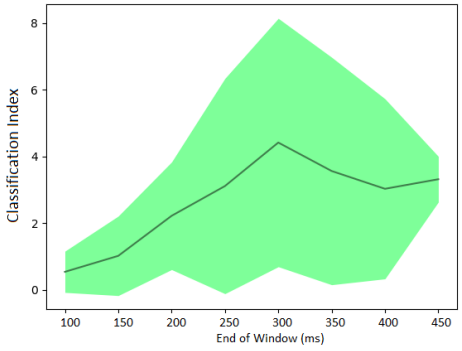
Fig. 12: The estimated mean of the classification Indexes for all articulatory and auditory features of subject B using 600 ms window. The shaded area represents the 95% confidence interval. The X-axis represents the shift by 50 ms of the features window. The Y-axis represents the classification index.

articulation. In general, the articulatory features and auditory features show higher classification indexes in the case of the 600 ms window's length compared to the 300 ms window's length. In other words, both representations are distributed in a prolonged-time window and they are represented in a time interval larger than 300 ms. Nevertheless, the 600 ms window seems to be too long since the curves usually showed a steady-state classification index when shifting the window farther, for instance, subject C consonants (10c) and manners of articulation (10a) classification indexes curves, in addition to subject A places of articulation classification indexes curve (8b). Further analysis must be done to capture the best window length for each representation.

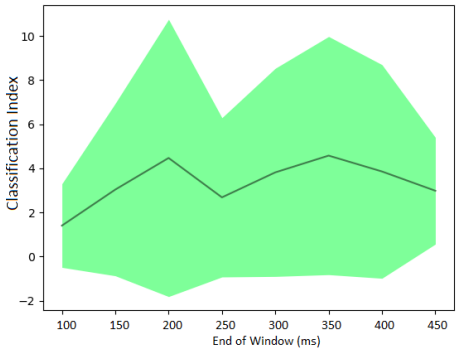
In general, the temporal characterization of the auditory features has a more consistent classification indexes curve. On the other hand, the temporal characterization of the articulatory features, which are represented by place and manners articulations, varied across subjects. More specifically, subjects A and C have consistent classification indexes for the places of articulation, but subject B has more consistent classification indexes for the manners of articulation. Fig. 13 shows the average and standard deviation of all representations across the three subjects. The Consonants curve, shown in Fig. 13a shows that the classification indexes increase as the time window is shifted toward speech perception. Places and manners of articulation show a high standard deviation since the classification indexes for these two representations significantly varied across subjects. However, since places of articulation were better represented for two subjects and manners of articulation were better represented for a single subject, the standard deviation of the places of articulation is less than the standard deviation of manners of articulation across subjects.



(a) Average and Standard Deviation of Consonants Classification Indexes across Subjects



(b) Average and Standard Deviation of Places of Articulation Classification Indexes across Subjects



(c) Average and Standard Deviation of Manners of Articulation Classification Indexes across subjects

Fig. 13: The estimated mean of the classification Indexes for all articulatory and auditory features of subject B using 600 ms window. The shaded area represents the 95% confidence interval. The X-axis represents the shift by 50 ms of the features window. The Y-axis represents the classification index.

TABLE 4: The best time interval where each auditory and articulatory features classification indexes were maximized in ms

Subject	Articulatory Features	Auditory Features
A	[-300,300]	[-200,400]
B	[-400,200]	[-250,350]
C	[-300,300]	[-250,350]

Nonetheless, the final conclusion about the maximum classification indexes based on the temporal characteristics is provided in Table 4, which shows the time interval for each subject in which the classification indexes for each representation (i,e. auditory and articulatory representations) was maximized. Since the 600 ms window's length showed a better performance, the maximum classification index interval was set based on the 600 ms window's length analysis. Furthermore, since the manners of articulation are better represented for subject C and places of articulation are better represented for both subjects A and B, the manners of articulation were chosen as an articulatory features representative for subject C and places of articulation were chosen as a representative for subjects A and B.

Based on Table 4, the auditory and articulatory features exist in both stages: production and perception. Moreover, the articulatory features show up earlier than the auditory features by [50-150] ms. Table 4 also suggests that although both auditory and articulatory features are represented in both speech production and perception, the auditory features tend to be dominant in the speech perception stage.

4.3 MODELING SPEECH-RELATED NEURAL ACTIVITIES IN THE TEMPORAL LOBE

The electrodes located on the temporal lobe were visually selected. Table 5 shows the

number of the selected electrodes for each subject. Most of the selected electrodes of subject A were located at the inferior part of the temporal lobe, which is distant from the auditory cortex. Subject D has the best coverage of the temporal lobe, especially the auditory cortex. The critical value $\alpha_{.05}$ of the chance level Pearson correlation coefficient was estimated in order to have a simple and easy way to read results. The $\alpha_{.05}$ of the chance level Pearson correlation coefficient was calculated for each shift, subject, and frequency group as it was demonstrated in Section 3.2.4. That is, $4 \times 11 \times 7$ different critical values were obtained, where the first number refers to the number of subjects, the second refers to the number of shifts, and the third refers to the number of frequencies group. In order to understand how these critical values differ according to their parameters (i.e., subjects, frequency groups, and lags), they were grouped based on their shift. For instance, all critical values of -500 ms lag were best-fitted to a distribution. It was found that all shift-based groups followed a Normal distribution with very close means [.7,.10]. This means that the distributions of these grouped critical values are merely the same. Therefore, all critical values were best-fitted to a distribution, and it was found that they followed a Gaussian distribution with a mean of .09. This means that different parameters (i.e., subjects, frequency groups, and lags) do not affect the values of $\alpha_{.05}$. In other words, the distribution of the possible critical values is merely the same for all subjects, lags, and frequencies group. When having atypical values (e.g. .27), the analysis for that particular point was repeated using larger numbers of iterations since more iterations will indicate whether this atypical value is superficial. The triple or double of the original iteration number was chosen. These abnormal values would be changed to be consistent with the general samples. Based on this, the final critical

