UNIVERSITY OF CALIFORNIA SAN DIEGO

Analysis of intracranial electrophysiology

during vowel production and perception for speech prosthesis design

A Thesis submitted in partial satisfaction of the requirements
for the degree Master of Science

in

Electrical Engineering (Signal & Image Processing)

by

Alexandra MIKHAEL

Committee in charge:

Professor Vikash Gilja, Chair
Professor Gal Mishne
Professor Truong Nguyen

2024

The Thesis of Alexandra MIKHAEL is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

BCI Brain Computer Interface

EEG Electroencephalography

ECoG Electrocorticography

fMRI Functional Magnetic Resonance Imaging

SVM Support Vector Machine

ML Machine Learning

AI Artificial Intelligence

iEEG Intracranial Electroencephalography

sEEG Stereotactic Electroencephalography

BMI Brain Machine Interface

Ctx Cortex

Rh Right Hemisphere

Lh Left Hemisphere

Temp Temporal

Oc Occipital

Cingul Cingular

Ant Anterior

Hippoc Hippocampus

Mid Middle

Parahip Para hippocampus

Lat Lateral

Orb Orbital

Front	Frontal
Inf	Inferior
Sup	Superior
Ins	Insular
Transv	Transversal

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

There are so many people that made this work happen and helped me throughout my years as a master's student.

First, I would like to thank my advisor Professor Vikash Gilja for his guidance and support during my time in his lab. I owe him for making me a better researcher and critical thinker. I hope to apply what he taught me, and work with the same dedication and passion for research as him.

Second, I would like to thank my mentor Professor Massoud Khraiche for his guidance, advice and support through the ups and downs ever since I was a first-year undergraduate student. Thank you for making me the researcher I am today and teaching me how to think critically and scientifically, never to give up on my ideas and to pursue my dreams.

Third, I would like to thank all my colleagues in the Translational Neural Engineering lab, especially Sophia (Jingya) Huang for her support and help on this project. Thank you for helping with theoretical knowledge and always being available. Thank you to my best friend Siddharth who listened to all my research problems and went out with me all around San Diego.

Finally, I would like to thank my parents for their support and unconditional love. They made me the person I am today, and I would not be here without their help and trust. Thank you for believing in me, always pushing me to be the best version of myself and reminding me that I can do anything I set my mind to. Big thanks also to my partner who has been with me through those two years and never stopped supporting me.

VITA

2018-2022	Research Assistant, American University of Beirut
2019	International Honors Program, Stanford University
2021	Summer Research Intern, University of California San Diego
2022	Bachelor of Engineering in Electrical and Computer Engineering with Minor in Biomedical Engineering, American University of Beirut
2022-2024	Graduate Student Researcher, University of California San Diego
2024	Master of Science in Electrical Engineering (Signal & Image Processing), University of California San Diego

ABSTRACT OF THE THESIS

Analysis of intracranial electrophysiology

during vowel production and perception for speech prosthesis design

by

Alexandra MIKHAEL

Master of Science in Electrical Engineering (Signal & Image Processing)

University of California San Diego, 2024

Professor Vikash Gilja, Chair

Speech is the basis of human communication and yet, the neural foundation of speech production and perception are still far from understood. Multiple proof of concept studies have demonstrated the potential to provide a fully closed loop speech prosthesis for people suffering

from anarthria and other speech disabilities. In this thesis, we explore the value of intracranial electrophysiology to capture the variability of neural signals during vowel production and perception for speech prosthesis design. Using stereotactic electroencephalography, we analyze neural recordings to investigate how vowels are encoded in different brain regions and how vowel production correlates to neural activity. To enable these analyses, we first design audio signal processing algorithms and develop metrics to detect speech onset and to extract the first two formant frequencies from an input audio signal. We confirm that these two extracted formant frequencies can be used to uniquely identify vowels in the formant space. Second, we investigate the behavior of neural activity related to speech production by comparing it to that of silence and conclude that there exist significant differences between neural activity during speech and silence. Furthermore, we show that different brain regions respond differently to speech production revealing that there exists a spatially specific modulation relative to vowel production in the brain. This spatially distinct modulation has dynamics that correspond to vowel production onset time. In the final section of this work, we investigate the neural correlation to specific phoneme production and start to implement a classifier to decode vowel identity based on neural activity.

Chapter 1 INTRODUCTION

1.1 Brain Computer Interface

In literature, brain computer interfaces (BCI) are defined as the direct connection between the brain and external devices aimed at enhancing human cognitive and sensory-motor capabilities by translating brain signals into commands. This technique holds promise for circumventing normal pathways that may be dysfunctional due to brain or spinal cord injuries or diseases. BCI usually involves multiple stages for brain-computer interactions. [1-3] The first stage involves signal recording using various techniques such as EEG, ECoG, fMRI, … depending on the application. In the second stage, the recorded signals are processed for feature extraction, dimensionality reduction, and classification. This stage also incorporates pre-processing strategies such as denoising and signal segmentation or trialization. Features can be extracted in either the time domain or frequency domain, depending on the application. Dimensionality reduction methods involve techniques such as Principal Component Analysis (PCA), while classification methods involve Support Vector Machine (SVM), Artificial Neural Networks (ANN) or Hidden Markov Models (HMM). In the third stage, these classified signals are translated into control signals for various devices, interfaces, prosthetics, … .

BCI technologies are advancing rapidly and finding applications in various fields such as healthcare and rehabilitation. Traditional BCIs use motor imagery [4] to control a cursor or to choose between a selected number of options, while others leverage event-related potentials (ERPs) [5] or steady-state evoked potentials [6] to spell out texts. However, traditional BCI face challenges when it comes to accuracy, usability, and efficiency. Emerging trends such as AI and machine learning hold promise for enhancing BCI performance.

1.1.1 Neural Recording Technologies

Advancements in medical devices, nanomaterials and microfabrication techniques have enabled neuroscientists to record from the brain with increasingly higher spatial and temporal resolutions [7]. The miniaturization of the electrodes and the use of novel flexible materials are making these devices more conformable to the brain surface and easier to implant in the cortex, thereby reducing tissue scarring and immune response for long-term studies.

The technologies currently in use for neural recordings advance our understanding of the brain and provide a means to interface machines with humans. They can be broadly categorized into two groups : non-invasive and invasive, as shown in figure 1.1.

Non-invasive BCIs collect information from the brain without requiring brain surgery, utilizing methods such as EEG, Magnetoencephalography (MEG), fMRI and functional near-infrared spectroscopy (fNIRS). Among these, scalp EEG is particularly popular in BCI research due to its attractive features including fine temporal resolution, ease of use, portability, and lower cost.

Invasive neural recording technologies, such as ECoG, iEEG or sEEG, require neurosurgery and clinical care during and after monitoring but are crucial for targeted therapy (e.g., in epilepsy patients) and clinical localization of impaired brain regions [8, 9]. ECoG uses electrodes placed on the brain surface after a craniotomy, whereas sEEG uses burr holes to insert the electrode's shaft into deep regions of the cortex, allowing recordings from different brain regions not accessible by EEG and ECoG.

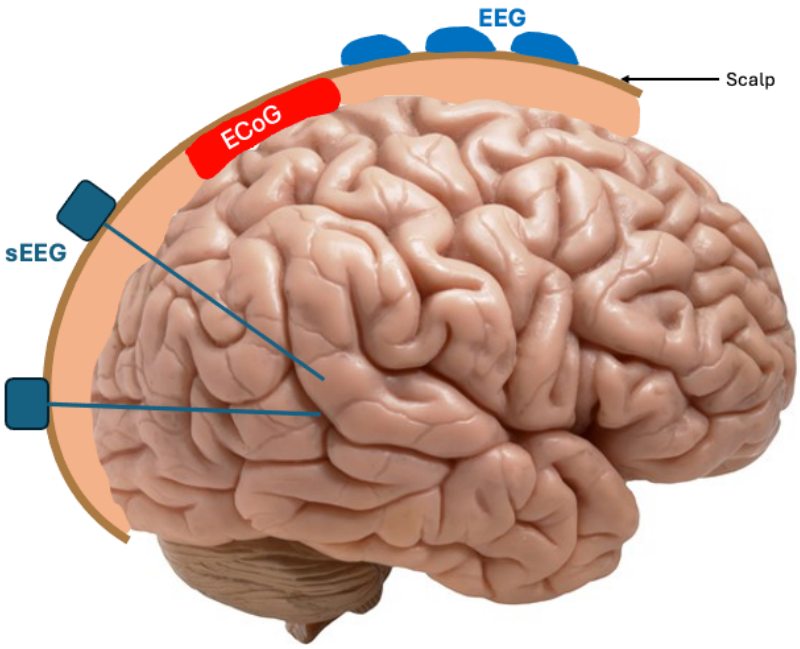

Figure 1-1 : Location of EEG, ECoG and sEEG electrode relative to the brain surface.

Intracranial recordings have offered an unprecedented venue for studying the fundamental neural processes underlying human behavior. Significant advances in human neuroscience have been made at various scales, from single neurons to field potentials, to elucidate diverse human behaviors encompassing perception, action, and cognition. Nevertheless, it is increasingly evident that there are notable gaps in current technologies, not only in electrode density, but also, and just as critically, in the extent of brain coverage and sampling. [10] Depending on the application, the measured signal and the required spatio-temporal resolution, different recording technologies can be used, as depicted in figure 1.2.

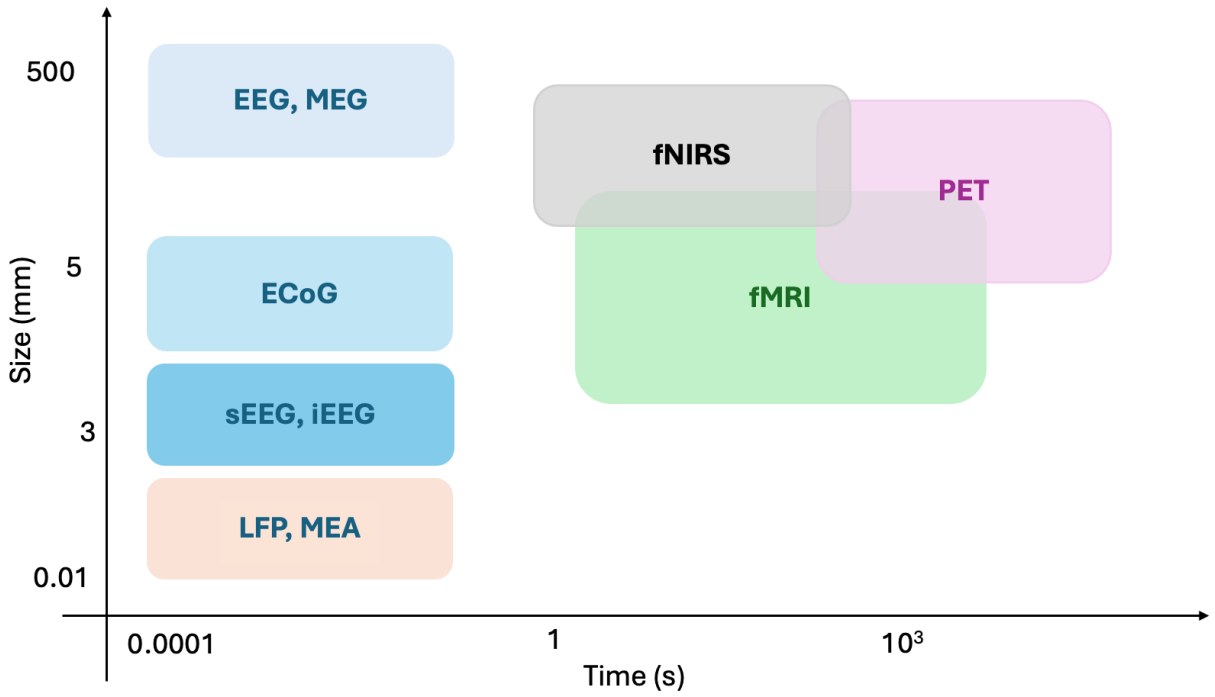

Figure 1-2 : Spatiotemporal domain of neuroscience and of the main methods available for the study of the nervous system.

1.1.2 Stereotactic Electroencephalography

Most recent studies on speech neuroprostheses use ECoG, an invasive neural recording technology that offers high temporal and spatial resolutions, high signal-to-noise ratio [12] and low sensitivity to movement artifacts. Intracortical electroencephalography (iEEG) has also been employed successfully for decoding speech [13, 14] and synthesizing formant frequencies. [15] An alternative method for measuring intracranial neural activity is stereotactic electroencephalography (sEEG), first developed by Talairach and Bancaud in Paris in the late 1950s. [16] Depth electrodes typically used for sEEG have 4–18 contacts spaced 2–10 mm apart and a diameter of 1 mm or less, giving millimeter and sub-millimeter resolution. [17]. These electrodes are either semi-rigid or flexible with a removable rigid stylet for insertion. Electrode shafts are implanted into the brain through small burr holes as shown in figure 1.3 [18,19]. sEEG

is minimally invasive, allowing extensive coverage of both hemispheres without performing a large craniotomy, mainly targeting deep structures with anatomical precision, and obviating the need for a second surgery to remove the electrodes. Similar to ECoG and unlike EEG, sEEG has the capability to identify two crucial aspects of intracranial recordings : broadband gamma activity and low-frequency oscillatory activity. Numerous investigations have indicated that broadband gamma activity (signal amplitude at frequencies exceeding 60 Hz) serves as a reliable marker of cortical activity at the population level, particularly in response to various motor, sensory, or cognitive tasks [20]. The primary drawback of sEEG is its limited functional mapping capabilities due to sparse recordings of contiguous cortical regions. In contrast to ECoG, which provides high-density coverage of specific regions, sEEG offers sparse sampling across multiple regions. This feature holds great potential for various BCI applications, not only because it targets sub-cortical brain regions but also because it allows the simultaneous recording of multiple regions.

sEEG has yielded valuable insights into the distributed neural representations of speech and language functions. By targeting specific lateral, medial, and ventral cortices, recent sEEG studies have been able to integrate data across large cohorts and thoroughly map speech production during cued word production and reading [21-24].

