Title

Cortical dynamics of speech motor sequencing and production

Author

Liu, Jessie Rachel

Publication Date

2023

Peer reviewed|Thesis/dissertation

Cortical dynamics of speech motor sequencing and production

by

Jessie R. Liu

DISSERTATION

Submitted in partial satisfaction of the requirements for degree of

DOCTOR OF PHILOSOPHY

in

Bioengineering

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO
AND
UNIVERSITY OF CALIFORNIA, BERKELEY

Approved:

DocuSigned by:
Edward F. Chang
4B5B40824E04415...

Edward F. Chang

Chair

DocuSigned by:
Jack L. GALLANT
[illegible]

Jack L. GALLANT

DocuSigned by:
John Houde
380962FCB1A7486...

John Houde

Committee Members

Acknowledgments

I first would like to thank my advisor and mentor Edward Chang. The chance I have been given to work and learn in this lab changed the trajectory of my career and transformed me as a scientist. Eddie is truly one of a kind—he is not only exceedingly talented in his career both as a neurosurgeon and as a scientist, but he is also a tremendously thoughtful and caring person. There are numerous moments where his enthusiasm and dedication, to the project and to mentoring me, kept me going through my PhD.

I'd also like to thank the members of my dissertation committee, Jack Gallant and John Houde, for their scientific guidance and advice as I wrote this dissertation.

I'd like to thank my talented colleagues—Margaret Seaton, Sean Metzger, Alexander Silva, Ilina Bhaya-Grossman, Maximilian Dougherty, Patrick Hullett, Lingyun Zhao, Deborah Levy, and so many others. The amount of talent that surrounds me is just outstanding and I know you will all continue to excel. What a bonus to find camaraderie, friendship, and joy in working with you all. I also thank Viv Her for all the financial and logistical support that has enabled me to perform and share my research with the broader neuroscience community.

I profusely thank the patients who participated in our research, and especially to those who participated in the stimulation research. This would not have been possible without you. It cannot be overstated how critical your hard work, patience, and generosity has been to all of our research.

To Pancho, Robby, and Ann—I will never be able to thank you enough. Your dedication to the project, your hard work, and patience with us as we navigate uncharted waters, are the key components to the success of this project. We would not have been able to achieve any of this research without you. Though few people may read these acknowledgements, it's important to me that I impress upon anyone who does, just how special this group of people are. They are some of the most dedicated, talented, and kind-hearted people I have ever met. Simply working with them day to day, even when the decoders weren't working or when we encountered numerous technical difficulties, was a joy in and of itself. It doesn't hurt that we are so lucky to have participants with such great senses of humor, which ensures our recording sessions are full of smiles. I hope you take immense pride in yourselves and everything we've done together. I can't wait to see what we do next!

I'd also like to thank the many professors from the University of Pittsburgh, my alma mater, who played a role in my growth as a scientist. In particular, Carsten Stuckenholz, Arash Mahboobin, and George Stetten. I'd also like to specifically thank Aaron Batista for first introducing me to computational neuroscience and brain computer interface research, and for encouraging me to pursue this research. I'd also like to thank Sanjeev Shroff whose sage advice of "If all else is equal, go for the new experience" cemented my decision to attend this program, and is advice I come back to often. I'd like to thank Mike Modo who gave me my first research opportunity and gave me an incredible amount of freedom to learn about all parts of the research process. I'd also like to thank my Modo Lab colleagues Francesca Nicholls and Harman Ghuman who provided valuable camaraderie and mentorship.

I am indebted to my parents, Amanda and Dongping, who have unwaveringly supported me. My parents have always given me the freedom to pursue whatever career I wanted and never made me feel like any choice would be inadequate. My work ethic, my ambition, and my drive I get from them both. They have played a critical role in my success. I love you both, so much. I'd like to thank my siblings, whose friendship and visits to San Francisco kept me sane. In particular, I thank Jaimie who can always make me laugh. I also thank my grandmother Lois, my grandfather James, and Yeye who instilled in me the importance of education and encouraged my interests in engineering.

I thank my friends Adam Smoulder, Andrew Sivaprakasam, Sarah Shaykevich, and Henry Phalen—The Beansz. 6am beans feels just like yesterday and also like a lifetime ago. Though much has changed in our lives since the days of B09 and 320 shenanigans, I am so grateful that the beans are a constant. I also thank my day 1 Cohort 2017 friends–Louise Hansen and Alex Beltran. I am so grateful for your friendship and companionship as we navigated grad school.

And finally, I thank this beautiful earth, in particular Mount Tamalpais, Point Reyes National Seashore (especially Sky Trail and Kelham Beach), and Grand Canyon National Park. Their beauty not only kick-started my love of hiking and being outdoors, but granted me mental refuge during periods where this dissertation and my PhD felt insurmountable.

Contributions

This thesis contains material that has been previously published in peer-reviewed journals. Namely, Chapter 1 is directly adapted from:

> David A. Moses*, Sean L. Metzger*, **Jessie R. Liu***, et al. (2021). Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria. *New England Journal of Medicine*, 385(3), 217-227. doi: 10.1056/NEJMoa2027540.

Chapter 2 is directly adapted from:

> Sean L. Metzger*, **Jessie R. Liu***, David A. Moses*, et al. (2022). Generalizable spelling using a speech neuroprosthesis in an individual with severe limb and vocal paralysis. *Nature Communications*, 13(6510). doi: 10.1038/s41467-022-33611-3.

* Denotes equal contributions.

Finally, Chapter 3 is directly adapted from a currently unpublished manuscript. Personal contributions are further detailed at the start of each chapter with a disclaimer concerning previous or future publication.

Abstract

Cortical dynamics of speech motor sequencing and production

Jessie R. Liu

Speech is one of the most efficient and effortless ways to communicate. Producing speech requires planning speech targets, sequencing speech-motor movements, and coordinating a dynamic system of articulators to shape breath in real time, generating the sounds we perceive and interpret as needs, ideas, and emotions. Loss of this ability, through neurodegenerative disease or paralysis, is devastating and reduces self-reported quality of life. For many with this condition, the cortical signals to control their articulators still persist–however these signals cannot be communicated to their vocal tract, due to injured or diseased descending pathways leading to vocal tract paralysis. Recent advances in neural recording hardware, our understanding of how speech production is controlled in the brain (specifically in the ventral sensorimotor cortex), and machine learning have enabled the development of speech brain computer interfaces (BCI). A direct speech BCI would translate neural signals into intended speech, restoring the ability to communicate to these individuals. This body of work first demonstrates a proof of concept that a direct speech BCI consisting of a 50-word vocabulary can be developed using high-density neural recording hardware, called electrocorticography, in a participant who cannot speak due to severe paralysis. We then built upon this proof-of-concept by developing a spelling-based speech BCI that could be controlled by silently

attempted speech. The methods we used for these studies were based on our understanding of how articulatory movements are controlled by neural signals. However, much less is known about how the brain controls the upstream processes of planning and sequencing these movements. Motivated by the success of translating neuroscientific findings into BCI development, we next sought to understand how speech is sequenced in the brain. Using a task where healthy speakers spoke syllable sequences of varying complexity, we found both neural activity specific to production and widespread sustained activity associated with planning syllable sequences. This network, consisting both of areas classically considered to be involved in speech planning, such as Broca's area, as well as more novel regions like the middle precentral gyrus (mPrCG), was modulated by the complexity of the sequences. However, only the mPrCG demonstrated robust sustained activity, sequence complexity encoding, and was correlated with the participants reaction time, suggesting that this area's role in speech planning is specific to speech-motor sequencing. We confirmed this by using direct cortical stimulation, which induced speech errors only during complex sequences, in the absence of direct motor or perceptual effects. This work establishes the mPrCG as a critical node of speech-motor sequencing, redefining traditional notions of how the brain sequences and produces speech. Together, these studies demonstrate the potential of speech brain computer interfaces for restoring speech in paralyzed individuals and puts forth new possibilities for neurobiologically informed algorithms for decoding speech.