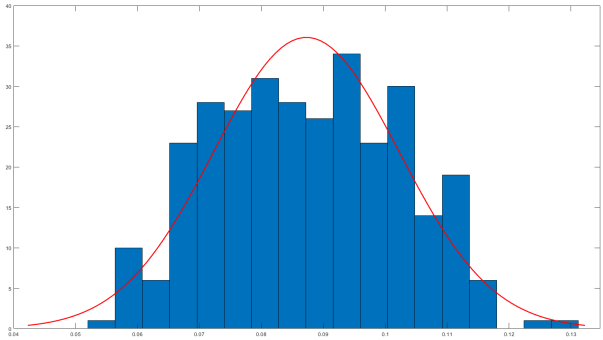


Fig. 14: The Historgram of the critical values obtained from different shifts, frequency groups, and subjects.

value $\alpha_{.05}$ was estimated by fitting all obtained critical values to a Gaussian distribution and then calculating the critical value at .05 based on this distribution. This calculation yields $\alpha_{.05} = .0.1145$, which would be considered one of the largest extreme values that a level of chance correlation could take. Fig 14 shows the distribution of all the estimated critical values.

Fig. 15 shows the mean, which is represented by the black curve, and the standard deviation, which is represented by the gray shaded area, of the correlation coefficients over different lags across the four subjects. The lowest frequency group [1-3]Hz starts from the top, and the higher frequency group [19-21]Hz ends down at the bottom. The x-axis represents the lags starting from -500 ms and up to 500 ms (11 shifts). The red horizontal line represents the critical value $\alpha_{.05}$ of the level of chance. Based on Fig. 15, which shows the Pearson correlation coefficients from -500 ms to 500 ms lags with 100 ms increase, the propagation curves start to increase from -.5 s to a point in the interval [.1,.2] s and then start to decrease where speech-related activity starts to diminish. This observation holds for all frequency groups. Lower frequency groups (frequency groups with higher power)

TABLE 5: The number of electrodes that are mainly located in the temporal lobe for each subject.

Subject	**Electrodes number**
A	43
B	34
C	20
D	70

have stronger correlations than high frequency groups (frequency groups with lower power). Since the variance is very high, this means that the performance varied significantly across subjects. The next section will draw conclusions based on subject D since this subject has the best electrodes coverage and the best performance.

Fig. 16 shows the Pearson correlation coefficients of subject D. Each curve represents a frequency group. The analysis starts at 500 ms before the onset and ends at 500 ms after the onset with an increase of 100 ms. Thus, the continuous curve was interpolated using Matlab software. All values on the curve are statistically significant above the level of chance ($p \leq .001$) except when speech leads ECoG signals by 500 ms (i.e., the last point on the curve).

The power of each frequency group is positively correlated with the ability to be decoded from the temporal lobe neural activities. For instance, the first group (f1 in Fig. 7), which is the mean of values at the integer frequencies power in the interval [1-3], has the best correlation with the actual signal and the second higher correlation is assigned to the second group of frequencies [4-6] and so on.

These curves are very similar to the activation index curve of speech-related activities in the temporal cortex in a study that was done by Brumberg *et al.* (2016), where they used the

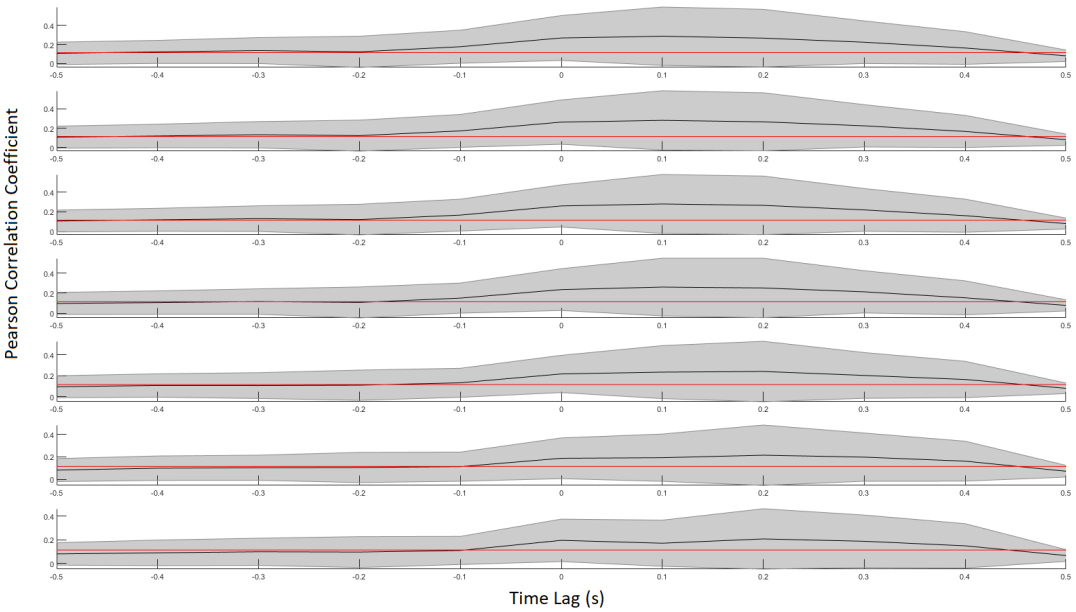


Fig. 15: The Mean and standard deviation of the average correlation coefficient across the four subjects for each frequency group between predicted and the actual output of the testing data. The first group of frequencies [1-3] starts from the top, down to the seventh group of frequencies [19-21]. The red reference line represents the estimated critical value $\alpha_{.05}$.