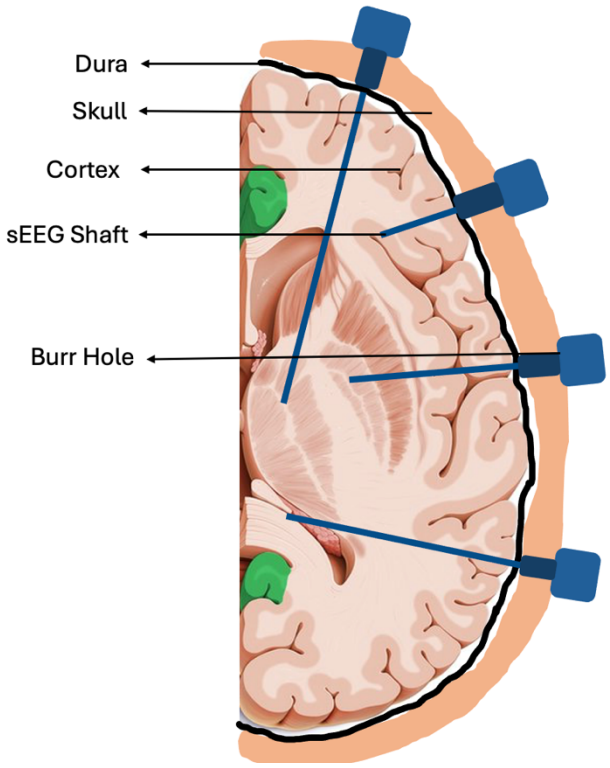

Figure 1-3 : A schema illustrating the concept of sEEG.

1.1.3 Speech Brain Computer Interface

Brain-computer interfaces (BCIs) represent cutting-edge technology with immense promise for restoring speech communication in individuals with speech impairments or disabilities. These interfaces establish a direct link between the brain and external devices, translating neural signals associated with speech production, comprehension, or perception into actionable commands for speech synthesis or recognition systems. By leveraging advances in neuroscience, signal processing, and machine learning, BMIs offer a unique pathway to decode the complex patterns of neural activity underlying speech processes. This interdisciplinary field not only seeks to understand the fundamental mechanisms of brain activity during speech but also aims to develop innovative solutions that can bridge the gap between neural signals and intelligible speech output. Through the integration of neural decoding algorithms and advanced hardware

technologies, speech BMIs hold great potential for revolutionizing assistive technologies and enabling individuals with speech disabilities to communicate effectively and autonomously.

Previous neuroscientific studies provided evidence for neural representations of speech, such as phones and phonetic features during speech perception [25-27] and production [28, 29]. Most studies to date employ a whole-word approach, classifying cortical activation patterns primarily based upon the differences between full words [30]. In a study, Kellis et. al successfully identified at best less than half of ten words in one patient using micro-ECoG electrodes over facial motor cortex [31]. Although these whole-word studies show some preliminary success in speech decoding, these success rates cannot be extrapolated to more complex speech. Moreover, the current most efficient BCI for communication reports information rates of 2.1 bits s^{-1} [32], much lower than the average natural efficiency of human speech production at 25 bits s^{-1} [33].

One strategy to enhance information rates could involve focusing on decoding the smallest isolated segments of speech, known as phonemes. This approach would use phonemes, rather than words, as the 'events' around which to analyze changes in brain signal. Speech BCIs using intracortical recordings to decode phonemes have achieved up to 21% classification success of all phonemes [15]. Similar studies using ECoG succeeded in classifying four [34] and two [35] phonemes, isolated from the context of words. These approaches demonstrate the potential to decode phonemes from cortical signals, which can be used to enhance speech BCIs.

Great advances have been made in the field of speech neuroprostheses where the decoding of a textual representation by decoding phonemes [36,37], phonetic [38] and articulatory [39] features, words [31] and sentences [40-43] is possible from neural recordings during actual speech production.

1.2 Speech Production and Feedback Control

1.2.1 Anatomy of Articulators

Speech is a dynamic process involving both a speaker and a listener. At the physiological level for the speaker, the brain generates electrical signals that activate muscles in the vocal tract and vocal cords. These muscle movements produce pressure changes at the lips, initiating a sound wave that propagates through air to the listener's eardrum. The vibration of the eardrum triggers electric signals that travel along sensory nerves to the listener's brain, where speech recognition and understanding take place. In the listener's speech pathway, feedback through the ear allows for monitoring and correction of one's own speech.

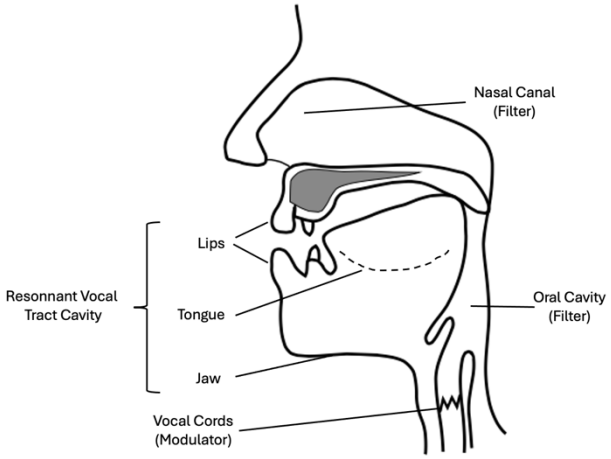

Figure 1-4 : Key speech articulators in the human vocal tract.

Human speech consists primarily of two classes : vowels, also called phonemes, and consonants, each characterized by specific acoustic and spectral properties. During vowel production, air is expelled from the lungs through muscle contraction, leading to airflow through

the vocal cords. This results in the periodic vibration of the vocal cords, whose rate gives the pitch of the sound.

Articulation is well described as a source–filter model of speech production, where the vocal cords (the source) vibrate to generate sound, which is then modulated by the positions of the articulators (the filter) over time to produce phonetic sounds. Different production models can be built for the different phonemes.

Historical studies using functional lesioning and various neuroimaging techniques have mapped the brain areas responsible for processing different aspects of speech and the overall dynamics of their interactions. Indefrey's model [44], delineates speech production into six distinct stages: conceptual preparation, lemma retrieval, phonological code retrieval, phonological encoding, phonetic encoding, and articulation.

This sequence of processes can be conceptualized as the successive conversion of speech information through a continuous "speech pathway." However, in natural speech, this transformation along the pathway is cascading rather than strictly sequential, characterized by significant temporal overlap, parallel processing, and cortical network feedback. Nevertheless, the functional-anatomic compartmentalization of the neural speech pathway suggests that neural speech decoding could intercept this stream at many different points along the pathway with different tradeoffs.

The analysis and presentation of speech signals in the frequency domain are crucial for studying the characteristics and acoustic properties of speech. A significant aspect of the speech signal spectrum comprises formants, which align with the resonant frequencies of the vocal tract and represent peaks in the vowel spectra. These frequencies correspond to where the concentration of acoustic power is the largest. The accurate determination of formant frequencies greatly impacts

the performance of key systems in speech recognition, speech identification, and formant-based speech synthesis. For vowels, the tongue varies in height as the jaw is raised/lowered, which correlates inversely with the center frequency of the lowest-frequency resonance of the vocal tract, called the first formant (F1). (F*i* means the center frequency of the *i* th resonance.) Lateral tongue position affects all formants but correlates primarily with the second formant (F2). To handle possible difficult environmental conditions, the frequency ranges with strongest energy are likely acoustic aspects to aim for, as they stand out amid distortions and have strong correlations to the vocal tract's shape.

Literature has shown that formant values, especially the first and second formants, can be used to uniquely identify vowels, as shown in figure 1.5. [45] The formant values for the vowels vary across languages depending on their phonetic structures.

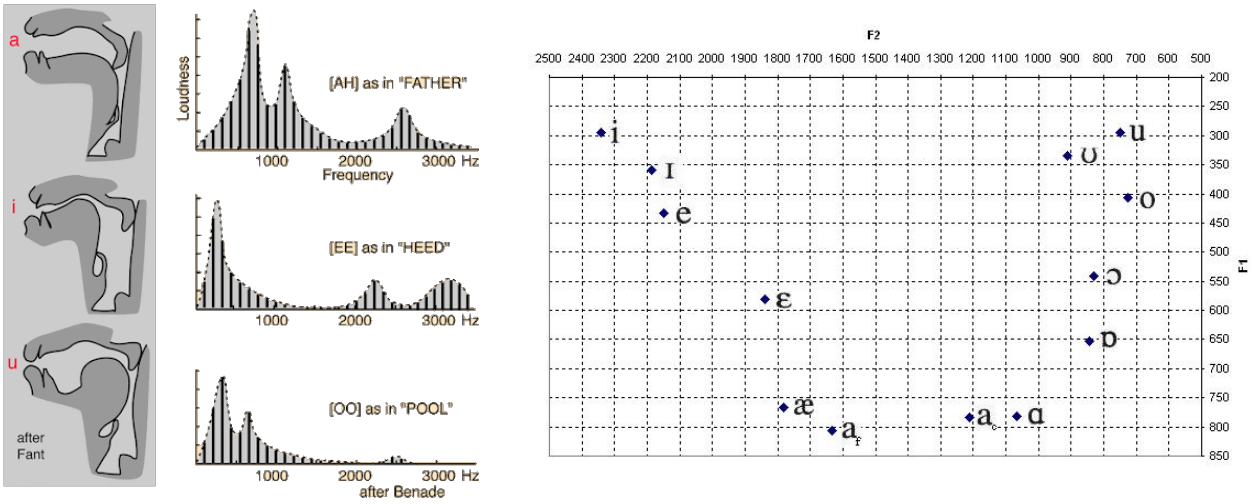

Figure 1-5 : a) Sketches taken from x-rays of the head during the production of the vowels 'a', 'i' and 'u' and b) corresponding vowel spectra [46] c) American English vowels in the F1/F2 formant space. [45]

1.2.2 Feedback Control

The speech network is divided into two separate anatomical streams that arise from the posterior superior temporal gyrus and appear to be specialized in complementary functions [39]. On one hand, the dorsal stream oversees translation of sensory/acoustic speech signals into motor-articulatory representations, known as auditory-motor integration that is required for speech production and verbal repetition [48]. On the other hand, the ventral stream is mainly involved in the mapping of sensory speech signals into conceptual and semantic representations for speech comprehension, as shown in figure 1.6. Neural signals are continuously integrated with sensory feedback, including proprioceptive and auditory information, allowing real-time adjustments in speech production.

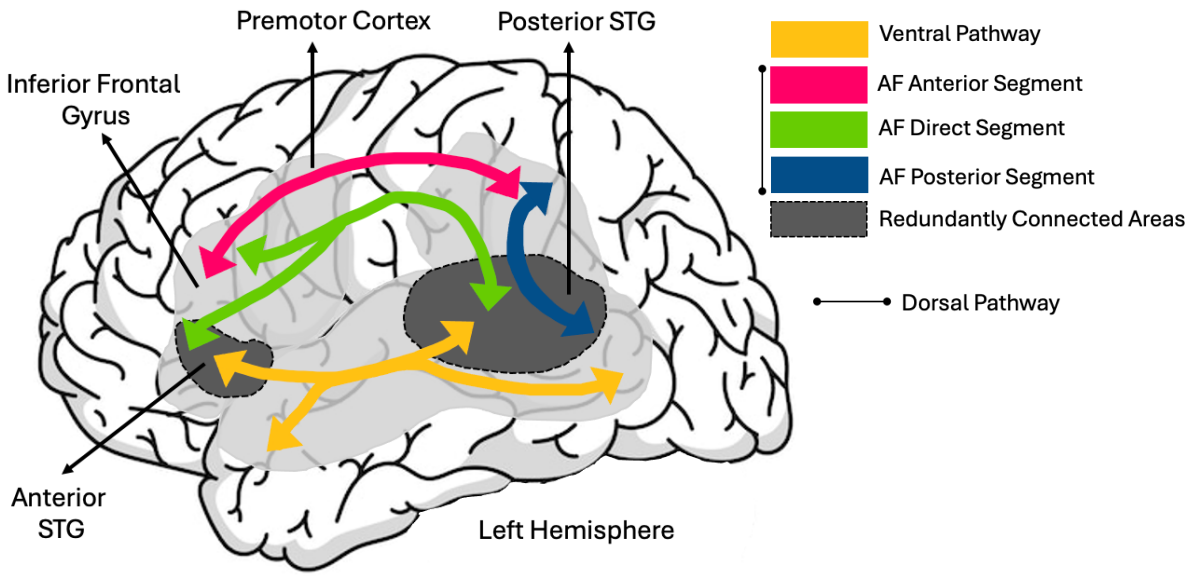

Figure 1-6 : Division of labor of the ventral and dorsal pathways for language.

While most motor acts aim to achieve goals in three-dimensional space (e.g., reaching, grasping, throwing, walking), the primary goal of speech is to transmit an acoustic signal to a

listener via the auditory system. When a movement results in error, the nervous system amends the motor commands that generate the subsequent movement. This error is determined by visual feedback of the subject and plays an important role in monitoring performance and improving skill level [47]. Similarly, auditory feedback plays an important role in monitoring vocal output, achieving verbal fluency and correct speech production. In the early 1900s, Lombard showed evidence of the influence of auditory feedback on speech by conducting a study demonstrating that speakers modify the intensity of their speech in noisy environments [52]

Auditory feedback influences human speech production, as demonstrated by studies using rapid pitch and loudness changes [53]. Feedback has also been investigated using the gradual manipulation of formants for speech perception tasks [54]. These compensatory responses act to steer vocal output closer to the intended auditory target. Understanding the complex interplay of neural signals in speech production is fundamental for exploring speech disorders, enhancing interventions, and developing neuroprostheses technologies aimed at improving communication abilities.

1.3 Thesis Overview

We hope that this thesis offers an understanding for people who want to investigate phoneme and speech representation in neural data. We separate the body of this thesis into two main chapters investigating the following :

In Chapter 2, we show that the formants can be extracted from audio signals using Linear Predictive Coding. These formants can then be used to uniquely identify each vowel in the formant-space.

In Chapter 3, we start investigating the neural correlates of speech production. We start by showing that the brain states corresponding to periods of speech and silence are different. Then,

based on these results, we start investigating what are the specific neural signals and brain regions involved in the perception/production of speech sounds, and how do these signals vary across different vowels. For that we used the Hilbert transform and z-score to show different behaviors around speech onset for the different vowels across the brain. Lastly, we start implementing a State-Vector Machine (SVM) to try and decode the vowels based on the neural activity.

REFERENCES

[1] Ramadan RA, Vasilakos AV. Brain computer interface: control signals review. Neurocomputing. 2017;223:26–44

[2] Ramadan RA, Refat S, Elshahed MA, Ali RA. (2015). Basics of brain computer interface. Brain-Computer Interfaces: Current Trends and Applications. 31–50

[3] Chen X, Huang Y, Zhuang S. Current perspective of brain-computer Interface Technology on mild cognitive impairment. Highlights in Science Engineering and Technology. 2023;36:73–8

[4] McFarland, D. J., Miner, L. A., Vaughan, T. M., and Wolpaw, J. R. (2000). Mu and beta rhythm topographies during motor imagery and actual movements. *Brain Topogr*. 12, 177–186. Doi: 10.1023/A:1023437823106

[5] Farwell, L. A., and Donchin, E. (1988). Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalogr. Clin. Neurophysiol*. 70, 510–523. Doi: 10.1016/0013-4694(88)90149-6

[6] Sutter, E. E. (1992). The brain response interface: communication through visually induced electrical brain responses. *J. Microcomput. Appl*. 15, 31–45. Doi: 10.1016/0745-7138(92)90045-7

[7] Viventi, J., Kim, D.H., Vigeland, L., Frechette, E.S., Blanco, J.A., Kim, Y.S., Avrin, A.E., Tiruvadi, V.R., Hwang, S.W., Vanleer, A.C. and Wulsin, D.F., 2011. Flexible, foldable, actively multiplexed, high-density electrode array for mapping brain activity in vivo. *Nature neuroscience*, *14*(12), pp.1599-1605.