Contents

List of Figures

List of Tables

Introduction

The ability to communicate is a critical part of life and the most natural and efficient way that we accomplish this is by speaking. Speaking allows us not only to express our emotions and build relationships with loved ones, but it is also critical for advocating for ourselves and is one tool that enables us to have agency in our environments. Producing speech involves dynamically controlling the muscles in our vocal tract which in turn shape the air that we carefully exhale. Though producing speech is an incredibly complex process we are able to initiate and articulate speech almost instantly.

Losing the ability to speak, such as through neurodegenerative disease or severe paralysis, can be devastating. Existing alternative or augmentative communication techniques are much slower and often fatiguing or frustrating to use for long periods of time. This loss or severe reduction in one's ability to communicate reduces self-reported quality of life (Felgoise et al. 2016). In many cases, individuals are cognitively intact and the brain signals required to produce speech still persist. It is because of diseased or injured descending pathways that these brain signals are not properly communicated to the vocal tract, leading to vocal tract paralysis and the inability to speak intelligibly. Given that these brain signals still exist, technology that could translate these signals into the intended speech (such as phrases or sentences that the person was attempting to say) has the potential to restore communica-

tion in individuals with this condition. Developing this kind of technology presents several obstacles including how to record neural activity, which brain areas to record from, and the appropriate algorithms to use to translate the activity into speech. While engineering advances are undoubtedly crucial to the first and third hurdles, understanding the system also informs the latter two.

Academics and neurosurgeons have long hypothesized about how the brain controls this dynamic system, with seminal case studies as early as 1861 (Broca 1861; Penfield et al. 1937). Early on, models of what areas of the brain played a role in speech production relied on lesion studies, where post-mortem studies of patients with language disorders were used to link brain damage with observed deficits (Broca 1861). Early models of speech production were built on behavioral and linguistic studies that importantly began to hypothesize about what specific processes make up speech production (Levelt 1993; MacNeilage 1998). Though both direct neurosurgical case studies and behavioral models are important to understanding speech production, the ability to simultaneously record cortical activity and behavior during speech production has been critical in advancing our understanding of how the brain facilitates speech production. Noninvasive methods, like functional magnetic resonance imaging (fMRI), have enabled us to measure whole brain activity with varying levels of spatial resolution. An advantage of noninvasive methods is that any person's brain can be measured, with no medical risk, and a high degree of coverage can be achieved. However, speech is a fast and dynamic process, and noninvasive methods can be limited by signal-to-noise ratio and temporal resolution (Chang 2015).

Invasive recording methods, such as electrocorticography (ECoG), require neurosurgery to place but yield neural signals with high signal-to-noise ratio at a high temporal resolution (Chang 2015). Placement of ECoG grids is often used for the surgical treatment of drug resistant epilepsy, in order to precisely map where a patient's seizures are originating from and to determine whether surgical intervention is in danger of disrupting speech or motor function. Temporary placement (intraoperatively or for acute hospital stays on the order of one week) of these grids and patients' consent to participate in research offers a rare opportunity to record both high fidelity neural activity and behavior simultaneously. Studies using this type of data have significantly advanced our understanding of how speech production is represented in the brain.

One area important to controlling articulation is the ventral sensorimotor cortex (vSMC). The vSMC comprises the ventral portions of both the pre- and postcentral gyri. This area has been linked to the control of articulation at various levels, such as continuous articulatory movements, discrete phonemes, and articulatory gestures (Chartier et al. 2018; Mugler et al. 2018). In parallel to advancing the neuroscientific understanding of speech production, these representations can be leveraged to decode intended speech from brain signals. Speech decoding was first investigated in healthy speakers, with some models explicitly leveraging articulatory representations (Makin et al. 2020; Sun et al. 2020; Anumanchipalli et al. 2019; Herff et al. 2015). However, it was unknown whether these cortical articulatory representations would persist years after paralysis and whether these methods would generalize to individuals who are unable to fluently coordinate their articulatory movements due to severe

paralysis. To this end, we started a clinical trial (the BCI Restoration of Arm and Voice, or BRAVO) to study a chronic ECoG-based speech BCI in participants who are unable to speak.

In **Chapter 1**, we describe a proof-of-concept study demonstrating the first successful direct-word speech BCI (this chapter is directly adapted from Moses*, Metzger*, Liu*, et al. 2021). This work involved our first clinical trial participant, Bravo-1, who had an ECoG grid surgically placed over primarily the vSMC in the spring of 2019. Due to a brainstem stroke, Bravo-1 is both severely paralyzed and has been diagnosed with anarthria, or the inability to articulate speech. Bravo-1 has extremely limited control over his facial muscles and vocal tract and can only make unintelligible noises at a slow rate. We recorded neural activity while Bravo-1 attempted to say 50 English words out loud. Using artificial neural networks, we trained two models–a speech detector and a speech classifier. The speech detection model predicted when Bravo-1 was attempting to say a word, only using neural activity. When a speech attempt was detected, this would pass the relevant window of neural activity to the classifier which would predict which of the 50 words was being said. Finally, we also leveraged the statistical structure of English by applying a natural language model to the predicted sequences of words. This model would correct unlikely sequences of decoded words.

Our decoding framework enabled fast, flexible decoding of a limited vocabulary that can be invoked and disengaged volitionally just by Bravo-1's natural speech attempts. This served as an important proof-of-concept in several regards. First, all previous human BCI studies used intracortical arrays, where the implanted recording hardware consists of small

needle-like electrodes that penetrate the cortex and record from single neurons. While this greatly enhances spatial resolution, these arrays cover much smaller areas of cortex and are often prone to signal instability. ECoG had been theorized to have greater signal stability, but had not yet been tested for efficacy in BCI studies targeting speech. Here, we demonstrate that it indeed feasible to have a chronic ECoG-based BCI. Second, there had previously been two intracortical BCI studies decoding speech, but the accuracies were too low to be clinically viable. Here, for a limited vocabulary, we show that high accuracy can be achieved. Finally, this work demonstrated that even after 15 years of paralysis, neural signals correlated with attempted articulation still persist.

Though this first work was an important milestone, it only applied to a small vocabulary and required Bravo-1 to attempt to vocalize, which is fatiguing for him. In **Chapter 2**, we investigated whether we could use silent speech attempts to control a spelling system (this chapter is directly adapted from Metzger*, Liu*, Moses* et al. 2022). The advantages here are two-fold. Silently attempted speech refers to speech attempts where one's mouth could be moving, but there is no vocalization. This is different from completely imagined speech attempts, which involve no orofacial movement at all. Using silent speech attempts is advantageous because they are not as effortful and therefore are often faster. Second, using a spelling system has the potential to generalize to an unlimited vocabulary instead of being limited to a finite-sized vocabulary.