same data and paradigm. This analysis shows that there is stronger speech-related neural activity in the very early stage of speech production (-500 ms prior) whereas the previous analysis showed there is either no or very weak speech-related activity in the temporal lobe before -220 ms with respect to the speech onset. It can be interpreted by the LSTM-RNN model is able to capture the nonlinear correlations between the ECoG and speech signals since the prior work was based on Pearson correlation coefficients between speech and neural activities. In order to test this hypothesis, a linear regression instead of LSTM-RNN was built. The performance of the linear regression along with the performance of the LSTM-RNN model is shown in Fig. 17. For time lags -200 ms prior, the linear model shows much weaker correlation coefficients compared with LSTM-RNN model. This may indicate the

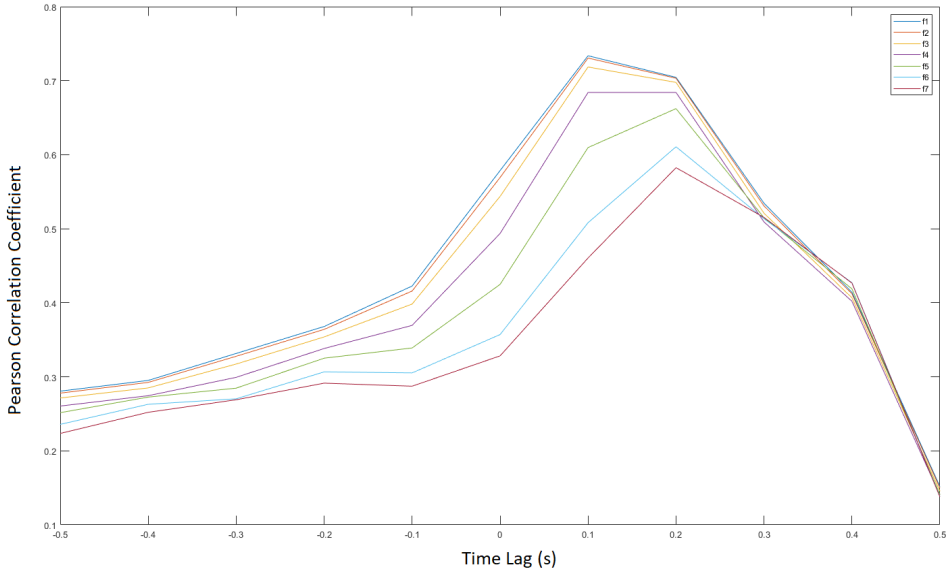


Fig. 16: Decoding the frequencies of speech envelope from subject D temporal lobe using the gamma envelope based on LSTM-RNN model. The x-axis represents the time differences between speech and ECoG signals in seconds, where negative numbers point to when the ECoG signals lead the speech signal and positive numbers point to when speech leads ECoG signals. Y-axis represents the Pearson correlation coefficient between the predicted and true values based on testing data. F1 stands for the mean of the power of the first three integers [1-3], f2 stands for the mean of the power of the next three integers [4-6] and so on.

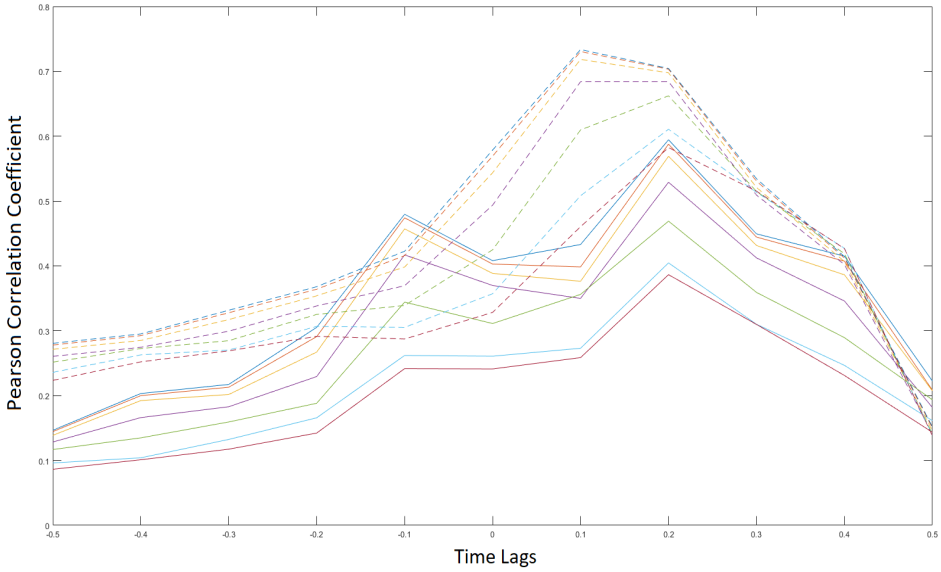


Fig. 17: Linear Model Performance VS LSTM-RNN performance: The testing performance of linear model is shown in solid lines and the testing performance of LSTM-RNN is shown in dashed lines.

nonlinear relationship between the neural activity in the temporal lobe in the very early the speech production stage and the speech activity. Fig. 17 also shows the better performance of LSTM-RNN over the linear model. The previous hypothesis may also be validated by duplicating the work of Brumberg *et al.* using a nonlinear correlation technique instead of the Pearson correlation coefficient. Another interpretation is that, in this work, the spectral power of the speech envelope provides a better representation than the raw envelope used in the prior study.