[8] Scott B Wilson, Christine A Turner, Ronald G Emerson, and Mark L Scheuer. Spike detection ii: automatic, perception-based detection and clustering. Clinical neurophysiology, 110(3):404–411, 1999

[9] Osorio, I., Frei, M.G., Giftakis, J., Peters, T., Ingram, J., Turnbull, M., Herzog, M., Rise, M.T., Schaffner, S., Wennberg, R.A. and Walczak, T.S., 2002. Performance reassessment of a real-time seizure-detection algorithm on long ECoG series. *Epilepsia*, *43*(12), pp.1522-1535.

[10] Engel, A.K., Moll, C.K.E., Fried, I., and Ojemann, G.A. (2005). Invasive recordings from the human brain: clinical insights and beyond. Nat. Rev. Neurosci. 6, 35–47

[11] Sejnowski, T., Churchland, P. & Movshon, J. Putting big data to good use in neuroscience. *Nat Neurosci* 17, 1440–1441 (2014). https://doi.org/10.1038/nn.3839

[12] Parvizi, J. & Kastner, S. Promises and limitations of human intracranial electroencephalography. *Nature neuroscience* 21, 474–483 (2018)

[13] Stavisky, S. D., Rezaii, P., Willett, F. R., Hochberg, L. R., Shenoy, K. V., & Henderson, J. M. (2018, July). Decoding speech from intracortical multielectrode arrays in dorsal "arm/hand areas" of human motor cortex. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 93-97). IEEE.

[14] Stavisky, Sergey D., Francis R. Willett, Donald T. Avansino, Leigh R. Hochberg, Krishna V. Shenoy, and Jaimie M. Henderson. "Speech-related dorsal motor cortex activity does not interfere with iBCI cursor control." *Journal of neural engineering* 17, no. 1 (2020): 016049.

[15] Brumberg, J., Wright, E., Andreasen, D., Guenther, F. & Kennedy, P. Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech motor cortex. *Frontiers in Neuroscience* 5, 65, https://doi.org/10.3389/fnins.2011.00065 (2011)

[16] Bancaud, J. (1959). Apport de l'exploration fonctionnelle par voie stéréotaxique à la chirurgie de l'épilepsie. *Neurochirurgie* 5, 55–112

[17] Herff, C., Krusienski, D. J. & Kubben, P. The potential of stereotactic-EEG for brain-computer interfaces: Current progress and future directions. *Frontiers in Neuroscience* 14, 123 (2020)

[18] van der Loo, Lars E., Olaf EMG Schijns, Govert Hoogland, Albert J. Colon, G. Louis Wagner, Jim TA Dings, and Pieter L. Kubben. "Methodology, outcome, safety and in vivo accuracy in traditional frame-based stereoelectroencephalography." *Acta neurochirurgica* 159 (2017): 1733-1746.

[19] Iida, K. & Otsubo, H. Stereoelectroencephalography: indication and efficacy. *Neurologia medico-chirurgica* 57, 375–385 (2017)

[20] Gaona, Charles M., Mohit Sharma, Zachary V. Freudenburg, Jonathan D. Breshears, David T. Bundy, Jarod Roland, Dennis L. Barbour, Gerwin Schalk, and Eric C. Leuthardt. "Nonuniform high-gamma (60–500 Hz) power changes dissociate cognitive task and anatomy in human cortex." *Journal of Neuroscience* 31, no. 6 (2011): 2091-2100.

[21] Forseth K J, Kadipasaoglu C M, Conner C R, Hickok G, Knight R T and Tandon N 2018 A lexical semantic hub for heteromodal naming in middle fusiform gyrus Brain 141 2112–26

[22] Forseth K J, Pitkow X, Fischer-Baum S and Tandon N 2021 What the brain does as we speak bioRxiv preprint (https:// doi.org/10.1101/2021.02.05.429841v1) (Accessed 17 August 2022)

[23] Murphy E, Forseth K J, Donos C, Rollo P S and Tandon N 2022 The spatiotemporal dynamics of semantic integration in the human brain bioRxiv Preprint (https://doi.org/ 10.1101/2022.09.02.506386v1) (Accessed 13 September 2022)

[24] Woolnough O, Donos C, Curtis A, Rollo P S, Roccaforte Z J, Dehaene S, Fischer-Baum S and Tandon N 2022 A spatiotemporal map of reading aloud J. Neurosci. 42 5438–50

[25] F. Edward, Chang, J.W. Rieger, K. Johnson, M.S. Berger, N.M. Barbaro, R.T. Knight, Categorical speech representation in human superior temporal gyrus. Nat. Neurosci. **13**(11), 1428–1432 (2010)

[26] J. Kubanek, P. Brunner, A. Gunduz, D. Poeppel, G. Schalk, The tracking of speech envelope in the human cortex. PloS ONE **8**(1), e53398 (2013)

[27] N. Mesgarani, C. Cheung, K. Johnson, E.F. Chang, Phonetic feature encoding in human superior temporal gyrus. Science 1245994 (2014)

[28] M. Fukuda, R. Rothermel, C. Juhász, M. Nishida, S. Sood, E. Asano, Cortical gamma-oscillations modulated by listening and overt repetition of phonemes. Neuroimage **49**(3), 2735–2745 (2010)

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

[30] Pei X, Leuthardt E C, Gaona C M, Brunner P, Wolpaw J R and Schalk G 2010 Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition *Neuroimage* **54** 2960–72

[31] Kellis S, Miller K, Thomson K, Brown R, House P and Greger B 2010 Decoding spoken words using local field potentials recorded from the cortical surface *J. Neural Eng.* **7** 056007

[32] Bin G, Gao X, Wang Y, Li Y, Hong B and Gao S 2011 A high-speed BCI based on code modulation VEP *J. Neural Eng.* **8** 025015

[33] Reed C M and Durlach N I 1998 Note on information transfer rates in human communication *Presence: Teleoperators Virtual Environ.* **7** 509–18

[34] Blakely T M, Miller K J, Rao R P N, Holmes M D and Ojemann J G 2008 Localization and classification of phonemes using high spatial resolution electrocorticography (EcoG) grids *EMBS'08: Proc. 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* Pp 4964–7

[35] Leuthardt E C, Gaona C, Sharma M, Szrama N, Roland J, Freudenberg Z, Solis J, Breshears J and Schalk G 2011 Using the electrocorticographic speech network to control a brain-computer interface in humans *J. Neural Eng.* **8** 036004

[36] Mugler, Emily M., James L. Patton, Robert D. Flint, Zachary A. Wright, Stephan U. Schuele, Joshua Rosenow, Jerry J. Shih, Dean J. Krusienski, and Marc W. Slutzky. "Direct classification of all American English phonemes using signals from functional speech motor cortex." *Journal of neural engineering* 11, no. 3 (2014): 035015.

[37] Ramsey, N. F., Salari, E., Aarnoutse, E. J., Vansteensel, M. J., Bleichner, M. G., & Freudenburg, Z. V. (2018). Decoding spoken phonemes from sensorimotor cortex with high-density ECoG grids. *Neuroimage*, *180*, 301-311.

[38] Lotte, F., Brumberg, J. S., Brunner, P., Gunduz, A., Ritaccio, A. L., Guan, C., & Schalk, G. (2015). Electrocorticographic representations of segmental features in continuous speech. *Frontiers in human neuroscience*, *9*, 97.

[39] Mugler, E. M., Tate, M. C., Livescu, K., Templer, J. W., Goldrick, M. A., & Slutzky, M. W. (2018). Differential representation of articulatory gestures and phonemes in precentral and inferior frontal gyri. *Journal of Neuroscience*, *38*(46), 9803-9813.

[40] Herff, C., Heger, D., De Pesters, A., Telaar, D., Brunner, P., Schalk, G., & Schultz, T. (2015). Brain-to-text: decoding spoken phrases from phone representations in the brain. *Frontiers in neuroscience*, *8*, 141498.

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

[42] Moses, D. A., Leonard, M. K. & Chang, E. F. Real-time classification of auditory sentences using evoked cortical activity in humans. *Journal of neural engineering* **15**, 036005 (2018).

[43] Makin, J. G., Moses, D. A. & Chang, E. F. Machine translation of cortical activity to text with an encoder–decoder framework. *Tech. Rep*., Nature Publishing Group (2020).

[44] P.B. Denes and E.N. Pinson, The Speech Chain: The Physics and Biology of Spoken Language, Anchor Press-Doubleday, Garden City, NY, 1973

[45] Indefrey, Peter. "The spatial and temporal signatures of word production components: a critical update." *Frontiers in psychology* 2 (2011): 255.

[46] L. Rabiner and B. Juang, Fundamentals of speech recognition. Prentice Hall, 1993

[47] Benade, Arthur H., Fundamentals of Musical Acoustics, Oxford University Press, 1976

[48] Friederici, A. D., and Gierhan, S. M. (2013). The language network. *Curr. Opin. Neurobiol.* 23, 250–254. Doi: 10.1016/j.conb.2012.10.002

[49] Saur, D., Kreher, B.W., Schnell, S., Kümmerer, D., Kellmeyer, P., Vry, M.S., Umarova, R., Musso, M., Glauche, V., Abel, S. and Huber, W., 2008. Ventral and dorsal pathways for language. *Proceedings of the national academy of Sciences*, *105*(46), pp.18035-18040.

[50] López-Barroso, Diana, and Ruth de Diego-Balaguer. "Language learning variability within the dorsal and ventral streams as a cue for compensatory mechanisms in aphasia recovery." *Frontiers in Human Neuroscience* 11 (2017): 476.

[51] Huang, Vincent S., and Reza Shadmehr. “Evolution of motor memory during the seconds after observation of motor error.” *Journal of neurophysiology* 97.6 (2007): 3976-3985.

[52] Lombard, Etienne. “Le signe de Televation de la voix.” *Annu. Maladies oreille larynx nez pharynx* 27 (1911): 101-119.

[53] Xu, Y., Larson, C. R., Bauer, J. J., & Hain, T. C. (2004). Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences. *The Journal of the Acoustical Society of America*, *116*(2), 1168-1178.

[54] Purcell, David W., and Kevin G. Munhall. “Compensation following real-time manipulation of formants in isolated vowels.” *The Journal of the Acoustical Society of America* 119.4 (2006): 2288-2297

Chapter 2 BEHAVIORAL ANALYSIS OF SPEECH PRODUCTION

2.1 Introduction

Speech recognition involves capturing speech patterns from a person's voice, processing them through a computer, and identifying their content. This interdisciplinary field draws upon various disciplines, including artificial intelligence, phonetics, linguistics, signal processing, and information theory.

Speech recognition technology is widely integrated in our daily lives through applications like iOS Siri, Google, Amazon Alexa, etc. However, challenges persist in developing speech prosthetics due to diverse language dialogues and human speech accents. Many academics have worked on voice recognition technologies over the last few decades. Through years of research and trials, vowel recognition has emerged as one of the most efficient speech recognition method, given that vowel's formant frequencies are the most identifiable in spectrogram observations [1].

Speech production is a complex motor function that involves the coordination of various anatomical structures and physiological processes. In human speech, formant frequencies are key acoustic features that characterize the resonant properties of the vocal tract during the production of vowels. When we articulate vowels, the vocal tract shapes and filters the airflow from our lungs, resulting in specific resonances that correspond to formant frequencies. The first two formants, denoted as F1, F2, are particularly important in vowel perception and classification. Each vowel in speech is characterized by a unique combination of formant frequencies, which allow us to distinguish and identify different vowels. Formant frequencies are crucial in speech recognition and synthesis systems because they provide essential acoustic cues for decoding spoken language.

A vowel extraction system comprises two important components : formant extraction and classification. Formant extraction is a key step in visualizing vowel behaviors. Current research

predominantly utilizes two approaches for formant extraction: Linear Prediction Coefficient (LPC)-based formant extraction and Mel-frequency cepstral coefficients (MFCC)-based formant extraction [2].

Speech signals are non-stationary, meaning their characteristics vary over time based on the speaker's vocal tract. Short-Time Fourier Transform (STFT) has been widely used in literature to study speech in the frequency domain, since it has been shown that a frame of 20 - 30 ms could be considered as time invariant for analysis [3]. Linear prediction coding (LPC) is a time-domain based signal-source modeling, which makes it a powerful tool to characterize vowels behavior during the speech signal [4].

2.2 Dataset Description

This work is conducted in collaboration with Prof. Christian Herff at Maastricht University, whose team conducts the experiments and provides us with the recordings. In this study, the dataset includes recordings from three subjects with pharmaco-resistant epilepsy with hospital ID codes corresponding to KH28, KH30, and KH31. During the experiment, participants are presented with a randomly selected vowel displayed on a screen from a set of six predetermined Dutch vowels, and instructed to produce it after a cue is given to them. Each session consists of a total of around 70 trials, where one trial is one vowel production and two consecutive trials are separated by 2 seconds.

These subjects have been implanted with sEEG electrodes in different brain regions and audio recordings are captured synchronously with the sEEG data, sampled at 1024 Hz. The number of trials for each vowel ranges from 5 to 8 trials per subject, varying between subjects, as shown in table 2.1. Although the duration of vocalization differs across trials, the subject and the vowel

produced remain consistent across all trials for a specific vowel. Individual formant values and pitch are specific to each participant. Given the consistency of the subject across trials, we observe that the pitch remains constant across trials, and we expect minimal variation the extracted formant values for a specific vowel across trials.