We designed a spelling system similar to our decoding framework in **Chapter 1**, using a speech detector, speech classifier, and language modeling, with some key differences.

In this work, we had Bravo-1 use the NATO phonetic alphabet to spell out words. The NATO alphabet was developed with phonetic discriminability in mind (e.g. "Alpha" for "A", "Bravo" for "B", and so on), and so we hypothesized that speech attempts using this set, as opposed to singular letters (e.g. "A") would yield better discriminability of the neural data. Additionally, we incorporated more neural features in our decoding. Previous ECoG work predominantly used neural activity in the high-gamma range (from 70 to 150 Hz) though other frequency bands are also known to hold useful information (Mugler et al. 2018; Sun et al. 2020; Proix et al. 2022; Anumanchipalli et al. 2019). Here, we included low-frequency signals (from 0.3 to 100 Hz) in addition to high-gamma activity. And finally, all of Bravo-1's speech attempts were silently attempted. We found that even when not vocalizing, silently attempted speech evoked neural activity patterns in the vSMC. Further, we found that NATO code words could be decoded from neural activity with high accuracy, aided by the combination of low and high frequency signals. Though we only tested the system in real time with a vocabulary of about 1000 words, offline simulations showed that we could generalize to vocabularies of over 9000 words with no significant loss of accuracy.

These first two chapters served as important groundwork for establishing ECoG-based methods of speech decoding for participants who cannot speak. Part of what contributed to the success of these methods was our understanding of how articulation is represented in the vSMC. While there are several engineering avenues to improving speech BCI methods, including improving hardware and decoding algorithms, improving our understanding of the system we're trying to mimic, the process of speech production, can also be important.

While we know much more today about how articulation might be controlled in the brain, we know far less about the potential upstream processes involved before the actual articulation of intended speech targets. There are many potential processes that are crucial to speech production and occur upstream of articulation, including conception of ideas and lexical access (e.g. word choice), but we chose to focus on speech-motor planning as it may be the process just before continuous articulation (Levelt 1993; Guenther et al. 2016). Though there are many linguistic theories of how speech-motor movements are planned and sequenced, we know very little about what brain areas and temporal dynamics facilitate this process.

In **Chapter 3**, we investigated the process of speech-motor sequencing in healthy speakers. Using ECoG grids placed for the surgical treatment of epilepsy, we recorded from multiple cortical areas implicated in various aspects of speech production. During recording, participants read a target sequence presented to them on a computer screen, then waited a short delay, before being given a go-cue to repeat what they had read. We found that this task evoked not only phasic activity associated with articulation, but also widespread sustained activity. That is, we identified a network across multiple cortical areas where activity that was evoked by reading the sequence, remained sustained throughout the delay period, during the period just before they started speaking, as well as during speech. This network involved regions classically thought to be involved in speech planning and sequencing, such as Broca's area (Bohland et al. 2006; Guenther et al. 2016; Hickok et al. 2022), but remarkably it was the precentral gyrus with the most robust sustained activity. Further,

we specifically modulated the sequence complexity at the syllable and phoneme levels of the target sequences, referred to as sequence and articulatory complexity respectively. We found that an area in the middle of the precentral gyrus (which we term the mPrCG) superior to the vSMC, most consistently encoded sequence complexity, while other areas more transiently encoded sequence complexity. Importantly, this area had not been previously considered to be involved in speech-motor planning. It was this area alone that, in addition to sustained activity encoding sequence complexity, also correlated with behavioral aspects of speech production such as reaction time. These results suggest that sequence complexity in the mPrCG is specific to speech-motor sequencing, rather than reflecting higher order levels of processing.

These results represent new insights into the neural correlates of speech-motor sequencing. With ECoG, we have the additional advantage of being able not only to record neural activity but also to deliver stimulation. Stimulation mapping is commonly used to map out which brain areas are critical to speech or motor functions (Lu et al. 2021; Leonard et al. 2019)—delivering stimulation disrupts normal brain circuits, thus identifying which areas are necessary and causal for a particular function. In four of our participants, we stimulated putative sequencing sites in the mPrCG. Remarkably, we found that stimulation caused speech errors only on complex syllable sequences, directly mirroring the sequence complexity encoding we had observed, in the absence of direct motor effects, perceptual effects, or other known effects such as speech arrest. In contrast, stimulation in other brain areas with sequence complexity encoding did not yield speech errors. Strikingly, speech errors

induced by stimulation of the mPrCG resembled those found in apraxia of speech (AOS), which is a clinical speech disorder hypothesized to occur from a deficit in speech motor programming or sequencing (Strand et al. 2014). Though AOS was originally thought to result from damage to the insula or to Broca's area, more recent lesion studies and resection case studies have instead suggested that AOS arises from damage to the mPrCG (Itabashi et al. 2016; Graff-Radford et al. 2014; Chang et al. 2020; Levy et al. 2023). We provide, to our knowledge, the first neurobiological link between speech-motor sequencing, the mPrCG, and AOS. These results force us to reconsider classical models of speech production that do not account for sustained activity or that attribute sequencing to Broca's area.

These works demonstrate a key proof of concept that speech BCIs can be used to restore speech in people with paralysis and highlight a novel speech-motor circuit that offers potential to further improve naturalistic speech decoding. Together, they represent a culmination of machine learning, neurobiological, and neurosurgical techniques that have made it possible to investigate these topics. These results set a foundation for future studies of speech BCI in people who cannot speak, and for developing neurobiologically informed methods of decoding.

References

Anumanchipalli, Gopala K., Josh Chartier, and Edward F. Chang (Apr. 2019). "Speech synthesis from neural decoding of spoken sentences". *Nature* 568.7753, pp. 493–498. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-019-1119-1.

Bohland, Jason W. and Frank H. Guenther (Aug. 2006). "An fMRI investigation of syllable sequence production". *NeuroImage* 32.2, pp. 821–841. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2006.04.173.

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

Chang, Edward F. (Apr. 2015). "Towards Large-Scale, Human-Based, Mesoscopic Neurotechnologies". *Neuron* 86.1, pp. 68–78. ISSN: 08966273. DOI: 10.1016/j.neuron.2015.03.037.

Chang, Edward F., Garret Kurteff, John P. Andrews, et al. (Sept. 1, 2020). "Pure Apraxia of Speech After Resection Based in the Posterior Middle Frontal Gyrus". *Neurosurgery* 87.3, E383–E389. ISSN: 0148-396X, 1524-4040. DOI: 10.1093/neuros/nyaa002.

Chartier, Josh, Gopala K. Anumanchipalli, Keith Johnson, and Edward F. Chang (2018). "Encoding of Articulatory Kinematic Trajectories in Human Speech Sensorimotor Cortex". *Neuron* 98.5, 1042–1054.e4. DOI: 10.1016/j.neuron.2018.04.031.