Also, another analysis was conducted that included all the frequencies [0-50]Hz. The results were similar to the curve in Fig. 16 for high-power frequency components. For low-power frequencies, the correlation was very low and statistically insignificant. Moreover, the analysis of classification indexes was applied but it is presented because the high values

of correlations makes the classification indexes nearly constant, not showing any differences in the decoding capabilities at different lags.

CHAPTER 5

CONCLUSIONS

The first aim of this thesis was to decode the auditory and articulatory features and characterize the temporal propagation of each representation in the motor cortex. The second aim was to decode speech envelop frequencies from the temporal lobe and characterize speech-related activity in the temporal lobe during speech production and perception. This chapter concludes the thesis with a summary of the main contributions and several possible future directions for this research.

5.1 MAIN CONTRIBUTIONS

Places of articulation and manners of articulation were selected as relevant articulatory features and they were successfully decoded (i,e. statistically significant above the chance level) from the motor cortex. Consonants were chosen as relevant auditory features and they were also successfully decoded from the motor cortex. Temporal characterization of auditory features suggests that they are better represented in the speech perception stage and in the very late speech production stage (150-100 ms before the onset). The articulatory features (either places or manners of articulation) appear before the auditory features by 50 to 150 ms. The temporal lobe can provide a predictive model of speech envelope frequencies during speech production and perception.

The Contribution of this Work in Speech-based BCI

In addition to providing new knowledge about the speech-related neural activities for advancing the speech-based BCI, this work can be applied directly in a speech-based BCI system. The two main results of the thesis indicate that, firstly, the articulatory features appear before the auditory features in the motor cortex by 50 to 150 ms, and auditory features are most relevant to the speech perception stage. Secondly, the temporal lobe is able to predict speech information in the production stage. These two results suggest that multiple decisions can be taken from different regions across different time intervals. Combining these decisions will improve the reliability of the BCI system. For instance, a speech-based BCI can detect the articulatory features from the motor cortex and auditory features from the temporal lobe at the very early stage of speech production. After that, the auditory features are detected from both the motor cortex and the temporal lobe in the late speech production and speech perception stage. Finally, these decisions are combined together to reduce the error and maximize the probability that a detection is correct since more knowledge minimizes the error of machine learning models. This means that a speech-based BCI system can be composed of multiple modules, where each one works on a specific feature representation (e.g., articulatory and auditory) from a specific brain region and at a specific time interval. In other words, each module is specialized in the representation-spatial-temporal decoding technique. Such implementation would lead to improving the real-time speech-based BCI system in a way that if the decision in the early production stage is very confident (i.e., probability of error is too low) then this will help to reduce the response time (i.e., time required to issue a command). This thesis contributes to

giving possible prototypes of such modules. Another possible usage is that if a speech-based BCI system is mainly implemented to decode the auditory features (e.g., phonemes), an articulatory features-based BCI system can provide support when the former system is confused, in a way that opposes or supports the decision of the auditory features-based system.

5.2 FUTURE WORK

In this thesis, the representations of speech were chosen to be the power of the frequencies of the speech envelope. However, potentially better representations can be tested, for instance, a representation based on the human auditory system, as presented by Chi *et al.* (2015) [87]. Also, two different window lengths were evaluated, assuming the duration of speech-related activity is greater than 300 ms. On the other hand, 600 ms might be too long. Determining the optimal temporal durations for auditory and articulatory features representations in the motor cortex will help to increase both the ITR and efficiency of a speech-based BCI system.

Both representations of the articulatory features in this thesis were essentially based on the auditory features (i.e., phonemes). However, the articulatory features are related to both the timing of articulators' kinematics and the articulators involved in the uttering. In this work, the choice of articulatory features is based on the articulators involved in uttering but not on the timing of the articulators' kinematics since the collected data does not support measuring the kinematics in an accurate way. However, deep learning speech recognition research has started to explore the possibility of decoding the timings of the articulatory features based on speech signals. Nonetheless, a baseline to validate the output of these deep

networks is not provided, but this seems a promising tool for decoding articulatory features from neural data. Taking the timings of the articulators' kinematics into consideration will represent the articulatory features in a more accurate way, hence, better characterizing the propagation of the articulatory features through time. It is hypothesized in this thesis that the LSTM-RNN is able to predict a nonlinear correlation between ECoG and speech signals. This assumption can be validated by replicating the work of Brumberg *et al.* by using a nonlinear correlation technique.

5.3 DISCUSSION

The best representation for neural signals in the motor cortex during speech production and perception in terms of auditory and articulatory features has been discussed in the literature. However, no conclusive results were presented regarding this research problem. As was discussed in chapter 2, there are reports that have indicated that both representations exist. For instance, Cheung *et al.* (2016) suggested the auditory features are well-represented in the motor cortex while listening and articulatory features exist during speech production [55]. Mugler *et al.* suggested in two studies that the articulatory features are superior over auditory features in the motor cortex [85, 22]. Also, Conant *et al.* (2018) showed that kinematics of articulatory features are well-represented in vSMC which is part of the motor cortex during vowel production [51]. In this thesis, the temporal differences of auditory and articulatory representations in the motor cortex are presented. This thesis also presents the decoding of the auditory and articulatory features of speech based on neural activities in the motor cortex.