Table 2-1 : Dataset Description

Subject Identification	Total number of trials in audio recording	Total number of electrodes in sEEG recordings
KH-28	74	97
KH-30	71	127
KH-31	70	130

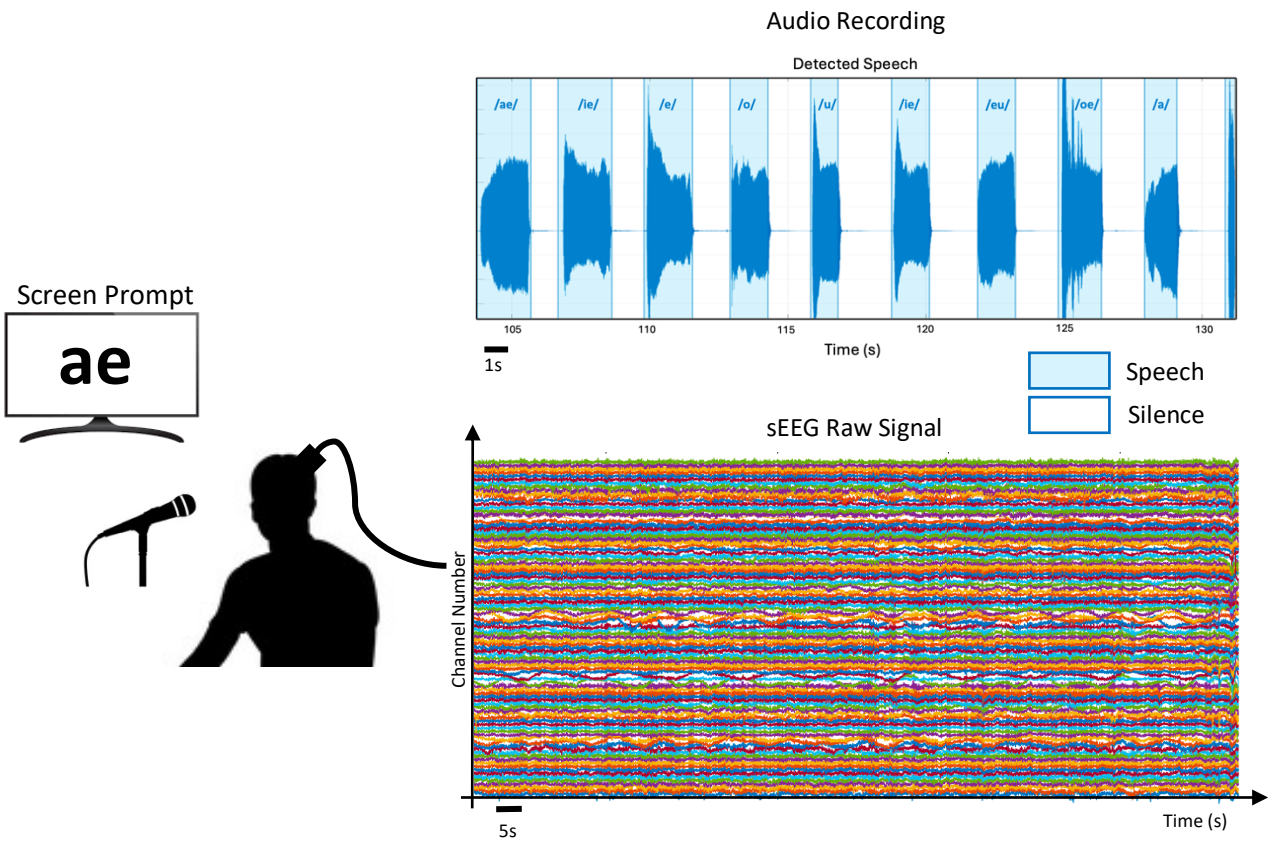

Figure 2-1 : Experimental Setup

2.3 Experiment Design

The goal of the project is to design a perturbation task to gain insight about neural mechanisms underlying auditory feedback and error propagation in vowel production. This includes three main tasks summarized in figure 2.2.

In the first task, baseline task, we want to study vowel production. In particular, we want to first extract the vowel parameters, i.e. the formants and estimate the production variability in each subject. The second task is focused on perception, where the goal is to understand how sensitive individuals are to errors introduced in vowel production space. Specifically, we're

interested in errors that lead to confusion between two vowels. To achieve this, we will take pairs of vowels that subjects produced in the first task and morph them at different percentages.

Now human audio perception of vowels doesn't follow a linear pattern but rather a sigmoid function. So, our objective is to estimate subject-specific psychometric curves for recognizing these perceptually relevant errors. This helps us gain insights into the perturbation boundary for the third task, the perturbation task, that is still being prepared in the EMU at UCSD. Here, we want to introduce errors into the audio recording extracted in tasks 1 and 2 and play it back to the subject. The goal here is to gain insight into the neural mechanism underlying auditory feedback and error propagation to construct more informative loss functions that enable richer neural decoder representations that more closely correlate to naturally perceived error.

The work in this thesis relates to task 1 and start of task 2, where the goal is to investigate how neural signals control speech production and try to identify specific neural markers for speech production.

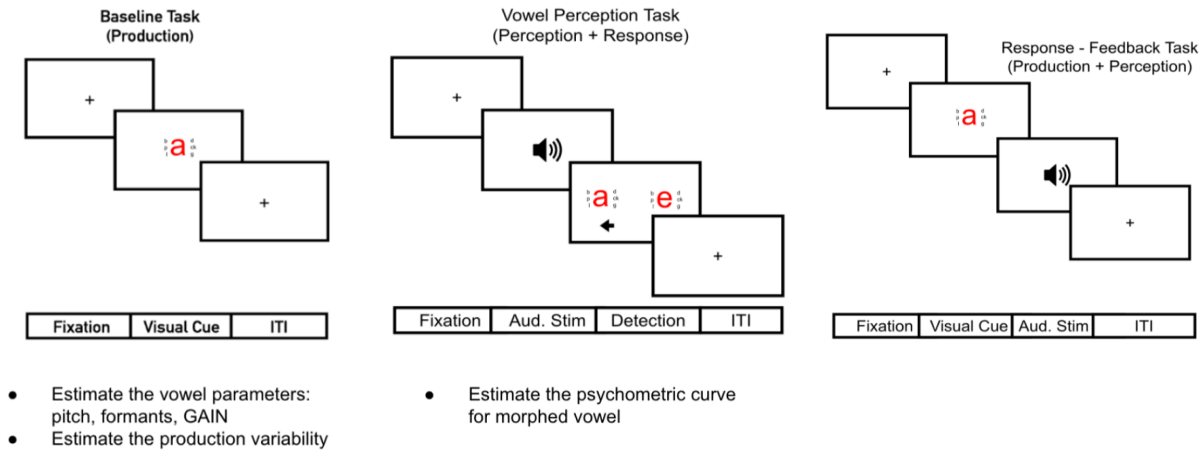

Figure 2-2 : Experiment Design Blocks

2.4 Audio Feature Extraction

2.4.1 Speech Onset Detection

To reduce the processing time and errors in formant extraction, I developed a way to detect the onset of speech production. By doing so, the extraction algorithm can be run only on the voiced part of the input signal (speech part), which will resolve the issue of detecting random formant values in the silence part.

The detection algorithm is based on the energy and zero-crossing rate of the input signal. First, during speech, the energy of the signal is high compared to periods of silence and by setting a threshold, the onset of speech production can be detected.

Average energy can be defined as:

$$E = \sum_{m=0}^{N-1}[w(m)x(n-m)]^2\,,\ \ 0 \le m\ \le N-1 \quad (1)\ [5]$$

where x(n) is the speech signal, N the length of frame, m is the frame shift and w(m) is the window function.

For the purpose of this work, we compared two windowing methods : Hamming window and Hanning window. The Hamming window can be described as follows :

$$w(n) = \begin{cases} 0.54 - 0.46 * \cos\frac{2\pi n}{N-1}\,,\ 0 \le n \le N-1 \\ 0,\ \ otherwise \end{cases} \quad (2)\ [6]$$

The Hanning window can be described as :

$$w(n) = \begin{cases} 0.5 - 0.5 * \cos\frac{2\pi n}{N-1}, & 0 \le n \le N-1 \\ 0, & otherwise \end{cases} \quad (3)[6]$$

The length of the window is crucial for frequency analysis since if the window is large in the time domain, it will be narrow in the frequency domain and vice versa. After comparison, it was concluded that both windows have the characteristics of low pass and symmetry. The main lobe of Hamming window is the widest, has the lowest side lobe level, has relatively stable spectrum for speech signal, and helps to enhance the characteristics of the central section of signal. So the Hamming window was used for speech onset detection.

Another metric that has been used to identify the onset of speech is zero-crossing rate (ZCR). It shows how many times the x-axis is crossed by a frame of voiced signal. One of the most straightforward techniques for time domain speech analysis is zero crossing analysis.

ZCR can be defined as [7] :

$$ZCR = \frac{1}{2}\sum_{m=0}^{N-1}|sgn[x(m)] - sgn[x(m-1)]| * w(n-m) \quad (4)$$

During speech, the ZCR of the speech signal is low compared to periods of silence as shown in figure 2.3. So, by combining these two metrics, I was able to extract the speech frames from the input audio signal and classify each frame as voiced or silence and extract the time indices for speech onset and offset that will be used later in the neural signal pipeline. Once a frame is detected to be voiced, i.e. speech, the formant extraction algorithm will then run on this specific frame.

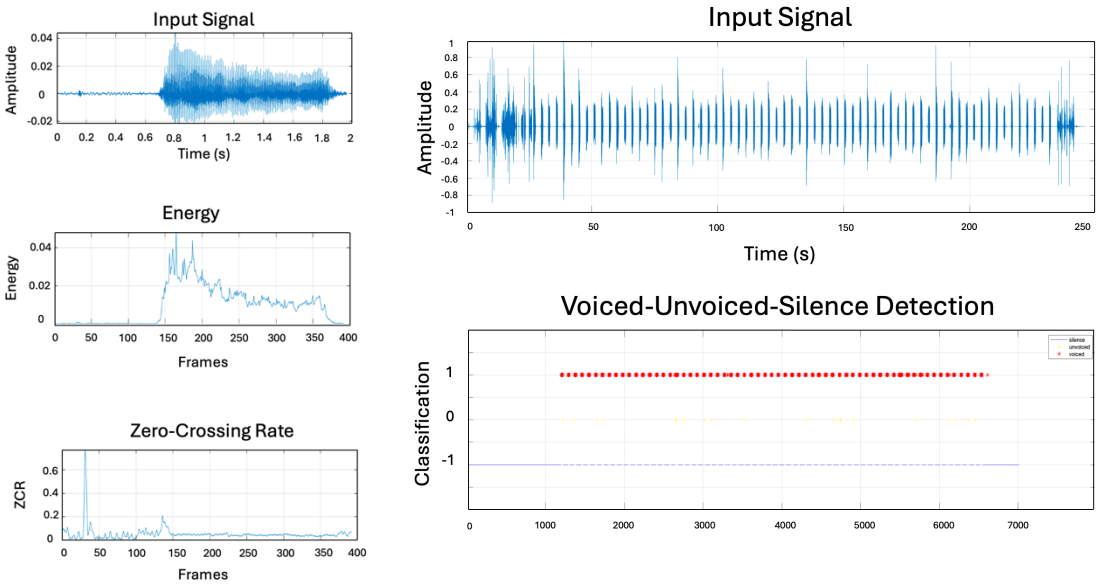

Figure 2-3 : Energy, ZCR and Classification results.

2.4.2 Formant Extraction

One of the most significant and popular speech analysis approaches is linear predictive analysis. This method's significance is rooted in both its relative speed of computation and its capacity to deliver precise estimates of the speech parameters. It is an all-pole model in the Z transform domain since the linear predictive analysis method assumes that the speech can be described by a predictor model that only considers previous values of the output. It is a time-domain based signal-source modeling where the source, $e(n)$, models the vocal cords, while the

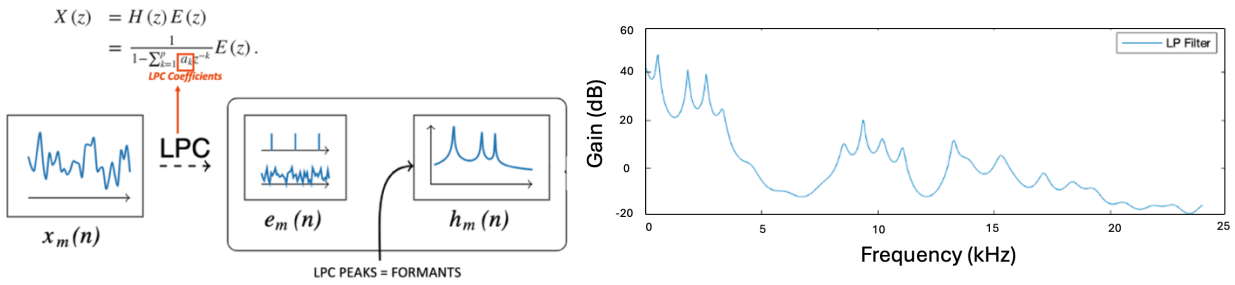

Figure 2-4 : LPC method diagram and resulting spectrum for one voiced frame

resonant filter, $h(n)$, models the vocal tract. This makes it a powerful tool to characterize vowels behavior during the speech signal. Due to the error minimized by LP, spectral peaks are emphasized in the envelope, as they are in the auditory system and correspond to the formant frequencies. The order of the LPC model is crucial to get detailed coefficients of the input signal and be able to accurately extract the frequency components. Multiple orders of LPC were compared, ranging from order 3 to 50, and it was concluded that an LPC model of order 50 approximated the speech signal best and was able to extract formant values fast and accurately. After running the algorithm, we get an LPC spectrum for each voiced frame as shown in figure 2.4, where each peak represents a formant frequency.