Felgoise, Stephanie H., Vincenzo Zaccheo, Jason Duff, and Zachary Simmons (May 18, 2016). "Verbal communication impacts quality of life in patients with amyotrophic lateral sclerosis". *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* 17.3, pp. 179–183. ISSN: 2167-8421, 2167-9223. DOI: 10.3109/21678421.2015.1125499.

Graff-Radford, Jonathan, David T. Jones, Edythe A. Strand, et al. (Feb. 2014). "The neuroanatomy of pure apraxia of speech in stroke". *Brain and Language* 129, pp. 43–46. ISSN: 0093934X. DOI: 10.1016/j.bandl.2014.01.004.

Guenther, Frank H. and Gregory Hickok (2016). "Neural Models of Motor Speech Control". *Neurobiology of Language*. Elsevier, pp. 725–740. ISBN: 978-0-12-407794-2.

Herff, Christian, Dominic Heger, Adriana de Pesters, et al. (2015). "Brain-to-text: decoding spoken phrases from phone representations in the brain". *Frontiers in Neuroscience* 9 (June), pp. 1–11. DOI: 10.3389/fnins.2015.00217.

Hickok, Gregory, Jonathan Venezia, and Alex Teghipco (Nov. 30, 2022). "Beyond Broca: neural architecture and evolution of a dual motor speech coordination system". *Brain*, awac454. ISSN: 0006-8950, 1460-2156. DOI: 10.1093/brain/awac454.

Itabashi, Ryo, Yoshiyuki Nishio, Yuka Kataoka, et al. (Jan. 2016). "Damage to the Left Precentral Gyrus Is Associated With Apraxia of Speech in Acute Stroke". *Stroke* 47.1, pp. 31–36. ISSN: 0039-2499, 1524-4628. DOI: 10.1161/STROKEAHA.115.010402.

Leonard, Matthew K., Ruofan Cai, Miranda C. Babiak, et al. (June 2019). "The peri-Sylvian cortical network underlying single word repetition revealed by electrocortical stimulation

and direct neural recordings". *Brain and Language* 193, pp. 58–72. ISSN: 0093934X. DOI: 10.1016/j.bandl.2016.06.001.

Levelt, Willem J. M. (1993). *Speaking: From Intention to Articulation*. The MIT Press. ISBN: 978-0-262-27822-5. DOI: 10.7551/mitpress/6393.001.0001.

Levy, Deborah F., Alexander B. Silva, Terri L. Scott, et al. (Mar. 27, 2023). "Apraxia of speech with phonological alexia and agraphia following resection of the left middle precentral gyrus: illustrative case". *Journal of Neurosurgery: Case Lessons* 5.13, CASE22504. ISSN: 2694-1902. DOI: 10.3171/CASE22504.

Lu, Junfeng, Zehao Zhao, Jie Zhang, et al. (Sept. 4, 2021). "Functional maps of direct electrical stimulation-induced speech arrest and anomia: a multicentre retrospective study". *Brain* 144.8, pp. 2541–2553. ISSN: 0006-8950, 1460-2156. DOI: 10.1093/brain/awab125.

MacNeilage, Peter F. (Aug. 1998). "The frame/content theory of evolution of speech production". *Behavioral and Brain Sciences* 21.4, pp. 499–511. ISSN: 0140-525X, 1469-1825. DOI: 10.1017/S0140525X98001265.

Makin, Joseph G., David A. Moses, and Edward F. Chang (Apr. 2020). "Machine translation of cortical activity to text with an encoder–decoder framework". *Nature Neuroscience* 23.4, pp. 575–582. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/s41593-020-0608-8.

Metzger, Sean L., Jessie R. Liu, David A. Moses, et al. (Nov. 8, 2022). "Generalizable spelling using a speech neuroprosthesis in an individual with severe limb and vocal paralysis". *Nature Communications* 13.1, p. 6510. ISSN: 2041-1723. DOI: 10.1038/s41467-022-33611-3.

Moses, David A., Sean L. Metzger, Jessie R. Liu, et al. (July 15, 2021). "Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria". *New England Journal of Medicine* 385.3, pp. 217–227. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMoa2027540.

Mugler, Emily M., Matthew C. Tate, Karen Livescu, et al. (2018). "Differential Representation of Articulatory Gestures and Phonemes in Precentral and Inferior Frontal Gyri". *The Journal of Neuroscience* 4653, pp. 1206–18. DOI: 10.1523/JNEUROSCI.1206-18.2018.

Penfield, Wilder and Edwin Boldrey (1937). "Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation". *Brain* 60.4, pp. 389–443. ISSN: 0006-8950, 1460-2156. DOI: 10.1093/brain/60.4.389.

Proix, Timothée, Jaime Delgado Saa, Andy Christen, et al. (Jan. 10, 2022). "Imagined speech can be decoded from low- and cross-frequency intracranial EEG features". *Nature Communications* 13.1, p. 48. ISSN: 2041-1723. DOI: 10.1038/s41467-021-27725-3.

Strand, Edythe A., Joseph R. Duffy, Heather M. Clark, and Keith Josephs (Sept. 2014). "The apraxia of speech rating scale: A tool for diagnosis and description of apraxia of speech". *Journal of Communication Disorders* 51, pp. 43–50. ISSN: 00219924. DOI: 10.1016/j.jcomdis.2014.06.008.

Sun, Pengfei, Gopala K Anumanchipalli, and Edward F Chang (Dec. 1, 2020). "Brain2Char: a deep architecture for decoding text from brain recordings". *Journal of Neural Engineering* 17.6, p. 066015. ISSN: 1741-2560, 1741-2552. DOI: 10.1088/1741-2552/abc742.

Chapter 1

Neuroprosthesis for decoding speech in a paralyzed person with anarthria

Disclaimer: This chapter is a direct adaptation of the following article. Supplementary material is not included in this adaptation, but is available online.

David A. Moses*, Sean L. Metzger*, **Jessie R. Liu***, et al. (2021). Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria. *New England Journal of Medicine*, 385(3), 217-227. doi: 10.1056/NEJMoa2027540.

* Denotes equal contribution.

Personal contributions: I trained and developed the real-time speech detection models and performed speech detection performance, electrode contribution, and neural stability analyses. With David A. Moses and Sean L. Metzger, I collected data, edited all figures, and with Edward F. Chang we wrote the original draft of the manuscript with input from all authors.

1.1 Abstract

Background: Technology to restore communication for paralyzed patients who have lost the ability to speak has the potential to improve autonomy and quality of life. Decoding words and sentences directly from the neural activity of a paralyzed individual who cannot speak may be an improvement over existing methods for assisted communication.

Methods: We implanted a high-density, subdural multi-electrode array over the speech motor cortex of a person with anarthria, the loss of the ability to articulate speech, and spastic quadriparesis caused by brainstem stroke. Across 48 sessions, we recorded 22 hours of cortical activity while the participant attempted to say individual words from a 50-word vocabulary. Using deep learning, we created computational models to detect and classify words from patterns in the recorded cortical activity. We applied these models and a language model, which describes how frequently certain word sequences occur in natural language, to decode full sentences as he attempted to say them.

Results: We decoded sentences from the participant's cortical activity in real time at a median rate of 15 words per minute with a median word error rate of 26%. In post-hoc analyses, we detected 98% of individual word production attempts and classified words with 47% accuracy using cortical signals that were stable throughout the 81-week study period.