The characterization of speech-related neural activities in the temporal lobe during

speech production and perception is presented. While the temporal lobe is well known to be correlated with speech in the perception stage, its role is not well known in the production stage. Since the motor cortex has a role in speech perception, we hypothesized that the temporal lobe also has a role in speech production. Moreover, this hypothesis is supported by a very recent fMRI study by Okada *et al.* (2018) where they concluded that the temporal lobe can provide a predictive model during speech production [50]. These thesis results were consistent with this prior work. However, this was observed in a single subject who had the best coverage for the temporal lobe, especially the auditory cortex. Having subjects with similar electrode coverage would be beneficial to generalize this conclusion to more subjects.

For some subjects, the location of the electrodes on both the motor cortex and the temporal lobe were chosen by visual inspection, which may introduce error. Also, an experimental paradigm of single phoneme/syllable uttering would be more beneficial to be used in characterizing the auditory and articulatory features since the neural activity of each phoneme will be definitely isolated from other phonemes' activities. For the monosyllabic word dataset, there were subjects who were not presented in this analysis since they did not show statistically significant performance.

The results obtained in this thesis will contribute to advancing speech-based BCI and the communication disorders field. It will contribute to providing tools for people who suffer from neuromuscular diseases to improve their quality of life. Moreover, the results of this thesis contribute to the speech neurophysiology field, where it gives a better understanding of speech processes in the brain.

REFERENCES

[1] A. Lécuyer, F. Lotte, R. B. Reilly, R. Leeb, M. Hirose, and M. Slater, "Brain-computer interfaces, virtual reality, and videogames," *Computer*, vol. 41, no. 10, 2008.

[2] J. L. Park, M. M. Fairweather, and D. I. Donaldson, "Making the case for mobile cognition: Eeg and sports performance," *Neuroscience & Biobehavioral Reviews*, vol. 52, pp. 117–130, 2015.

[3] C. G. Lim, T. S. Lee, C. Guan, D. S. S. Fung, Y. Zhao, S. S. W. Teng, H. Zhang, and K. R. R. Krishnan, "A brain-computer interface based attention training program for treating attention deficit hyperactivity disorder," *PloS one*, vol. 7, no. 10, p. e46692, 2012.

[4] G. Schalk and E. C. Leuthardt, "Brain-computer interfaces using electrocorticographic signals," *IEEE Reviews in Biomedical Engineering*, vol. 4, pp. 140–154, 2011.

[5] X.-J. Wang, "Neurophysiological and computational principles of cortical rhythms in cognition," *Physiological reviews*, vol. 90, no. 3, pp. 1195–1268, 2010.

[6] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767 – 791, 2002.

[7] C. Herff, D. Heger, A. De Pesters, D. Telaar, P. Brunner, G. Schalk, and T. Schultz, "Brain-to-text: decoding spoken phrases from phone representations in the brain," *Frontiers in neuroscience*, vol. 9, p. 217, 2015.

[8] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer interfaces for communication and control," *Clinical neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[9] J. J. Vidal, "Toward direct brain-computer communication," *Annual review of Biophysics and Bioengineering*, vol. 2, no. 1, pp. 157–180, 1973.

[10] M. Cheng, X. Gao, S. Gao, and D. Xu, "Design and implementation of a brain-computer interface with high transfer rates," *IEEE transactions on biomedical engineering*, vol. 49, no. 10, pp. 1181–1186, 2002.

[11] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain-computer communication," *Proceedings of the IEEE*, vol. 89, no. 7, pp. 1123–1134, 2001.

[12] E. W. Sellers and E. Donchin, "A p300-based brain–computer interface: initial tests by als patients," *Clinical neurophysiology*, vol. 117, no. 3, pp. 538–548, 2006.

[13] J. Jin, B. Z. Allison, E. W. Sellers, C. Brunner, P. Horki, X. Wang, and C. Neuper, "An adaptive p300-based control system," *Journal of neural engineering*, vol. 8, no. 3, p. 036006, 2011.

[14] Y. Wang, R. Wang, X. Gao, B. Hong, and S. Gao, "A practical vep-based brain-computer interface," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 14, no. 2, pp. 234–240, 2006.

[15] M. Wang, I. Daly, B. Z. Allison, J. Jin, Y. Zhang, L. Chen, and X. Wang, "A new hybrid bci paradigm based on p300 and ssvep," *Journal of neuroscience methods*, vol. 244, pp. 16–25, 2015.

[16] D. J. McFarland, L. A. Miner, T. M. Vaughan, and J. R. Wolpaw, "Mu and beta rhythm topographies during motor imagery and actual movements," *Brain topography*, vol. 12, no. 3, pp. 177–186, 2000.

[17] N. Dronkers and J. Ogar, "Brain areas involved in speech production," 2004.

[18] P. D. Cheney, "Role of cerebral cortex in voluntary movements: A review," *Physical therapy*, vol. 65, no. 5, pp. 624–635, 1985.

[19] X. Chen, Y. Wang, M. Nakanishi, X. Gao, T.-P. Jung, and S. Gao, "High-speed spelling with a noninvasive brain–computer interface," *Proceedings of the national academy of sciences*, vol. 112, no. 44, pp. E6058–E6067, 2015.

[20] E. M. Mugler, J. L. Patton, R. D. Flint, Z. A. Wright, S. U. Schuele, J. Rosenow, J. J. Shih, D. J. Krusienski, and M. W. Slutzky, "Direct classification of all american english phonemes using signals from functional speech motor cortex," *Journal of neural engineering*, vol. 11, no. 3, p. 035015, 2014.