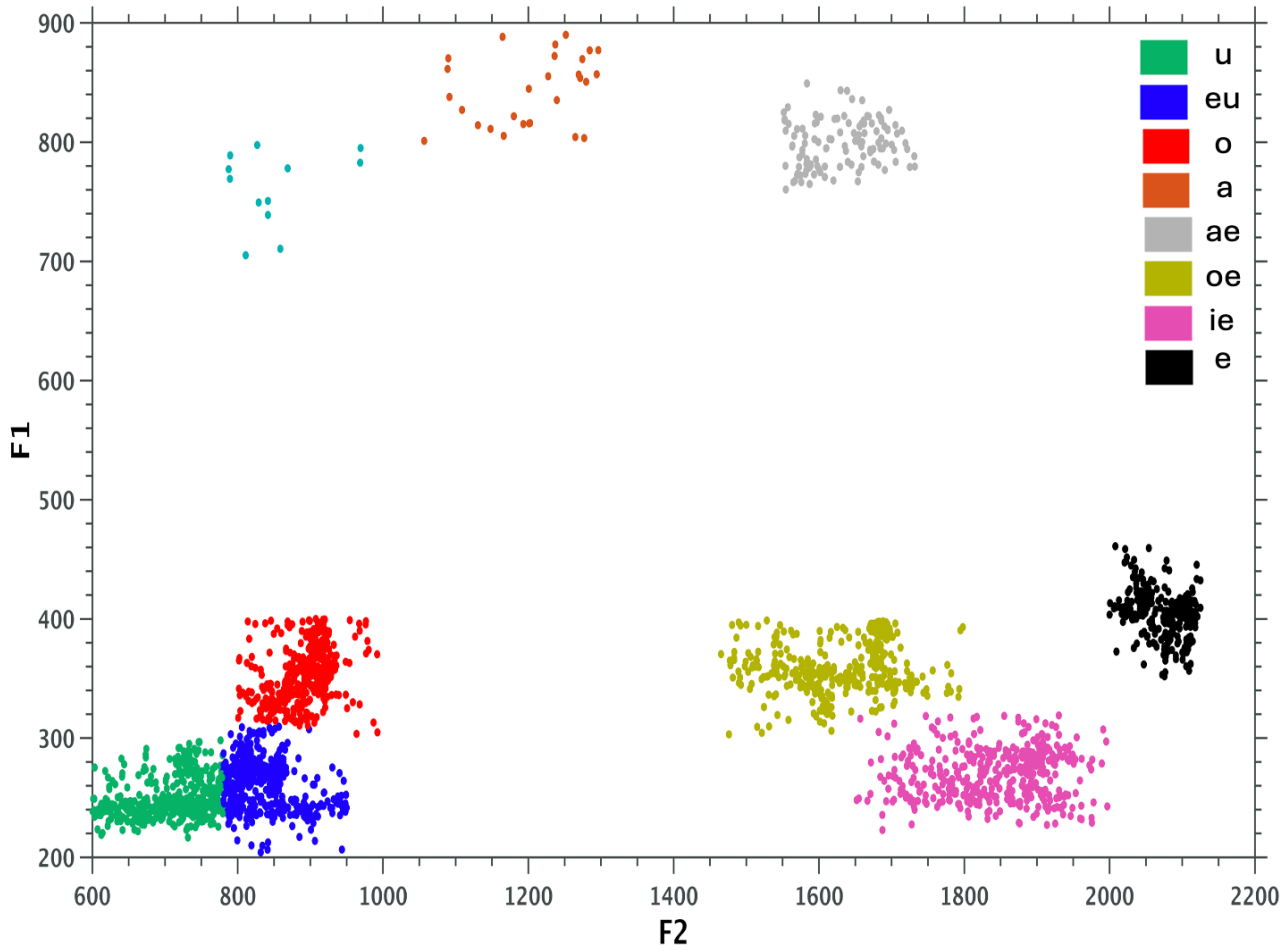

Figure 2-5 : Extracted vowel identity results in the formant space for subject kh-28

Using this algorithm, the first two formants, F1 and F2 were extracted for each trial and each subject. Figure 2.5 represents the resulting values in the formant space for one subject and shows that the vowels can be uniquely identified. In the remaining parts of the project, I only focused on the following six vowels : /o/, /u/, /eu/, /ie/, /e/, /oe/. This is because, for the other subjects, vowels /a/ and /ae/ were not present.

2.5 Structured Perturbation for Speech Production

Miller et. al. [8] explain the notion of perceptually relevant error, divinding it into two approaches. The first one is quantifying the quality of synthesized speech, by quantifying distance between true and synthesized audio with distance metric on spectrogram or spectral feature including but not limited to L2, L1, cc. However, one problem with this approach is that there are no established metrics for quantifying decoded speech when the ground truth is not known – that is, when the BCI user can't speak as is the case with anarthria. The second approach, which is applicable to a synthesized-only situation, is to have human listeners report if they understood the speech. However, this is imprecise, labor-intensive, and cannot be used as a cost function for the BCI. The ASR assessment is also problematic since the synthesized speech is often unintelligible.

For the third task, our approach is to raise and decrease formant values towards another vowel, method called vowel morphing, and add random noise, as control condition. We are interest in how behaviorally human adapt to those errors and how error is perceived and integrated into the speech production neural pathway. In this work, vowel morphing strategies have been implemented as a first step towards completing these tasks.

Current theoretical models of speech production suggest that vowel production is strongly reliant on internally represented speech goals [9]. Although the exact nature of the speech goals is unclear, it has been suggested that the speech motor system (SMS) may use perceptual goals (e.g.,

auditory goals) to determine errors in its motor output. These models posit that during production, the SMS compares auditory feedback of the produced speech with its auditory goals; when the auditory feedback resides outside the auditory goals (i.e., auditory error), the SMS generates corrective motor responses to reduce the perceived error. It has been showed that real-time auditory feedback perturbations (shifts in formant frequencies) of productions that were closer to the edge of the vowel category elicited larger compensatory responses relative to identical perturbations of productions closer to the center of the vowel. These results suggested that the SMS may use the perceptual boundary between two adjacent vowels to determine errors in its output.

As a first step to implement this method, I used linear interpolation in the formant space between two vowels, as shown in figure 2.6, to get new F1 and F2 values for different degrees of morphing. In order to synthesize a morphed vowel with the given morphed formant values, I used the Klatt formant synthesizer. It is an online tool that, given the values of the first two formants, can synthesize and audio signal of the desired vowel. The next steps are, first to investigate other techniques for vowel morphing, and second to use these audio recordings and play the morphed vowels to the subjects in the EMU and see how sensitive individuals are to errors introduced in vowel production space and get the CPB for the third task.

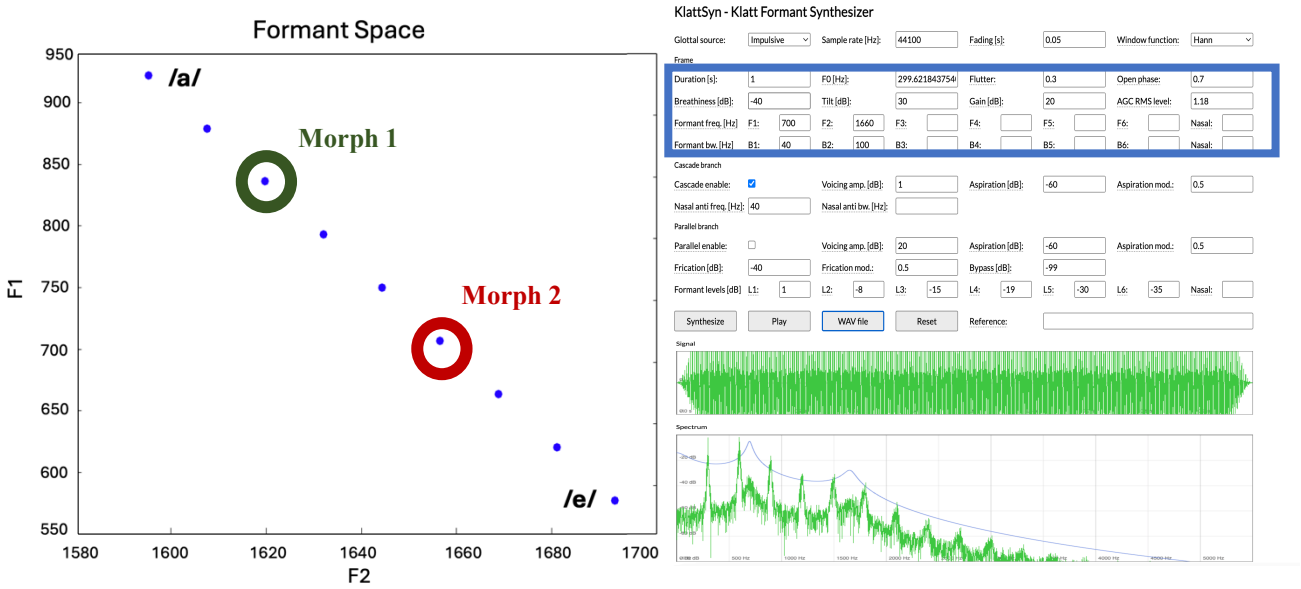

Figure 2-6 : Vowel morphing results and Klatt Synthesizer interface.

2.6 Discussion and Conclusion

Human listeners deal easily with variability in speech and can classify almost all vowels from any speaker of their native language correctly by perceptually “evening-out” differences between speakers. However, formant measurements show considerable variation related to anatomical/physiological differences between speakers, which becomes apparent when the first two formants, F1 and F2, corresponding to vowels produced by different speakers are plotted on the same formant-space plot.

It has been shown in literature that different vowels show different spectral characteristics [10]. A lot of phonetic studies thus model vowels as points sampled at steady state [11]. By contrast, vowel inherent spectral changes (VISC) refer to the changes in spectral properties over time of a specific vowel and are a characteristic of vowel identity [12]. However, VSIC focuses on the shape of formant contours throughout production rather than on static formant values. Thus, some changes in the vowel spectrum are not inherent but unplanned.

Rose (2002) states that speakers differ from each other in their voices (inter-speaker variation), but their voice also overlap considerably [13]. Also, speakers show variations within their own voice (intra-speaker variation). Speakers never produce the same vowel or utterance in the exact same way twice, thus acoustic differences between trials of specific vowel productions are always present [14]. Niziolek et. Al. proved that individuals correct their speech when a vowel perception is deviant from what is expected, based on the extent and direction of formant movement [15]. The findings suggest that vowels have static auditory targets in steady state.

Dutch vowels are generally voiced [16]. Formants have been shown to provide relatively high speaker-dependent information. McDougall et. al. showed that F2 provided more speaker-specificity than F1 [17]. They showed that F1 was not affected much by the intra-speaker variability whereas F2 was.

This is in accordance with our results where the range of frequencies for F1 around the mean value is smaller than that of F2 for a given vowel per subject, as shown in figure 2.5. Also, as shown in figure 2.7, F1 values remain consistent across time for a single trial whereas F2 values show more variations. We observe that it is more difficult for the subject to sustain constant F2 values for complex vowels like /oe/, /eu/ and /ie/, which emphasizes the fact that F2 is more susceptible to intra-speaker variabilities than F1. However, even with these intra-speaker variabilities, we are still able to uniquely identify vowels in the formant space, as shown in figure 2.5.

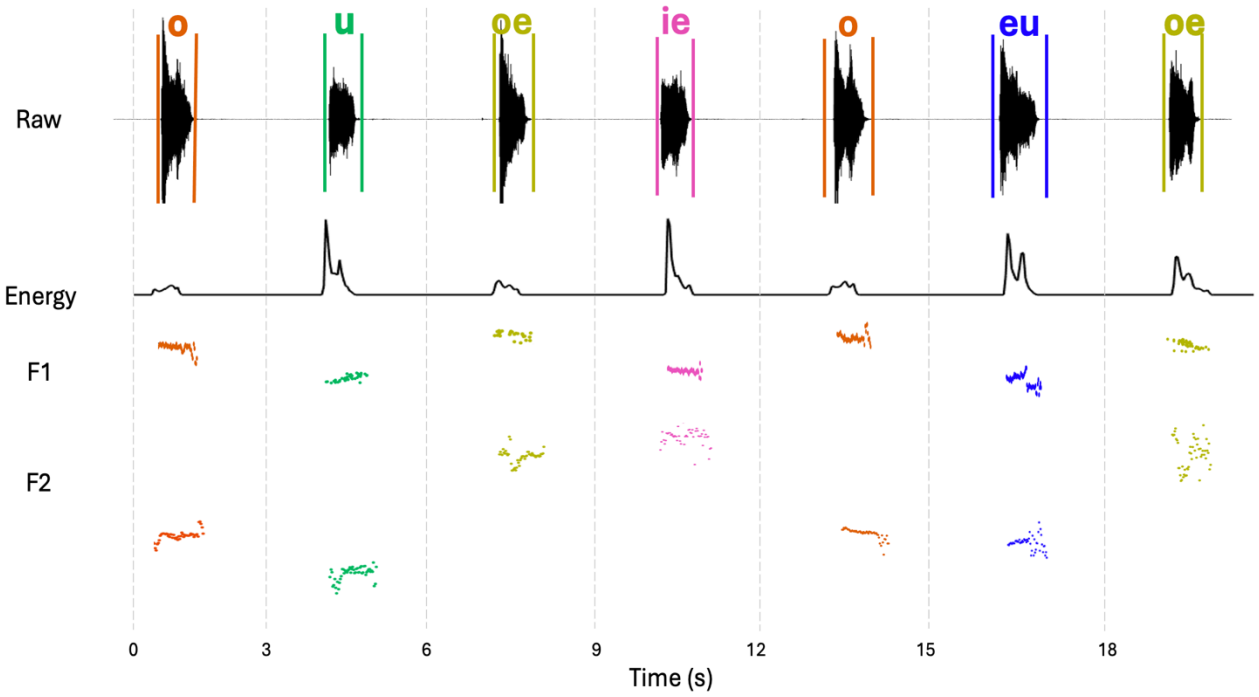

Figure 2-7 : Raw audio recording with corresponding energy plot and formant values across time

In conclusion, our pipeline was able to extract formant values from the input audio signal. These values show that we can uniquely identify vowels in the two-dimensional formant space, where each vowel is characterized by unique range of pair of F1-F2 values, even if we observe intra-speaker variabilities.

REFERENCES

[1] Stam, D. C. (n.d.). “Vowel recognition in continuous speech (thesis)”.

[2] Nehe, N. S., & Holambe, R. S. (2012). DWT and LPC based feature extraction methods for isolated word recognition. EURASIP Journal on Audio, Speech, and Music Processing, 2012(1).

[3] D. MANDALIA, P. GARETA, ”Speaker Recognition Using MFCC and Vector Quantization Model,” Institute of Technology, Nirma University, 2011.

[4] Shahriar, S., & Hoq, M. N. (2016). Evaluation of LPC trajectory for Vowel-Consonant-Vowel sequence. 2016 19th International Conference on Computer and Information Technology.

[5] Y. Yang, ”Research on Endpoint Detection of Speech,” Computer Systems & Applications , p. vol 21(6), 2012.

[6] Q. Xin och P. Wu, ”Research and Practice on Speaker Recognition Based on GMM,” Computer and Digital Engineering, p. vol 37(6), 2009. Y. Deng, X. Jing, H. Yang,

[7] Y. Cai, ”Research on end point detection of speech signal,” University of Jiangnan, nr TN912.3 Master thesis, 2008.

[8] S. Varshney, D. Farias, D. M. Brandman, S. D. Stavisky and L. M. Miller, "Using Automatic Speech Recognition to Measure the Intelligibility of Speech Synthesized From Brain Signals," *2023 11th International IEEE/EMBS Conference on Neural Engineering (NER)*, Baltimore, MD, USA, 2023, pp. 1-6, doi: 10.1109/NER52421.2023.10123751.

[9] Chao, Sara-Ching, Damaris Ochoa, and Ayoub Daliri. "Production variability and categorical perception of vowels are strongly linked." *Frontiers in Human Neuroscience* 13 (2019): 96.

[10] Lindblom, B. E., & Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. The Journal of the Acoustical society of America, 42(4), 830-843.

[11] Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. The Journal of the Acoustical society of America, 24(2), 175-184.

[12] Morrison, G. S. (2012). Theories of vowel inherent spectral change. Vowel inherent spectral change, 31-47.

[13] Rose, P. (2002). Forensic speaker identification. London : Taylor & Francis.

[14] Latinus, M., & Belin, P. (2011). Human voice perception. Current Biology, 21(4), 143-145.