Conclusions: In a person with anarthria caused by brainstem stroke, we used machine learning and a natural language model to decode words and sentences directly from cortical activity as the person attempted to speak.

1.2 Introduction

Anarthria is the loss of the ability to articulate speech. It can result from a variety of conditions, including stroke and amyotrophic lateral sclerosis (Beukelman et al. 2007). Patients with anarthria may have intact language and cognition, and some are able to produce limited oral movements and undifferentiated vocalizations when attempting to speak, but neuromuscular disorder prevents speech (Nip et al. 2017). For paralyzed individuals with severe movement impairment who are unable to operate assistive devices, it hinders communication with family, friends, and caregivers, reducing self-reported quality of life (Felgoise et al. 2016).

Advances have been made with typing-based brain-computer interfaces that allow speech-impaired individuals to spell out messages using cursor control (Sellers et al. 2014; Vansteensel et al. 2016; Pandarinath et al. 2017; Brumberg, Pitt, et al. 2018; Linse et al. 2018). However, letter-by-letter selection interfaces driven by neural signal recordings are slow and effortful. A more efficient and natural approach may be to directly decode whole words from brain areas that control speech. Our understanding of how the speech motor cortex orchestrates the rapid articulatory movements of the vocal tract has expanded (Bouchard et al. 2013; Lotte et al. 2015; Guenther and Hickok 2016; Emily M Mugler et al. 2014; Chartier et al. 2018; Salari et al. 2019). Engineering efforts have leveraged these findings and advances in machine learning to demonstrate that speech can be decoded from brain activity in people without speech impairments (Herff et al. 2015; Angrick et al. 2019; Anumanchipalli et al.

2019; David A. Moses, Metzger, et al. 2021; Makin et al. 2020).

For paralyzed individuals who cannot speak, neural activity cannot be precisely aligned with intended speech due to the absence of speech output, posing an obstacle for training computational models (Martin et al. 2018). In addition, it is unclear whether neural signals underlying speech control are still intact in individuals who have not spoken for years or decades. In earlier work, a paralyzed person used an implanted intracortical two-channel microelectrode device and an audiovisual interface to generate vowel sounds and phonemes but not full words (Guenther, Brumberg, et al. 2009; Brumberg, Wright, et al. 2011).

To determine if speech can be directly decoded from the neural activity of a person who is unable to speak, we tested real-time word and sentence decoding from the cortical activity of a person with limb paralysis and anarthria resulting from brainstem stroke.

1.3 Methods

Trial overview

This work was performed as part of the BRAVO study (BCI Restoration of Arm and Voice function, clinicaltrials.gov, NCT03698149), which is a single-institution clinical trial to evaluate the potential of electrocorticography, a method for recording neural activity from the cerebral cortex using electrodes placed on the surface of the brain, and custom decoding techniques for communication and movement restoration. The device used in this study

received Investigational Device Exemption approval by the United States Food and Drug Administration. At the time of writing, only one participant has been implanted with the device. Due to regulatory and clinical considerations concerning proper handling of the percutaneous connector, the participant did not have the opportunity to use the system independently for daily activities.

This work was approved by the UCSF Committee on Human Research and supported in part by a research contract under Facebook's Sponsored Academic Research Agreement. Only the authors were involved in the design and execution of the clinical trial; the collection, storage, analysis, and interpretation of the data; and the writing of the manuscript and decision to publish it. No study hardware or data were transferred to any sponsor, and we did not receive any hardware or software from a sponsor to use in this work. There were no agreements between the authors and any sponsor restricting the authors' analysis or publication of the data. All authors confirm that the clinical study, data, analyses, and reporting of outcomes are valid and adhere to the protocol.

Participant

The participant is a right-handed male who was 36 years old at the start of the study. At age 20, he suffered extensive bilateral pontine strokes associated with a right vertebral artery dissection, which resulted in severe spastic quadriparesis and anarthria as confirmed by a speech language pathologist and neurologists.

He is cognitively intact, scoring 26 out of 30 points on the Mini-Mental Status Exam and being physically incapable of scoring the remaining 4 points due to his paralysis. He is able to vocalize grunts and moans but unable to produce intelligible speech. He has unimpaired eye-movement control. He normally communicates using an assistive computer-based typing interface controlled by his residual head movements, with typing rates at approximately 5 correct words or 18 correct characters per minute.

Implant device

The neural implant used to acquire brain signals from the participant is a customized combination of a high-density electrocorticography electrode array (PMT Corporation, MN, USA) and a percutaneous connector (Blackrock Microsystems, UT, USA). The rectangular electrode array has a length of 6.7 cm, width of 3.5 cm, and thickness of 0.51 mm and consists of 128 flat, disc-shaped electrodes with 4-mm center-to-center spacing arranged in a 16-by-8 lattice formation. During surgical implantation, the participant was put under general anesthesia and the left-hemisphere speech sensorimotor cortex, identified using anatomical landmarks of the central sulcus, was exposed via craniotomy. The electrode array was then laid on the surface of the brain in the subdural space. The electrode coverage enabled sampling from multiple cortical regions that have been implicated in speech processing, including portions of the left precentral gyrus, postcentral gyrus, posterior middle frontal gyrus, and posterior inferior frontal gyrus (Bouchard et al. 2013; Chartier et al. 2018; Guenther and

Hickok 2016; Emily M. Mugler et al. 2018). The dura was sutured closed and the cranial bone flap was replaced. The percutaneous connector was placed extracranially on the contralateral skull convexity and anchored to the cranium. This percutaneous connector conducts cortical signals from the implanted electrode array through externally accessible contacts to a detachable digital link and cable, enabling transmission of the acquired brain activity to a computer (Figure 1.6). The participant underwent surgical implantation of the device in February 2019 and had no complications. The procedure lasted approximately 3 hours. We began collection of data for this study in April 2019. Neural data acquisition and real-time processing

Using a digital signal processing system (NeuroPort System, Blackrock Microsystems), signals from all 128 electrodes of the implant device were acquired and transmitted to a separate computer running custom software for real-time analysis (Figure 1.6, Figure 1.7 (David A Moses et al. 2018; David A. Moses, Leonard, et al. 2019). Informed by previous research that has correlated neural activity in the 70–150 Hz (high gamma) frequency range with speech motor processing (Bouchard et al. 2013; Chartier et al. 2018; Emily M. Mugler et al. 2018; David A. Moses, Leonard, et al. 2019; Salari et al. 2019), we measured high gamma activity for each channel on this separate computer to use in all subsequent analyses and during real-time decoding.

Task design

The study consisted of 55 sessions over 81 weeks and took place at the participant's residence or in a nearby office. The participant engaged in two tasks: an isolated word task and a sentence task (Figure 1.8).

On average, we collected approximately 27 minutes of neural data with these tasks during each session. In each trial of each task, the participant was visually presented with a target word or sentence as text on a screen and then attempted to produce (say aloud) that target.

In the isolated word task, the participant attempted to produce individual words from a set of 50 English words. This word set contained common English words that can be used to create a variety of sentences, including words that are relevant to caregiving and words requested by the participant. In each trial, the participant was presented with one of these 50 words, and, after a 2-second delay, he attempted to produce that word when the word text on the screen turned green. We collected a total of 9800 trials of the isolated word task with the participant across 48 sessions throughout the study period.