[21] C. Herff, D. Heger, A. De Pesters, D. Telaar, P. Brunner, G. Schalk, and T. Schultz, "Brain-to-text: decoding spoken phrases from phone representations in the brain," *Frontiers in neuroscience*, vol. 9, p. 217, 2015.

[22] E. M. Mugler, M. Goldrick, J. M. Rosenow, M. C. Tate, and M. W. Slutzky, "Decoding of articulatory gestures during word production using speech motor and premotor cortical activity," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pp. 5339–5342, IEEE, 2015.

[23] S. Kellis, K. Miller, K. Thomson, R. Brown, P. House, and B. Greger, "Decoding spoken words using local field potentials recorded from the cortical surface," *Journal of neural engineering*, vol. 7, no. 5, p. 056007, 2010.

[24] X. Pei, D. L. Barbour, E. C. Leuthardt, and G. Schalk, "Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans," *Journal of neural engineering*, vol. 8, no. 4, p. 046028, 2011.

[25] S. Martin, P. Brunner, C. Holdgraf, H.-J. Heinze, N. E. Crone, J. Rieger, G. Schalk, R. T. Knight, and B. N. Pasley, "Decoding spectrotemporal features of overt and covert speech from the human cortex," *Frontiers in neuroengineering*, vol. 7, p. 14, 2014.

[26] J. Kubanek, P. Brunner, A. Gunduz, D. Poeppel, and G. Schalk, "The tracking of speech envelope in the human cortex," *PloS one*, vol. 8, no. 1, p. e53398, 2013.

[27] S. Chakrabarti, D. J. Krusienski, G. Schalk, and J. S. Brumberg, "Predicting mel-frequency cepstral coefficients from electrocorticographic signals during continuous speech production," in *Abstract presented at Proceedings of the Sixth International IEEE/EMBS Neural Engineering Conference, San Diego, CA*, 2013.

[28] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang, "Reconstructing speech from human auditory cortex," *PLoS biology*, vol. 10, no. 1, p. e1001251, 2012.

[29] J. S. Brumberg, A. Nieto-Castanon, P. R. Kennedy, and F. H. Guenther, "Brain–computer interfaces for speech communication," *Speech communication*, vol. 52, no. 4, pp. 367–379, 2010.

[30] S. Martin, C. Mikutta, M. K. Leonard, D. Hungate, S. Koelsch, S. Shamma, E. F. Chang, J. d. R. Millán, R. T. Knight, and B. N. Pasley, "Neural encoding of auditory features during music perception and imagery," *Cerebral Cortex*, pp. 1–12, 2017.

[31] N. J. Hill, D. Gupta, P. Brunner, A. Gunduz, M. A. Adamo, A. Ritaccio, and G. Schalk, "Recording human electrocorticographic (ecog) signals for neuroscientific research and

real-time functional cortical mapping," *Journal of visualized experiments: JoVE*, no. 64, 2012.

[32] D. Wildgruber, H. Ackermann, U. Klose, B. Kardatzki, and W. Grodd, "Functional lateralization of speech production at primary motor cortex: a fmri study.," *Neuroreport*, vol. 7, no. 15-17, pp. 2791–2795, 1996.

[33] J. A. Jones and D. E. Callan, "Brain activity during audiovisual speech perception: an fmri study of the mcgurk effect," *Neuroreport*, vol. 14, no. 8, pp. 1129–1133, 2003.

[34] F. Babiloni, F. Cincotti, F. Carducci, P. M. Rossini, and C. Babiloni, "Spatial enhancement of eeg data by surface laplacian estimation: the use of magnetic resonance imaging-based head models," *Clinical Neurophysiology*, vol. 112, no. 5, pp. 724–727, 2001.

[35] K. Ibayashi, N. Kunii, T. Matsuo, Y. Ishishita, S. Shimada, K. Kawai, and N. Saito, "Decoding speech with integrated hybrid signals recorded from the human ventral motor cortex," *Frontiers in neuroscience*, vol. 12, p. 221, 2018.

[36] W. Penfield and E. Boldrey, "Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation," *Brain*, vol. 60, no. 4, pp. 389–443, 1937.

[37] J. E. Bogen and G. Bogen, "Wernicke's region–where is it?," *Annals of the New York Academy of Sciences*, vol. 280, no. 1, pp. 834–843, 1976.

[38] J. S. Brumberg, D. J. Krusienski, S. Chakrabarti, A. Gunduz, P. Brunner, A. L. Ritaccio, and G. Schalk, "Spatio-temporal progression of cortical activity related to continuous overt and covert speech production in a reading task," *PloS one*, vol. 11, no. 11, p. e0166872, 2016.

[39] M. K. Leonard, M. O. Baud, M. J. Sjerps, and E. F. Chang, "Perceptual restoration of masked speech in human cortex," *Nature communications*, vol. 7, p. 13619, 2016.

[40] C. R. Holdgraf, J. W. Rieger, C. Micheli, S. Martin, R. T. Knight, and F. E. Theunissen, "Encoding and decoding models in cognitive electrophysiology," *Frontiers in systems neuroscience*, vol. 11, p. 61, 2017.

[41] S. Martin, I. Iturrate, J. d. R. Millán, R. T. Knight, and B. N. Pasley, "Decoding inner speech using electrocorticography: progress and challenges toward a speech prosthesis," *Frontiers in neuroscience*, vol. 12, p. 422, 2018.

[42] S. Bouton, V. Chambon, R. Tyrand, A. G. Guggisberg, M. Seeck, S. Karkar, D. van de Ville, and A.-L. Giraud, "Focal versus distributed temporal cortex activity for speech sound category assignment," *Proceedings of the National Academy of Sciences*, vol. 115, no. 6, pp. E1299–E1308, 2018.