[15] Niziolek, C. A., & Guenther, F. H. (2013). Vowel category boundaries enhance cortical and behavioral responses to speech feedback alterations. Journal of Neuroscience, 33(29), 12090-12098

[16] Rietveld, A. C. M. & Van Heuven, V. J. (2013). Algemene fonetiek. Bussum : Coutinho.

[17] McDougall, K. And Nolan, F. (2007). Discrimination of speakers using the formant dynamics of /u:/ in British English. In Proceedings of the 16th International Congress of Phonetic Sciences (pp. 1825-1828)

Chapter 3 NEURAL ANALYSIS OF SPEECH PRODUCTION

3.1 Introduction

Perception and production of speech in the human brain requires the tightly coordinated activity of neurons across large portions of frontal and temporal cortices. This allows humans to produce a remarkably wide array of sounds that, when arranged together, make up words. These processes occur rapidly during normal speech and have been shown to recruit prefrontal regions known to be involved in word planning [1-9] and sentence construction [10-13]. Damage to one or more areas within this network can result in aphasia, dysarthria, or apraxia of speech. Bouchard et. al. demonstrated that phonetic features may be regionally organized and decoded from neural activity [14], suggesting the existence of an underlying cortical structure for phoneme production.

Previous studies on animals [15-17] and humans [18,19] showed how primary motor areas relate to vocalization movements and production of sound sequences such as songs. However, they do not reveal the neural processes involved in individual word production during natural speech. Other studies found a large regional overlap in areas involved in speech articulation and production [20-23]. In their study, Mugler et. al. were able to decode the entire set of phonemes from American English using ECoG [24] and successfully analyze and classify individual phonemes within word production, with 20.4% of all phonemes classified correctly. However, this study was only conducted over the primary motor cortex. Addition of more electrodes over other brain areas could improve the classifier performance.

All these studies show that there are strong neural correlates between phoneme production and neural activity. In this work, we investigate these correlations on a set of 6 Dutch vowels to see if we can find phoneme specific neural markers during speech production.

3.2 Dataset Description

3.2.1 Brain Areas Coverage

A thorough knowledge of the brain's speech production process, the specific timing of the brain areas involved, and the optimal locations for decoding them is necessary for the development of a speech neuroprostheses. A speech-BCI that relies on numerous areas within the speech production network, spanning both frontal and temporal cortices, may provide a robust basis for neuroprostheses applicable with broader indications. A recent proof of concept study illustrated that sEEG recordings are capable of decoding acoustic aspects of both overt and imagined speech [25]. Even when speech is silently 'imagined', it triggers activity across a broad network of brain areas despite the absence of audible sound. The decoding models identified activations from various regions within the frontal and temporal lobes and showed that the ventral portion of the sensory-motor cortex (vSMC) is a key brain area in the neural control of articulation [26]. Principal component analysis was utilized by Bouchard et al. [14] to convert the population brain activity into a "cortical state-space" offering a comprehensive representation of the cortical patterns linked with the produced syllables. Leveraging the superior temporal resolution of ECoG, they successfully distinguished the temporal aspects of cortical activity related to consonants and vowels. An analysis of the arrangement of both consonants and vowels within this cortical state-space unveiled that distinct phonemes were grouped based on the primary oral articulators involved in their production. Consequently, the spatial configurations of cortical activity across various speech articulators were utilized to elucidate the arrangement of phoneme representations throughout the vSMC network. Previous studies using ECoG and MEA showed that certain cortical regions contain representations for specific speech components. A somatotopic organization of place of articulation (POA), was shown in ventral sensorimotor cortex [14, 28-29],

while representations of manner of articulation (MOA) and phonemic components were localized more in lateral temporal cortex [29-32].

However, in a recent study, Thomas et. al. [33] showed that the decoding contributions of each speech component did not cluster within one specific cortical region. Instead, both articulatory and phonetic speech components were decoded using sEEG electrodes distributed across multiple areas known to play major roles in the speech production network. This may be due to the fact that this widespread coverage of sEEG also targeted locations involved in processes that precede speech onset, such as conceptualization, lexical access, and phonological formulation.

All these studies show that neural modulation of speech production is spread across a wide range of brain regions. sEEG is thus a good recording technology that can leverage this distribution and provide information about phonemic representation across neural populations.

In this work, three patients suffering from pharmaco-resistant epilepsy, native and fluent Dutch speakers, were implanted with sEEG electrodes in different brain regions as part of the clinical therapy for their epilepsy. Electrode locations were purely determined based on clinical necessity. The experiment and recording were done by our collaborator at Maastricht University where experiment design and data recording were approved by the Institutional Review Boards of both Maastricht University and Epilepsy Center Kempenhaeghe.

Participants were implanted with platinum-iridium sEEG electrode shafts (Microdeep intracerebral electrodes; Dixi Medical, Beçanson, France) with a diameter of 0.8 mm, a contact length of 2 mm and an inter-contact distance of 1.5 mm. Each electrode shaft contained between 5 and 18 electrode contacts. Two patients had electrode coverage in both hemispheres and one only in the left hemisphere. The number of shafts range from 9 to 12, and total number of electrodes

from 93 to 113 channels and anatomical labeling was done as described in [34]. An example of the brain areas covered for subject kh-28 is shown in figure 3.1.

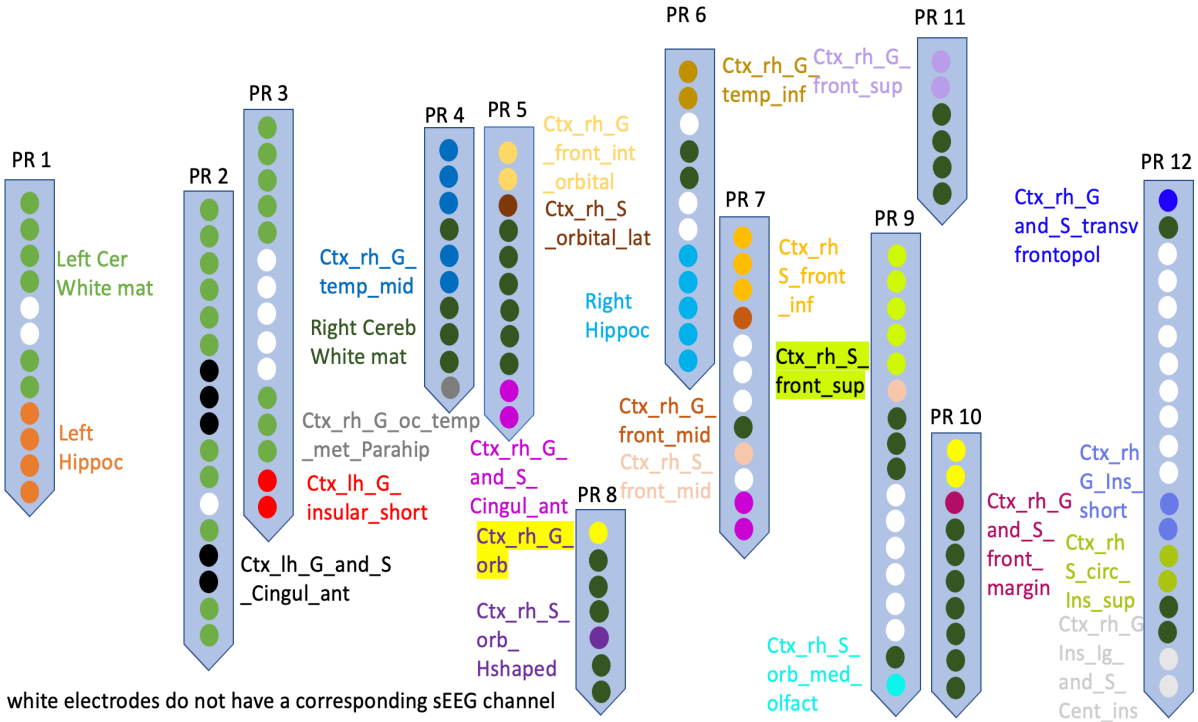

Figure 3-1 : sEEG brain regions coverage for subject Kh-28

3.3 Neural Signal Conditioning

As stated in [34], neural data was recorded in Maastricht by Prof. Herff using two or more Micromed SD LTM amplifier(s) (Micromed S.p.A., Treviso, Italy) with 64 channels each. Electrode contacts were referenced to a common white matter contact. Data were recorded at either 1024 Hz or 2048 Hz and subsequently downsampled to 1024 Hz. They used LabStreamingLayer [35] to synchronize the neural, audio and stimulus data.

3.3.1 Referencing Strategies

First, we used signal-to-noise ratio (SNR) and detected movement artifact as inclusion and exclusion criteria for channel used towards analysis. The remaining channels were high pass filtered with a cutoff frequency of 1 Hz to remove any underlying DC signal and then notch filtered with center frequencies at 50, 100 and 150 Hz to remove line noise and its harmonics.

It has been shown that speech activity is strongly represented in gamma and high-gamma frequency bands. The process of identifying broadband gamma and oscillatory activity starts with referencing a signal at a specific site against the signal at one or two reference locations during recording. Subsequently, a particular re-referencing method is typically applied, often in post-hoc analyses. The choice for referencing locations usually follows specific guidelines [36]. The benefits and shortcomings of different re-referencing techniques have been determined for ECoG [37] but not yet for sEEG. The most effective re-referencing approach for sEEG may vary from that used for ECoG, primarily due to sEEG geometry of the sEEG electrodes. In comparison to the surface electrode, sEEG is a depth electrode that allow direct contact across various brain structures, such as the cortex and white matter. These structures may exhibit differences in amplitudes, impedance levels, or other characteristics. In this work, two methods of referencing were investigated : Common Average referencing (CAR) and Grey-vs-White Matter referencing (GWM).

For CAR, the average neural activity was subtracted from each channel and done on a probe-by-probe basis. For GWM referencing, the white matter channels and grey matter channels were first separated into right-hemisphere and left-hemisphere channels. The average white matter activity of the corresponding hemisphere was subtracted from each grey matter channel, i.e. the

average activity of white matter channels in the right hemisphere was subtracted from every grey matter channel in the right hemisphere and the same for the left.

3.3.2 Task Structure and Trialization

As stated in section 2.3, each speech production trial was labeled with the corresponding vowel and all sEEG channels were labeled according to their location in the brain. The audio and sEEG recordings were conducted synchronously. This allows us to translate the time indices for onset and offset of speech corresponding to individual phonemes for each trial that were extracted from the audio signal during the formant extraction phase onto the sEEG data. Neural data was then grouped based on speech vs silence periods for the first analysis in section 3.4 independent of vowel identity and then grouped per vowel label for the vowel decoding in section 3.5.

3.4 Speech and Silence Decoding

3.4.1 Frequency Domain Analysis (Power Spectral Density)

The first step in the analysis of the neural data was to see if there was a difference between baseline, periods of silence, and periods of speech production independently of the vowel identity. For that, the pipeline shown in figure 3.2 was implemented.

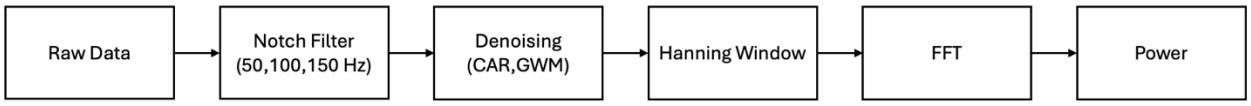

Figure 3-2 : Pipeline for Power analysis

The raw neural data was first passed through a notch filter at 50 Hz and its harmonics (100 Hz, 150 Hz, 200 Hz) to remove line noise. Then the pipeline was ran using both referencing

strategies, common average and grey-vs-white matter referencing, that will later be compared using a statistical t-test.

The neural data was then divided into speech segments and silent segments as shown in figure 3.3. One silent segment was taken 250 ms after the end of the previous vocalization with a length of 500 ms whereas one speech segment was taken 50 ms before speech onset (to consider neural activity around onset of speech due to speech preparation) with a length of 500 ms. The average length of silent segments in one recording session was 2 seconds and that of a speech segment was 1.3 seconds, so none of our silence segments overlap with the next speech segments.

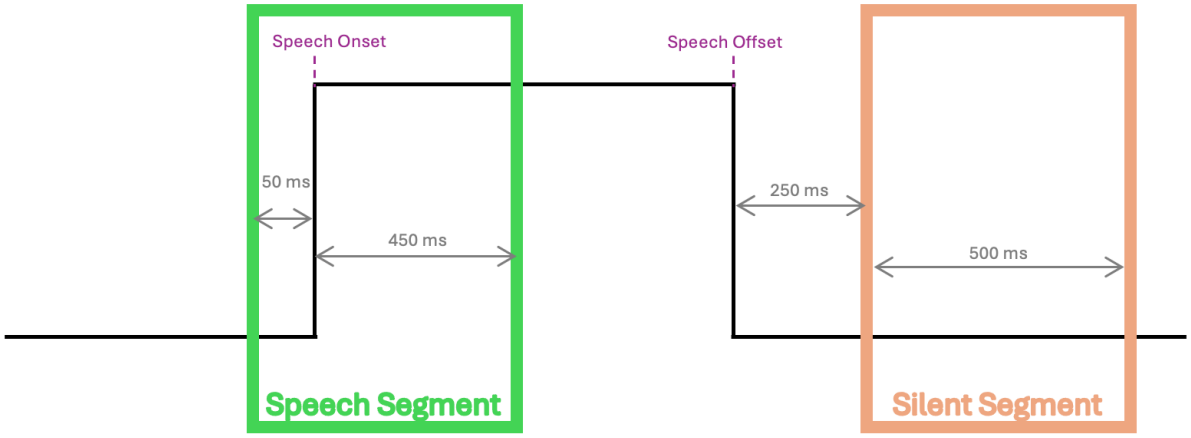

Figure 3-3 : Segmentation of speech and silence

After that, the Fast Fourier Transform (FFT) was computed using the Hanning window and power was calculated using (5) :

$$Power = \frac{1}{N*fs1} |X|^2 \text{ (dB)} \qquad (5)$$

where N is the length of the signal, fs1 is the sampling frequency (here 1024 Hz), and X is the output of the FFT. The resulting power was then plotted in the log domain for visualization as shown in figure 3.4. These results are consistent across channels and trials, where power of neural

data during speech is higher than that of silence for alpha (9-11 Hz), gamma (35-55 Hz), and high gamma (90-110 Hz) bands but the inverse applies in the beta band (15-30 Hz).