In the sentence task, the participant attempted to produce word sequences from a set of 50 English sentences consisting only of words from the 50-word set. In each trial, the participant was presented with a target sentence and attempted to produce the words in that sentence (in order) at the fastest rate that he was comfortably able to. Throughout the trial, the word sequence decoded from neural activity was updated in real time and displayed as feedback to the participant. We collected a total of 250 trials of the sentence task with

the participant across 7 sessions at the end of the study period. A conversational variant of this task, in which the participant was presented with prompts and attempted to respond to them, is depicted in Figure 1.1.

Modeling

We used neural activity collected during the tasks to train, optimize, and evaluate custom models. Specifically, we created speech detection and word classification models that both leveraged deep learning techniques to make predictions from the neural activity. To decode sentences from the participant's neural activity in real time during the sentence task, we used a decoding approach containing these two models, a language model, and a Viterbi decoder, which are all described below (Figure 1.1).

The speech detection model processed each time point of neural activity during a task and detected onsets and offsets of attempted word production events in real time (Figure 1.9). We fit this model using only neural data and task timing information from the isolated word task.

For each detected event, the word classification model predicted a set of word probabilities by processing the neural activity spanning from 1 second before to 3 seconds after the detected onset (Figure 1.10). The predicted probability associated with each word in the 50-word set quantified how likely it was that the participant was attempting to say that word during the detected event. We fit this model using neural data from the isolated word

task.

In English, certain sequences of words are more likely than others. To use this underlying linguistic structure, we created a language model that yielded next-word probabilities given the previous words in a sequence (Kneser et al. 1995; Chen et al. 1999).

We trained this model on a collection of sentences consisting only of words from the 50-word set, which was obtained using a custom task on a crowdsourcing platform.

We used a custom Viterbi decoder as the final component in the decoding approach, which is a type of model that determines the most likely sequence of words given predicted word probabilities from the word classifier and word sequence probabilities from the language model (Viterbi 1967, Figure 1.11). By incorporating the language model, the Viterbi decoder was capable of decoding more plausible sentences than what would result from simply stringing together the predicted words from the word classifier.

Evaluations

To evaluate the performance of our decoding approach, we analyzed the sentences that were decoded in real time using two metrics: word error rate and words per minute. The word error rate of a decoded sentence is defined as the number of word errors made by the decoder divided by the number of words in the target sentence. Words per minute is equal to the number of words that were decoded per minute of neural data.

To further characterize the detection and classification of word production attempts from

the participant's neural activity, we processed the collected isolated word data with the speech detection and word classification models in offline analyses. We measured classification accuracy as the percent of isolated word production attempts in which the word classifier correctly predicted the identity of the target word. We also measured electrode contributions as the impact that each individual electrode had on the predictions made by the detection and classification models (Simonyan et al. 2014; Makin et al. 2020).

To investigate the clinical viability of our approach for a long-term application, we evaluated the stability of the acquired cortical signals over time using the isolated word data. By sampling neural data from four different date ranges spanning the 81-week study period, we assessed if detection and classification performance on data in the final subset could be improved by including data from the three earlier subsets during model training, which would indicate that training data accumulated across months or years of recording would reduce the need for frequent model recalibration in practical applications of our approach.

Statistical analyses

Results for each experimental condition are presented with 95% confidence intervals when appropriate. No adjustments were made for experiment-wide multiple comparisons. Word error rate, words per minute, and classification accuracy evaluation metrics were prespecified before the start of data collection. Stability analyses were designed post hoc.

1.4 Results

Sentence decoding

During real-time sentence decoding, the median decoded word error rate across 15 sentence blocks (each block contained 10 trials) was 60.5% (95% confidence interval: 51.4% to 67.6%) without language modeling and 25.6% (95% confidence interval: 17.1% to 37.1%) with language modeling (Figure 1.2A). The lowest word error rate observed for a single test block was 6.98% (with language modeling). The median word error rate was 92.1% (95% confidence interval: 85.7% to 97.2%) when measuring chance performance with sentences randomly generated by the language model. Across all 150 trials, the median decoding rate was 15.2 words per minute when including all decoded words and 12.5 words per minute when only including correctly decoded words (with language modeling; Figure 1.2B). In 92.0% of trials, the number of detected words was equal to the number of words in the target sentence (Figure 1.2C). Across all 15 sentence blocks, 5 speech events were erroneously detected before the first trial in the block and were excluded from real-time decoding and analysis (all other detected speech events were included). For almost every target sentence, the average number of word errors decreased when the language model was used (Figure 1.2D). Furthermore, over half of the sentences were decoded without error (80 out of 150 trials; with language modeling). Use of the language model during decoding improved performance by correcting grammatically and semantically implausible word sequence predictions (Figure 1.2E).

Word detection and classification

In offline analyses with 9000 isolated word production attempts, the mean classification accuracy (described in the Modeling section) was 47.1% when using the speech detector and word classifier to predict the identity of the target word from cortical activity (chance was 2% accuracy; predictions were made without using the language model; see Figure 1.12 and Figure 1.13. for additional isolated word analysis results). 98% of these word production attempts were successfully detected (191 attempts were not detected), and 968 detected events were spurious (not associated with a speech attempt). Electrodes contributing to word classification performance were primarily localized to the ventral-most aspect of the ventral sensorimotor cortex, with electrodes in the dorsal aspect of the ventral sensorimotor cortex contributing to both speech detection and word classification performance (Figure 1.3A). Classification accuracy was consistent across the majority of the word targets (Figure 1.3B; 47.1% mean and 14.5% standard deviation of the classification accuracy along the diagonal of the row-normalized confusion matrix).

Long-term signal stability

Long-term stability of the speech-related cortical activity patterns recorded during isolated word production attempts enabled consistent model performance throughout the 81-week study period without requiring daily or weekly model recalibration (Figure 1.14). When using the speech detection and word classification models to analyze cortical activity recorded

at the end of the study period, classification accuracy increased when the training dataset included data recorded over a year prior to the test dataset (Figure 1.4).

1.5 Discussion

We demonstrated that high-density recordings of cortical activity in the speech motor area of an anarthric and paralyzed person can be used to decode full words and sentences in real time. Our deep learning models were able to use the participant's neural activity to detect and classify his attempts to produce words from a 50-word vocabulary, and we could use these models together with language modeling techniques to decode a variety of meaningful sentences. Enabled by the long-term stability of the implanted device, our models could use data accumulated throughout the 81-week study period to improve decoding performance on held-out data recorded at the end of the study.