[43] D. E. Callan, V. Tsytsarev, T. Hanakawa, A. M. Callan, M. Katsuhara, H. Fukuyama, and R. Turner, "Song and speech: brain regions involved with perception and covert production," *Neuroimage*, vol. 31, no. 3, pp. 1327–1342, 2006.

[44] F. Pulvermüller, M. Huss, F. Kherif, F. M. del Prado Martin, O. Hauk, and Y. Shtyrov, "Motor cortex maps articulatory features of speech sounds," *Proceedings of the National Academy of Sciences*, vol. 103, no. 20, pp. 7865–7870, 2006.

[45] X. Pei, E. C. Leuthardt, C. M. Gaona, P. Brunner, J. R. Wolpaw, and G. Schalk, "Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition," *Neuroimage*, vol. 54, no. 4, pp. 2960–2972, 2011.

[46] K. E. Bouchard, N. Mesgarani, K. Johnson, and E. F. Chang, "Functional organization of human sensorimotor cortex for speech articulation," *Nature*, vol. 495, no. 7441, p. 327, 2013.

[47] K. E. Bouchard and E. F. Chang, "Neural decoding of spoken vowels from human sensory-motor cortex with high-density electrocorticography," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pp. 6782–6785, IEEE, 2014.

[48] A. Flinker, A. Korzeniewska, A. Y. Shestyuk, P. J. Franaszczuk, N. F. Dronkers, R. T. Knight, and N. E. Crone, "Redefining the role of broca's area in speech," *Proceedings of the National Academy of Sciences*, vol. 112, no. 9, pp. 2871–2875, 2015.

[49] S. Martin, P. Brunner, I. Iturrate, J. d. R. Millán, G. Schalk, R. T. Knight, and B. N. Pasley, "Word pair classification during imagined speech using direct brain recordings," *Scientific reports*, vol. 6, p. 25803, 2016.

[50] K. Okada, W. Matchin, and G. Hickok, "Neural evidence for predictive coding in auditory cortex during speech production," *Psychonomic bulletin & review*, vol. 25, no. 1, pp. 423–430, 2018.

[51] D. F. Conant, K. E. Bouchard, M. K. Leonard, and E. F. Chang, "Human sensorimotor cortex control of directly-measured vocal tract movements during vowel production," *Journal of Neuroscience*, pp. 2382–17, 2018.

[52] N. E. Crone, D. Boatman, B. Gordon, and L. Hao, "Induced electrocorticographic gamma activity during auditory perception," *Clinical neurophysiology*, vol. 112, no. 4, pp. 565–582, 2001.

[53] R. T. Canolty, M. Soltani, S. S. Dalal, E. Edwards, N. F. Dronkers, S. S. Nagarajan, H. E. Kirsch, N. M. Barbaro, and R. T. Knight, "Spatiotemporal dynamics of word processing in the human brain," *Frontiers in neuroscience*, vol. 1, p. 14, 2007.

[54] E. F. Chang, C. A. Niziolek, R. T. Knight, S. S. Nagarajan, and J. F. Houde, "Human cortical sensorimotor network underlying feedback control of vocal pitch," *Proceedings of the National Academy of Sciences*, vol. 110, no. 7, pp. 2653–2658, 2013.

[55] C. Cheung, L. S. Hamilton, K. Johnson, and E. F. Chang, "The auditory representation of speech sounds in human motor cortex," *Elife*, vol. 5, p. e12577, 2016.

[56] E. F. Chang, J. W. Rieger, K. Johnson, M. S. Berger, N. M. Barbaro, and R. T. Knight, "Categorical speech representation in human superior temporal gyrus," *Nature neuroscience*, vol. 13, no. 11, p. 1428, 2010.

[57] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, p. 233, 2012.

[58] N. Mesgarani, C. Cheung, K. Johnson, and E. F. Chang, "Phonetic feature encoding in human superior temporal gyrus," *Science*, vol. 343, no. 6174, pp. 1006–1010, 2014.

[59] J. Berezutskaya, Z. V. Freudenburg, U. Güçlü, M. A. van Gerven, and N. F. Ramsey, "Neural tuning to low-level features of speech throughout the perisylvian cortex," *Journal of Neuroscience*, pp. 0238–17, 2017.

[60] L. Riecke, E. Formisano, B. Sorger, D. Başkent, and E. Gaudrain, "Neural entrainment to speech modulates speech intelligibility," *Current Biology*, vol. 28, no. 2, pp. 161–169, 2018.

[61] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS computational biology*, vol. 5, no. 3, p. e1000302, 2009.

[62] N. Ding, A. D. Patel, L. Chen, H. Butler, C. Luo, and D. Poeppel, "Temporal modulations in speech and music," *Neuroscience & Biobehavioral Reviews*, 2017.

[63] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, no. 5, pp. 2670–2680, 1994.

[64] S.-Y. Chang, E. Edwards, N. Morgan, D. Ellis, N. Mesgarani, and E. Chang, "Phone recognition for mixed speech signals: comparison of human auditory cortex and machine performance," tech. rep., Technical Report, 2015.

[65] N. E. Crone, A. Sinai, and A. Korzeniewska, "High-frequency gamma oscillations and human brain mapping with electrocorticography," *Progress in brain research*, vol. 159, pp. 275–295, 2006.