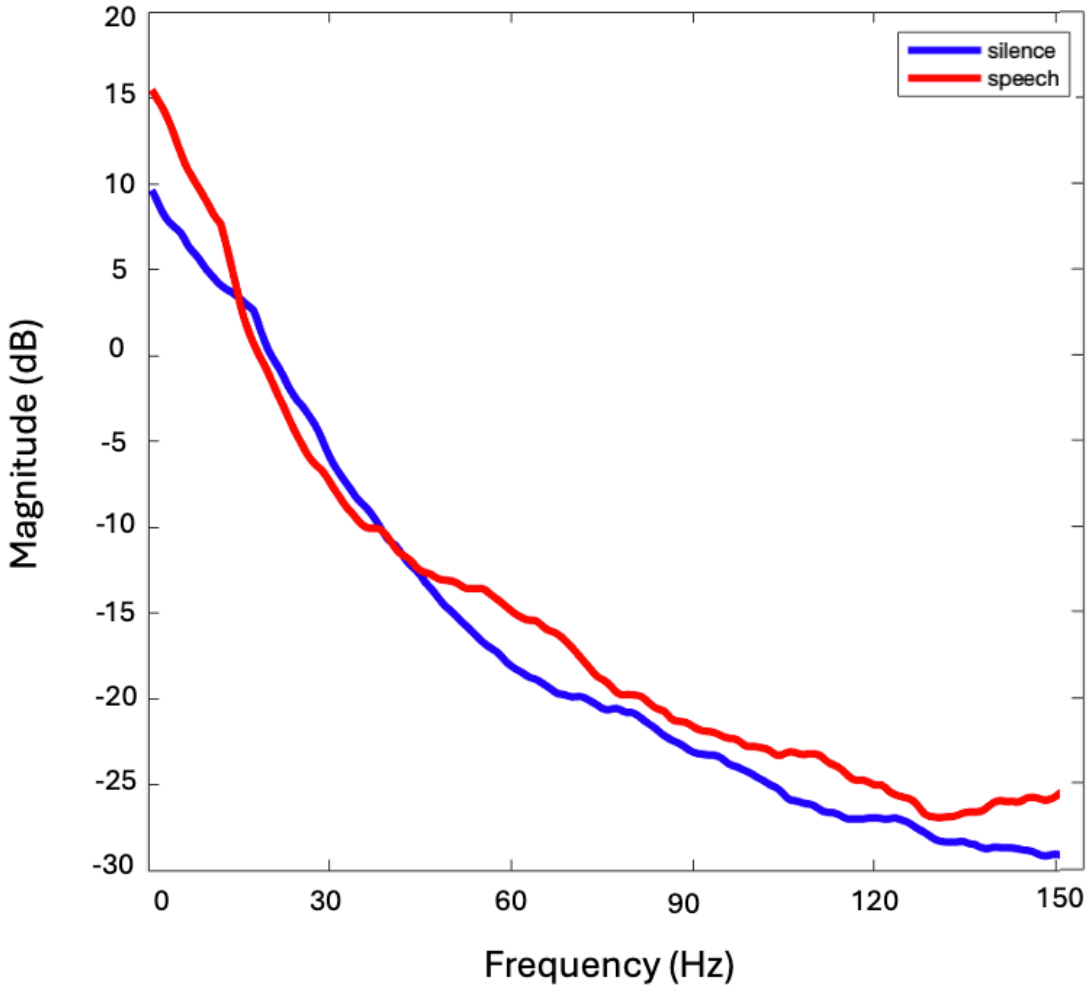

Figure 3-4 : Power Spectral Density of Speech and Silence in the log domain

3.4.2 Statistical Testing

In order to show that the results in section 3.4.1 are statistically significant and to compare the two referencing methods used, a statistical two-sample t-test with false discovery rate correction was performed on the resulting data for the different frequency bands (alpha, beta, gamma and high-gamma). This test is a method used to test whether the unknown population means of two groups are equal or not. The corresponding function used in MATLAB is [h,p] = ttest2(x,y). This function first returns a test decision for the null hypothesis that the data in vectors x and y comes from independent random samples from normal distributions with equal means and equal but unknown variances, where h=1 indicates the rejection of the null hypothesis at the Alpha significance level, given as 0.05. The second argument returned is *p*-value of the test, returned as a scalar value in the range [0,1]. p is the probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis. Small values of p cast doubt on the validity of the null hypothesis.

The hypothesis we want to test here is if the sEEG data from speech and silence come from different distributions for the different frequency bands. So, our null hypothesis is that the two data samples are from populations with equal mean. For our results to be significant we need to look at p-values < 0.05 when h=1.

The results show significance ($p < 0.05$) in the beta and especially high-gamma frequency bands, where more than 70% of the channels for all three subjects are significant, when using CAR. GWM referencing shows less channels with significance for all subjects, noting that we have less channels in total since the analysis is done on the grey-matter electrodes alone. We cannot use the white matter electrode since they are used as reference, one limit of this referencing

method. Figure 3.5 shows the resulting p-values for the different frequency bands for subject kh-28 for both referencing strategies.

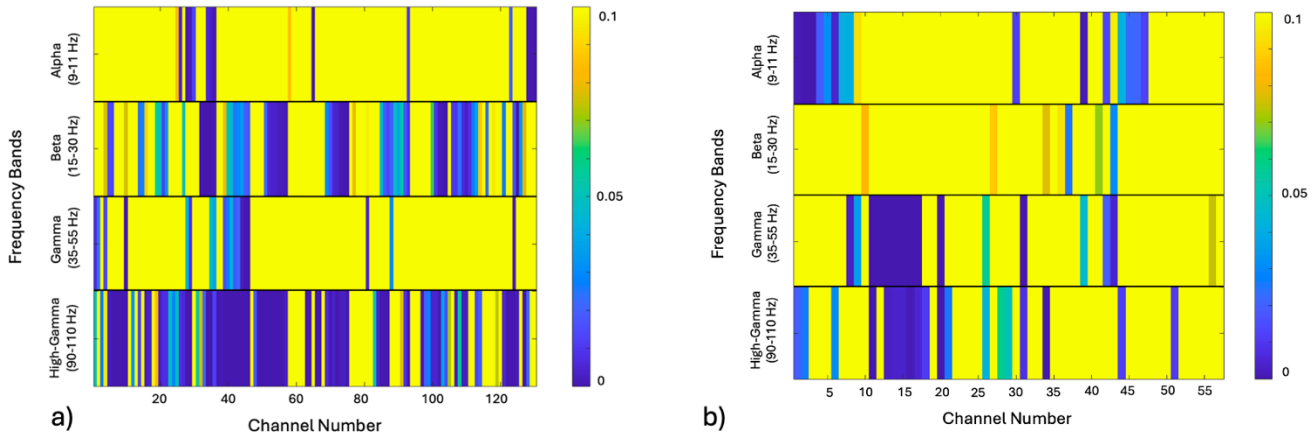

Figure 3-5 : T-Test p-value results after false discovery rate for subject kh-28 when using a) common average referencing, b) grey-vs-white matter referencing. The color bar indicates the p-values.

3.4.3 Neural Activity Around Speech Onset

In this part, we will focus on the time domain to see if different brain regions had specific and timed responses around speech onset for speech production, independently of vowel identity.

The raw neural data was first passed through a notch filter at 50 Hz and its harmonics (100 Hz, 150 Hz, 200 Hz) to remove line noise. Then CAR was used, and the data was bandpass filtered in the high gamma frequency range (90-110 Hz) using a 4^{th} order FIR Butterworth filter with zero phase. Using the time indices for onset and offset of speech, neural data corresponding to periods of speech were then passed through a Hilbert transform in the time domain and the power was calculated using :

$$Power = Hilb(x)^2 \qquad (6)$$

Where Hilb(.) is the output of the Hilbert Transform and x is the filtered neural data.

The last step was normalization using the z-score function (7) :

$$Z = (x - m)/s \quad (7)$$

Where x is the power calculated in (6), m and s are the mean and standard deviation across time per channel respectively.

The goal here is to study the neural behavior of speech production around onset. For that, we computed the z-scored power of the high-gamma activity for each vocalization trial. We then observed three distinct behaviors and grouped the channel in three categories based on the time of the peak power relative to speech onset, i.e. if peak power occurs before speech onset, around 100 ms after speech onset or late after speech onset as shown in figure 3.6a.

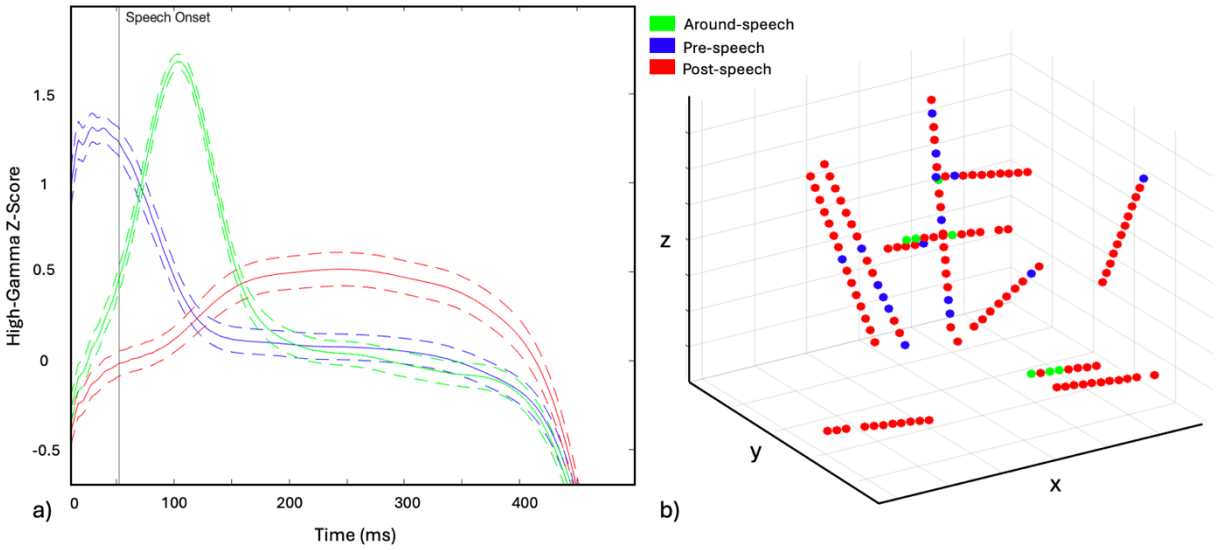

Figure 3-6 : a) Z-scored power of High-gamma activity across time during speech production, b) Spatial organization of channels in three-dimensions representing the grouping of behaviors

Another interesting visualization was to see the distribution of these responses across the brain. Figure 3.6b shows cortical activity at different electrode to visualize spatiotemporal patterns across the implanted sEEG electrodes in a three-dimensional plot, where the z=0 plane is the top of the brain.

3.4.3 Discussion

Our work demonstrated the feasibility and value of sEEG recordings in decoding speech production and enabled us to examine the spatial and temporal profiles of speech representations in the human brain. The frequency domain analysis first showed that the power of the neural signal in all brain regions covered in this work is different whether the subject is silent or producing speech. This shows that neural correlate for speech production exist across the brain, and we can decode speech production based on sEEG data.

Second, the statistical significance of channels between speech and non-speech was explored using a two-sampled t-test for both referencing strategies. For each participant, the significant channels (p-value ≤ 0.05) were similar across the frequency bands, and we can see that using CAR yields more channels that are statistically significant compared to GWM referencing. Thus, we will use CAR as our referencing strategy for the remaining of this work. It was also observed that significant channels were in both grey and white matter, which warrants further investigation as to the role of white matter in speech production.

Third, the t-test also showed that there are statistically significant differences in brain state associated with silence and speech production in both the beta band and high-gamma band. Since high-gamma band is more frequently employed as feature for speech BCIs in literature, this is the feature of our neural data that we will be using for vowel decoding in the next section.

Finally, theories of speech motor control and phonology have speculated that there is a hierarchical organization of phoneme representations, given the anatomical and functional dependencies of the vocal tract articulators during speech production [37-39].

Our results confirm that and seem to show that different brain regions respond not only differently but also at various stages of speech production, showing that cortical representations

are somatotopically organized, with individual sites tuned for a preferred phoneme and co-modulated by others. Most of the channels show peak activity late after speech onset for all the vocalization trials. However, some channels exhibit early responses to production that may be linked to speech preparation, or late responses that may be linked to either preparation for the next vocalization or other inherent processes. Also interesting to note that the distribution of these responses across the brain vary across trials. This shows that the activity not only depends on production but also on the nature of the speech being produced. We saw that for different vowels, most channels that were responding late after speech remained unchanged. However, the channels responding before and around speech onset were localized in different brain regions depending on the vowel and were consistent across trials. This shows that specific phonemes are encoded in precise brain regions, and important phonetic properties can be observed qualitatively from the spatiotemporal patterns across the brain. [14].

3.5 Vowel Identity Decoding

In this part of our work, we focus solely on the speech segments of input data. The electrodes used are the ones that showed significance between speech and silence in the t-test of section 3.4, and we focused on the z-scored high-gamma power for the remainder of this work. The same pipeline as section 3.4.3 was used here on each channel but now considering vowel

identity. We now want to push the analysis by adding vowel labels to vocalizations and see if we can decode individual vowels using neural high-gamma activity.

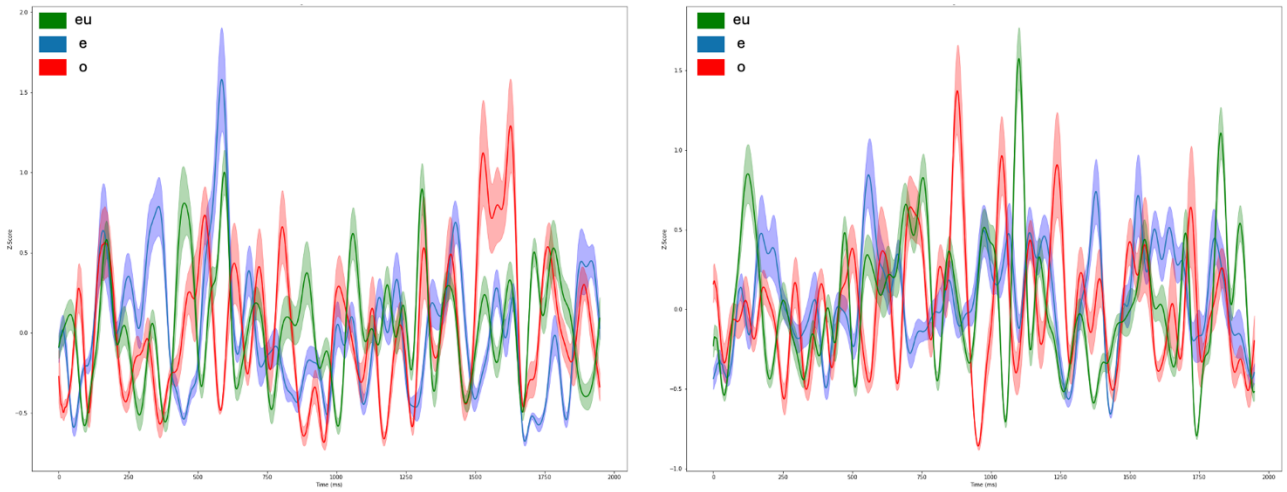

Figure 3-7 : Z-Scored power of high gamma activity and standard error for three vowels (/eu/, /e/, /o/). Channel locations are Left Hippocampus (left) and Right Cortex Temporal Middle (right)

As shown in figure 3.7, each vowel exhibits different neural responses not only across different brain regions but also when looking at a single channel. Just as focal stimulation is insufficient to evoke speech sounds, it is not any single region, but the coordination of multiple regions across the speech network that generates phonemes.