Previous demonstrations of word and sentence decoding from neural activity were conducted with participants who could speak and did not require assistive technology to communicate (Herff et al. 2015; Angrick et al. 2019; Anumanchipalli et al. 2019; David A. Moses, Leonard, et al. 2019; Makin et al. 2020). Similar to decoding intended movements from someone who cannot move, the lack of precise time alignment between intended speech and neural activity poses a challenge during model training. We managed this time-alignment problem with speech detection approaches (Kanas et al. 2014; David A. Moses, Leonard, et al. 2019; Dash et al. 2020) and classifiers that used machine learning techniques, such as

model ensembling and data augmentation, to increase tolerance to minor temporal variabilities (Sollich et al. 1996; Krizhevsky et al. 2012). Additionally, decoding performance was largely driven by neural activity patterns in ventral sensorimotor cortex, consistent with previous work implicating this area in intact speech production (Bouchard et al. 2013; Chartier et al. 2018; Emily M. Mugler et al. 2018). This finding informs electrode placement decisions for future studies and demonstrates the persistence of functional cortical speech representations after more than 15 years of anarthria, analogous to previous findings of limb-related cortical motor representations in tetraplegic individuals years after loss of limb movement (Shoham et al. 2001; Hochberg et al. 2006).

Incorporation of language modeling techniques reduced median word error rate by 35% and enabled perfect decoding in over half of the sentence trials. This improvement was facilitated by using all of the probabilistic information provided by the word classifier during decoding and by allowing the decoder to update previously predicted words each time a new word was decoded. These results demonstrate the benefit of integrating linguistic information when decoding speech from neural recordings. Speech decoding approaches generally become usable at word error rates below 30% (Watanabe et al. 2017), suggesting that our approach may be applicable in other clinical settings.

In previously reported brain-computer interface applications, decoding models often require daily recalibration prior deployment with a user (Pandarinath et al. 2017; Wolpaw et al. 2018), which can increase the variability of decoder performance across days and impede long-term adoption of the interface for real-world use (Wolpaw et al. 2018; Silversmith

et al. 2020). Due to the relatively high signal stability of electrocorticographic recordings (Chao et al. 2010; Vansteensel et al. 2016; Rao et al. 2017; Pels et al. 2019), we could accumulate cortical activity acquired by the implanted electrodes across months of recording to effectively train our decoding models. Overall, decoding performance was maintained or improved by accumulating large quantities of training data over time without daily recalibration, demonstrating the suitability of high-density electrocorticography for long-term speech neuroprosthetic applications.

These results demonstrate the early feasibility of direct word-based speech decoding from cortical signals in a paralyzed anarthric person.

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

1.6 Funding

This work was supported by a research contract under Facebook's Sponsored Academic Research Agreement, the National Institutes of Health (grant number NIH U01 NS098971-01), Joan and Sandy Weill, Bill and Susan Oberndorf, the William K. Bowes, Jr. Foundation, and the Shurl and Kay Curci Foundation.

1.7 Acknowledgments

The authors thank the study participant "Bravo-1" for his dedication and commitment; the members of Karunesh Ganguly's lab for help with the clinical trial; Mark Chevillet, Emily Mugler, Ruben Sethi, and Stephanie Thacker for support and feedback; Nick Halper and Kian Torab for hardware technical support; Mariann Ward for clinical nursing support; Matthew Leonard, Heather Dawes, and Ilona Garner for manuscript feedback; Viv Her for administrative support; Kenneth Probst for Figure 1.1 illustration; Todd Dubnicoff for video editing; and the participant's caregivers for logistical support.

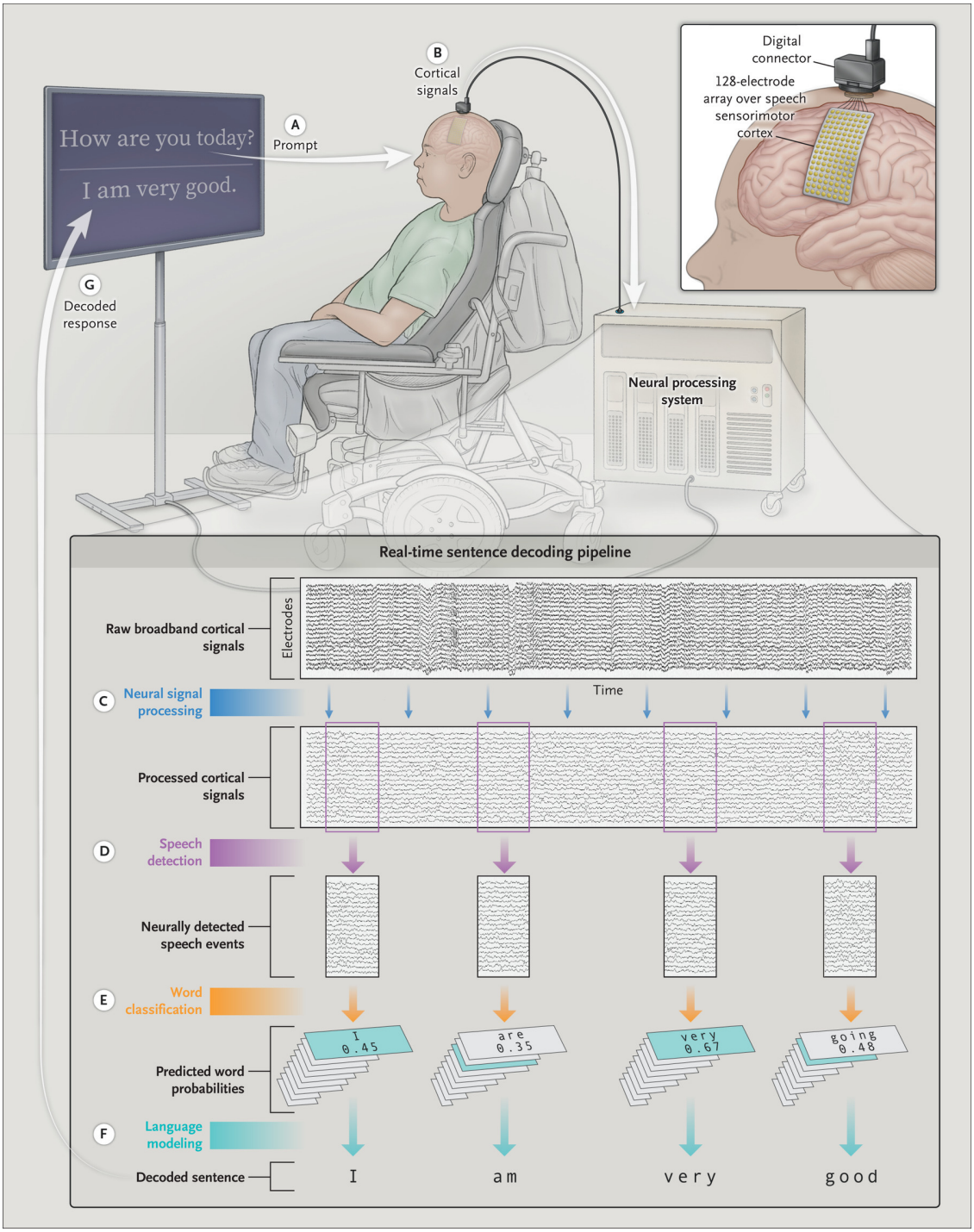
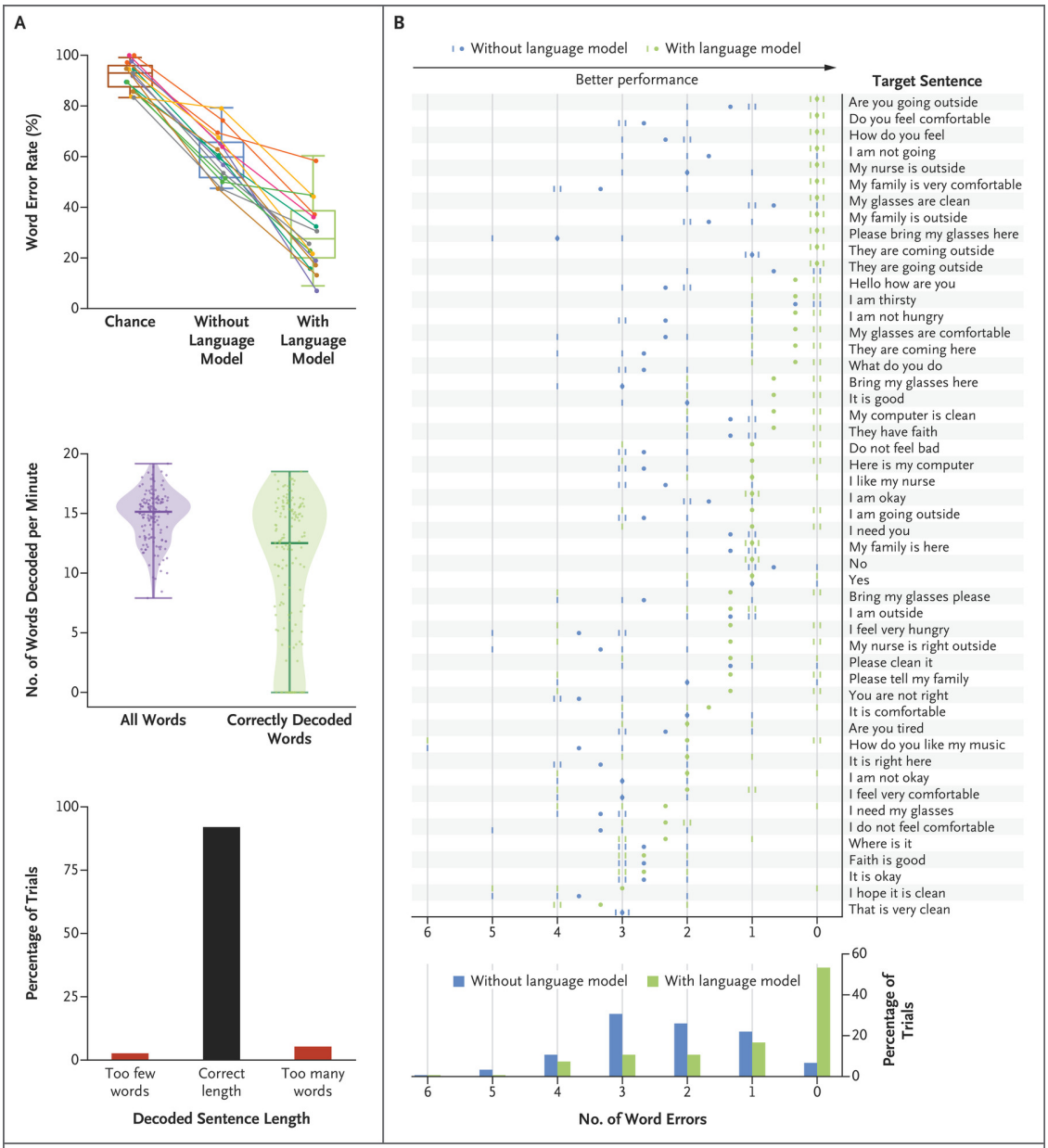