[66] T. Yanagisawa, M. Hirata, Y. Saitoh, T. Goto, H. Kishima, R. Fukuma, H. Yokoi, Y. Kamitani, and T. Yoshimine, "Real-time control of a prosthetic hand using human electrocorticography signals," *Journal of neurosurgery*, vol. 114, no. 6, pp. 1715–1722, 2011.

[67] G. Schalk, K. Miller, N. Anderson, J. Wilson, M. Smyth, J. Ojemann, D. Moran, J. Wolpaw, and E. Leuthardt, "Two-dimensional movement control using electrocorticographic signals in humans," *Journal of neural engineering*, vol. 5, no. 1, p. 75, 2008.

[68] E. Leuthardt, X.-M. Pei, J. Breshears, C. Gaona, M. Sharma, Z. Freudenburg, D. Barbour, and G. Schalk, "Temporal evolution of gamma activity in human cortex during an overt and covert word repetition task," *Frontiers in human neuroscience*, vol. 6, p. 99, 2012.

[69] Y. Liu, W. Zhou, Q. Yuan, and S. Chen, "Automatic seizure detection using wavelet transform and svm in long-term intracranial eeg," *IEEE transactions on neural systems and rehabilitation engineering*, vol. 20, no. 6, pp. 749–755, 2012.

[70] T. Blakely, K. J. Miller, R. P. Rao, M. D. Holmes, and J. G. Ojemann, "Localization and classification of phonemes using high spatial resolution electrocorticography (ecog)

grids," in *Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE*, pp. 4964–4967, IEEE, 2008.

[71] S. Chakrabarti, H. M. Sandberg, J. S. Brumberg, and D. J. Krusienski, "Progress in speech decoding from the electrocorticogram," *Biomedical Engineering Letters*, vol. 5, no. 1, pp. 10–21, 2015.

[72] E. C. Leuthardt, C. Gaona, M. Sharma, N. Szrama, J. Roland, Z. Freudenberg, J. Solis, J. Breshears, and G. Schalk, "Using the electrocorticographic speech network to control a brain–computer interface in humans," *Journal of neural engineering*, vol. 8, no. 3, p. 036004, 2011.

[73] D. Zhang, E. Gong, W. Wu, J. Lin, W. Zhou, and B. Hong, "Spoken sentences decoding based on intracranial high gamma response using dynamic time warping," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pp. 3292–3295, IEEE, 2012.

[74] F. Lotte, J. S. Brumberg, P. Brunner, A. Gunduz, A. L. Ritaccio, C. Guan, and G. Schalk, "Electrocorticographic representations of segmental features in continuous speech," *Frontiers in human neuroscience*, vol. 9, p. 97, 2015.

[75] V. G. Kanas, I. Mporas, H. L. Benz, K. N. Sgarbas, A. Bezerianos, and N. E. Crone, "Joint spatial-spectral feature space clustering for speech activity detection from ecog signals.," *IEEE Trans. Biomed. Engineering*, vol. 61, no. 4, pp. 1241–1250, 2014.

[76] B. Gick, I. Wilson, and D. Derrick, *Articulatory phonetics*. John Wiley & Sons, 2012.

[77] D. A. Moses, N. Mesgarani, M. K. Leonard, and E. F. Chang, "Neural speech recognition: continuous phoneme decoding using spatiotemporal representations of human cortical activity," *Journal of neural engineering*, vol. 13, no. 5, p. 056004, 2016.

[78] D. A. Moses, M. K. Leonard, and E. F. Chang, "Real-time classification of auditory sentences using evoked cortical activity in humans," *Journal of neural engineering*, vol. 15, no. 3, p. 036005, 2018.

[79] J. A. Livezey, K. E. Bouchard, and E. F. Chang, "Deep learning as a tool for neural data analysis: speech classification and cross-frequency coupling in human sensorimotor cortex," *arXiv preprint arXiv:1803.09807*, 2018.

[80] J. O'Sullivan, Z. Chen, J. Herrero, G. M. McKhann, S. A. Sheth, A. D. Mehta, and N. Mesgarani, "Neural decoding of attentional selection in multi-speaker environments without access to clean sources," *Journal of neural engineering*, vol. 14, no. 5, p. 056001, 2017.

[81] H. Akbari, B. Khalighinejad, J. Herrero, A. Mehta, and N. Mesgarani, "Reconstructing intelligible speech from the human auditory cortex," *bioRxiv*, p. 350124, 2018.

[82] G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw, "Bci2000: a general-purpose brain-computer interface (bci) system," *IEEE Transactions on biomedical engineering*, vol. 51, no. 6, pp. 1034–1043, 2004.

[83] A. S. House, C. Williams, M. H. Hecker, and K. D. Kryter, "Psychoacoustic speech tests: A modified rhyme test," *The Journal of the Acoustical Society of America*, vol. 35, no. 11, pp. 1899–1899, 1963.

[84] M. A. Mines, B. F. Hanson, and J. E. Shoup, "Frequency of occurrence of phonemes in conversational english," *Language and speech*, vol. 21, no. 3, pp. 221–241, 1978.

[85] E. M. Mugler, M. C. Tate, K. Livescu, J. W. Templer, M. A. Goldrick, and M. W. Slutzky, "Differential representation of articulatory gestures and phonemes in motor, premotor, and inferior frontal cortices," *bioRxiv*, p. 220723, 2017.

[86] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[87] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

VITA

Hassan Baker
Department of Electrical and Computer Engineering
Old Dominion University
Norfolk, VA 23529

Old Dominion University *December, 2018*
MSc in Electrical and Computer Engineering
Birzeit University, Palestine *June, 2015*
BSc in Computer Systems Engineering

Typeset using LaTeX.