The distributed organization of phoneme representations led us to propose that coordination of the multiple brain regions required for speech production would be associated with spatial patterns of cortical activity.

3.5.1 Support Vector Machine Implementation

The implementation of a phoneme decoder was started in this work where the focus was to work on a support vector machine (SVM). A SVM is a machine learning algorithm that employs supervised learning models to tackle complex classification, regression, and outlier detection

problems. It achieves this by optimally transforming data to determine boundaries between data points based on predefined classes, labels, or outputs. SVMs are widely used in fields such as healthcare, natural language processing, signal processing, and speech and image recognition. The primary objective of the SVM algorithm is to identify a hyperplane that distinctly separates data points of different classes. This hyperplane is positioned to maximize the margin between the classes, ensuring a clear distinction. The SVM can be divided into two categories: the SVM for multiclass classification and the SVM for two-class classification (or binary classification). In general, the multiclass SVMs are usually created using combinations of several two-class SVMs.

Support vector machines can be categorized into linear and nonlinear models. A linear support vector machine is used when the data can be separated by a straight line or hyperplane in its original domain. Conversely, if the data cannot be linearly separated in the original domain but can be transformed into a feature space where linear separation is possible, it is termed a nonlinear support vector machine. In this work, we implemented a multiclass SVM with linear kernel.

In this work, we want to see if we can decode the vowel label/identity based on the neural high gamma power extracted. In particular, we want to see if vowels are encoded in different brain regions and for that we are going to decode labels based on single channel activity.

The first step was to divide the neural data into training and testing data. For each subject, we have a total of 5 to 7 trials per vowel, and 93 to 113 sEEG channels. For each channel, 75% of trials were used for training and 25% for testing. The second step was to implement the classifier. We used the SVM multi-class classifier in MATLAB where the inputs were one two-dimensional matrix containing neural activity, and one vector containing the true label of the vowels, represented using one encoding. The classifier was trained and then tested on the remaining trials for individual channels. We also included cross-validation, a robust technique used to assess the

performance and generalizability of SVM classifiers. It helps in evaluating how the results of a statistical analysis will generalize to an independent dataset and in detecting overfitting. Since we have a limited number of training and testing data, the cross-validation technique used was the leave-one-out (LOO) cross-validation.

3.5.2 Results and Discussion

Our classifier was able to decode vowel identity with accuracy above chance level. In some brain regions like temporal middle and temporal frontal, we got an accuracy as high as 55% for vowels /e/ and /o/. We observed different classification accuracies for specific vowels depending on the brain region of the electrode where some brain regions were selective to specific vowels.

However, one limitation of this work is the amount of data that we are working with. In fact, the limited amount of training and testing data could be a reason for low performance of our classifier, where we only have three testing data per vowel per channel. One way to get around this issue is to group all data from channels located in the same brain region in order to get more trials to train and test the classifier. However, there is no solid proof in literature around the generalization of neural behavior across specific brain regions. For that reason, we can not generalize that two channels should exhibit the same behavior solely based on the fact that they are in the same brain region.

Another alternative would be to investigate non-linear kernels for the SVM classifier or even completely different machine learning algorithms for vowel decoding. One method used in literature for phoneme decoding was PCA analysis and LDA.

In conclusion, this work is a benchmark that shows different neural behaviors depending on the identity of vowel produced across the brain and shows promise for phoneme decoding using neural activity.

REFERENCES

[1] Bohland, J. W. & Guenther, F. H. An fMRI investigation of syllable sequence production. *NeuroImage* **32**, 821–841 (2006)

[2] Basilakos, A., Smith, K. G., Fillmore, P., Fridriksson, J. & Fedorenko, E. Functional characterization of the human speech articulation network. *Cereb. Cortex* **28**, 1816–1830 (2017)

[3] Tourville, J. A., Nieto-Castañón, A., Heyne, M. & Guenther, F. H. Functional parcellation of the speech production cortex. *J. Speech Lang. Hear. Res.* **62**, 3055–3070 (2019)

[4] Lee, D. K., Fedorenko, E., Simon, M. V., Curry, W. T., Nahed, B. V., Cahill, D. P., & Williams, Z. M. (2018). Neural encoding and production of functional morphemes in the posterior temporal lobe. *Nature communications*, *9*(1), 1877.

[5] Glanz, O., Hader, M., Schulze-Bonhage, A., Auer, P. & Ball, T. A study of word complexity under conditions of non-experimental, natural overt speech production using ECoG. *Front. Hum. Neurosci.* **15**, 711886 (2021)

[6] Yellapantula, S., Forseth, K., Tandon, N. & Aazhang, B. NetDI: methodology elucidating the role of power and dynamical brain network features that underpin word production. *eNeuro* 8, ENEURO.0177-20.2020 (2020)

[7] Hoffman, P. Reductions in prefrontal activation predict off-topic utterances during speech production. *Nat. Commun.* **10**, 515 (2019)

[8] Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M. and Smith, S.M., 2016. A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), pp.171-178.

[9] Chang, Edward F., Garret Kurteff, John P. Andrews, Robert G. Briggs, Andrew K. Conner, James D. Battiste, and Michael E. Sughrue. "Pure apraxia of speech after resection based in the posterior middle frontal gyrus." *Neurosurgery* 87, no. 3 (2020): E383-E389.

[10] Fedorenko, Evelina, Terri L. Scott, Peter Brunner, William G. Coon, Brianna Pritchett, Gerwin Schalk, and Nancy Kanwisher. "Neural correlate of the construction of sentence meaning." *Proceedings of the National Academy of Sciences* 113, no. 41 (2016): E6256-E6262.

[11] Nelson, M.J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., Cash, S.S., Naccache, L., Hale, J.T., Pallier, C. and Dehaene, S., 2017. Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, *114*(18), pp.E3669-E3678.

[12] Walenski, M., Europa, E., Caplan, D. & Thompson, C. K. Neural networks for sentence comprehension and production: an ALE-based meta-analysis of neuroimaging studies. *Hum. Brain Mapp.* **40**, 2275–2304 (2019)

[13] Elin, Kirill, Svetlana Malyutina, Oleg Bronov, Ekaterina Stupina, Aleksei Marinets, Anna Zhuravleva, and Olga Dragoy. "A new functional magnetic resonance imaging localizer for preoperative language mapping using a sentence completion task: Validity, choice of baseline condition, and test–retest reliability." *Frontiers in Human Neuroscience* 16 (2022): 791577.

[14] Bouchard, K. E., Mesgarani, N., Johnson, K. & Chang, E. F. Functional organization of human sensorimotor cortex for speech articulation. *Nature* **495**, 327–332 (2013)

[15] Coudé, Gino, Pier Francesco Ferrari, Francesca Rodà, Monica Maranesi, Eleonora Borelli, Vania Veroni, Fabio Monti, Stefano Rozzi, and Leonardo Fogassi. "Neurons controlling voluntary vocalization in the macaque ventral premotor cortex." *PloS one* 6, no. 11 (2011): e26822.

[16] Hahnloser, R. H. R., Kozhevnikov, A. A. & Fee, M. S. An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature* **419**, 65–70 (2002)

[17] Aronov, D., Andalman, A. S. & Fee, M. S. A specialized forebrain circuit for vocal babbling in the juvenile songbird. *Science* **320**, 630–634 (2008)

[18] Stavisky, S.D., Willett, F.R., Wilson, G.H., Murphy, B.A., Rezaii, P., Avansino, D.T., Memberg, W.D., Miller, J.P., Kirsch, R.F., Hochberg, L.R. and Ajiboye, A.B., 2019. Neural ensemble dynamics in dorsal motor cortex during speech in people with paralysis. *Elife*, *8*, p.e46015.

[19] Tankus, A., Fried, I. & Shoham, S. Structured neuronal encoding and decoding of human speech features. *Nat. Commun.* **3**, 1015 (2012)

[20] Vyas, S., Golub, M. D., Sussillo, D. & Shenoy, K. V. Computation through neural population dynamics. *Ann. Rev. Neurosci.* **43**, 249–275 (2020)

[21] Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Ryu, S. I. & Shenoy, K. V. Cortical preparatory activity: representation of movement or first cog in a dynamical machine? *Neuron* **68**, 387–400 (2010)

[22] Shenoy, K. V., Sahani, M. & Churchland, M. M. Cortical control of arm movements: a dynamical systems perspective. *Ann. Rev. Neurosci.* **36**, 337–359 (2013)

[23] Kaufman, M. T., Churchland, M. M., Ryu, S. I. & Shenoy, K. V. Cortical activity in the null space: permitting preparation without movement. *Nat. Neurosci.* **17**, 440–448 (2014)

[24] Mugler, Emily M., James L. Patton, Robert D. Flint, Zachary A. Wright, Stephan U. Schuele, Joshua Rosenow, Jerry J. Shih, Dean J. Krusienski, and Marc W. Slutzky. "Direct classification of all American English phonemes using signals from functional speech motor cortex." *Journal of neural engineering* 11, no. 3 (2014): 035015.

[25] Angrick, M., Ottenhoff, M.C., Diener, L., Ivucic, D., Ivucic, G., Goulis, S., Saal, J., Colon, A.J., Wagner, L., Krusienski, D.J. and Kubben, P.L., 2021. Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity. *Communications biology*, *4*(1), p.1055.

[26] Penfield W, Roberts L: Speech and Brain-mechanisms. Princeton University Press: Princeton; 1959, .

[27] Chartier J, Anumanchipalli G K, Johnson K and Chang E F 2018 Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex Neuron 98 1042–54.e4

[28] Conant D, Bouchard K E and Chang E F 2014 Speech map in the human ventral sensory-motor cortex Curr. Opin. Neurobiol. 24 63–67

[29] Lotte, Fabien, Jonathan S. Brumberg, Peter Brunner, Aysegul Gunduz, Anthony L. Ritaccio, Cuntai Guan, and Gerwin Schalk. "Electrocorticographic representations of segmental features in continuous speech." *Frontiers in human neuroscience* 9 (2015): 97.

[30] Moses D A, Mesgarani N, Leonard M K and Chang E F 2016 Neural speech recognition: continuous phoneme decoding using spatiotemporal representations of human cortical activity J. Neural Eng. 13 056004

[31] Chan, A.M., Dykstra, A.R., Jayaram, V., Leonard, M.K., Travis, K.E., Gygi, B., Baker, J.M., Eskandar, E., Hochberg, L.R., Halgren, E. and Cash, S.S., 2014. Speech-specific tuning of neurons in human superior temporal gyrus. *Cerebral Cortex*, *24*(10), pp.2679-2693.

[32] Mesgarani N, Cheung C, Johnson K and Chang E F 2014 Phonetic feature encoding in human superior temporal gyrus Science 343 1006–10

[33] Thomas, Tessy M., Aditya Singh, Latané P. Bullock, Daniel Liang, Cale W. Morse, Xavier Scherschligt, John P. Seymour, and Nitin Tandon. "Decoding articulatory and phonetic components of naturalistic continuous speech from the distributed language network." *Journal of Neural Engineering* 20, no. 4 (2023): 046030.

[34] Kothe, C. Lab streaming layer (LSL). https://github.com/sccn/labstreaminglayer**26**, 2015 (2014).

[35] Landré, Elisabeth, Mathilde Chipaux, Louis Maillard, William Szurhaj, and Agnès Trébuchon. "Electrophysiological technical procedures." *Neurophysiologie Clinique* 48, no. 1 (2018): 47-52.

[36] Liu, Y., W. G. Coon, A. De Pesters, P. Brunner, and G. Schalk. "The effects of spatial filtering and artifacts on electrocorticographic signals." *Journal of neural engineering* 12, no. 5 (2015): 056008.

[37] Browman, C. P. & Goldstein, L. Articulatory gestures as phonological units. *Haskins Laboratories Status Report on Speech Research* **99**, 69–101 (1989)

[38] Fowler, C. A., Rubin, P. E., Remez, R. E. & Turvey, M. T. in *Language Production*: *Speech and Talk* Vol. 1 (ed. Butterworth, B.) 373–420 (Academic Press, 1980)

[39] Saltzman, E. & Munhall, K. A dynamical approach to gestural patterning in speech production. *Ecol. Psychol.* **1**, 333–382 (1989)

Chapter 4 CONCLUSION

4.1 Conclusion

The work presented in chapters 2 and three serves multiple purposes for future research in speech prosthesis. Hopefully, other researchers are now encouraged to investigate the use of sEEG recordings for phoneme decoding, that could further our understanding of speech representation in the brain and also serve in speech BCI applications.

We have shown that speech onset can be extracted from audio signals using energy and zero crossing rate. Also, we demonstrated that linear predictive coding is able to extract the formants from the vowel's spectra with precision. By the end of chapter 2, our work showed that vowel identity can be uniquely identified by the first two formant frequencies.

In the second part of this work, we first showed that brain states correlated with periods of speech and silence are statistically different. We further concluded the different brain regions responded at different times around speech onset. This spatial-temporal distribution of neural activity across the brain proves the intricate work of different brain regions to give rise to speech production. Finally, we showed that there exists a correlation between the vowel being produced and the neural activity, and that we can decode with more than 50% accuracy the vowel identity from neural activity and that different brain regions are more selective to specific phonemes. Phoneme decoding can thus be the first step to construct a full speech decoding prosthesis.

4.2 Future Work

Regarding the audio signal processing, future work should be targeted 2 words the completion of tasks two and three. For task 2, the first step was implemented in this thesis. The formant frequency is gathered using linear interpolation can be used to generate different sounds and be played to the subjects. This will help us gain insight into the psychometric curves of the

vowels. Other interpolation methods should be investigated and compared to see the effect on speech perception. For task 3, if we think of the sensory motor system for speech as a control system, one theory suggests that there is an afferent copy of the produced vocalization that could be compared against the perceived sound to result in a feedback error signal to the motor system to guide next production. One hypothesis is that the mechanism for perceiving error and generating feedback control signal during production (task 3) and replay (task 2) are the same. If true, we could leverage the model learned in task 2 to infer the error signal.

Regarding neural decoding of vowel identity, other machine learning algorithms should be investigated to improve the classification accuracy. More data should be collected so we can appropriately train and test the different models. Nonetheless, phoneme representation in the brain shows promise and should be further investigated another natural progression is to study the neural activity when producing words and sentences in order to develop a fully functional closed loop speech prosthesis.