Figure 1.1. Schematic depiction of the spelling pipeline. (continued on next page).

(Previous page.) **Figure 1.1. Schematic depiction of the spelling pipeline.** Shown is how neural activity acquired from an investigational electrocorticography electrode array implanted in a clinical study participant with severe paralysis is used to directly decode words and sentences in real time. In a conversational demonstration, the participant is visually prompted with a statement or question (A) and is instructed to attempt to respond using words from a predefined vocabulary set of 50 words. Simultaneously, cortical signals are acquired from the surface of the brain through the electrode array (B) and processed in real time (C). The processed neural signals are analyzed sample by sample with the use of a speech-detection model to detect the participant's attempts to speak (D). A classifier computes word probabilities (across the 50 possible words) from each detected window of relevant neural activity (E). A Viterbi decoding algorithm uses these probabilities in conjunction with word-sequence probabilities from a separately trained natural-language model to decode the most likely sentence given the neural activity data (F). The predicted sentence, which is updated each time a word is decoded, is displayed as feedback to the participant (G). Before real-time decoding, the models were trained with data collected as the participant attempted to say individual words from the 50-word set as part of a separate task (not depicted). This conversational demonstration is a variant of the standard sentence task used in this work, in that it allows the participant to compose his own unique responses to the prompts.



Target Sentence Example	Decoded without Language Model	Decoded with Language Model
Hello how are you	Hungry how am you	Hello how are you
I like my nurse	I right my nurse	I like my nurse
They are going outside	They are going outside	They are going outside
My family is very comfortable	Glasses family is faith comfortable	My family is very comfortable
Bring my glasses please	Please my glasses please	Bring my glasses please
What do you do	What do I you	What do I do
How do you like my music	How do you like bad bring	How do you like my music

Figure 1.2. Decoding a Variety of Sentences in Real Time through Neural Signal Processing and Language Modeling. (continued on next page).

(Previous page.) **Figure 1.2. Decoding a Variety of Sentences in Real Time through Neural Signal Processing and Language Modeling.** Panel A shows the word error rates, the numbers of words decoded per minute, and the decoded sentence lengths. The top plot shows the median word error rate (defined as the number of word errors made by the decoder divided by the number of words in the target sentence, with a lower rate indicating better performance) derived from the word sequences decoded from the participant's cortical activity during the performance of the sentence task. Data points represent sentence blocks (each block comprises 10 trials); the median rate, as indicated by the horizontal line within a box, is shown across 15 sentence blocks. The upper and lower sides of the box represent the interquartile range, and the bars 1.5 times the interquartile range. Chance performance was measured by computing the word error rate on sentences randomly generated from the natural-language model. The middle plot shows the median number of words decoded per minute, as derived across all 150 trials (each data point represents a trial). The rates are shown for the analysis that included all words that were correctly or incorrectly decoded with the natural-language model and for the analysis that included only correctly decoded words. Each violin distribution was created with the use of kernel density estimation based on Scott's rule for computing the estimator bandwidth; the thick horizontal lines represent the median number of words decoded per minute, and the thinner horizontal lines the range (with the exclusion of outliers that were more than 4 standard deviations below or above the mean, which was the case for one trial). In the bottom chart, the decoded sentence lengths show whether the number of detected words was equal to the number of words in the target sentence in each of the 150 trials. Panel B shows the number of word errors in the sentences decoded with or without the natural-language model across all trials and all 50 sentence targets. Each small vertical dash represents the number of word errors in a single trial (there are 3 trials per target sentence; marks for identical error counts are staggered horizontally for visualization purposes). Each dot represents the mean number of errors for that target sentence across the 3 trials. The histogram at the bottom shows the error counts across all 150 trials. Panel C shows seven target sentence examples along with the corresponding sentences decoded with and without the natural-language model. Correctly decoded words are shown in black and incorrect words in red.

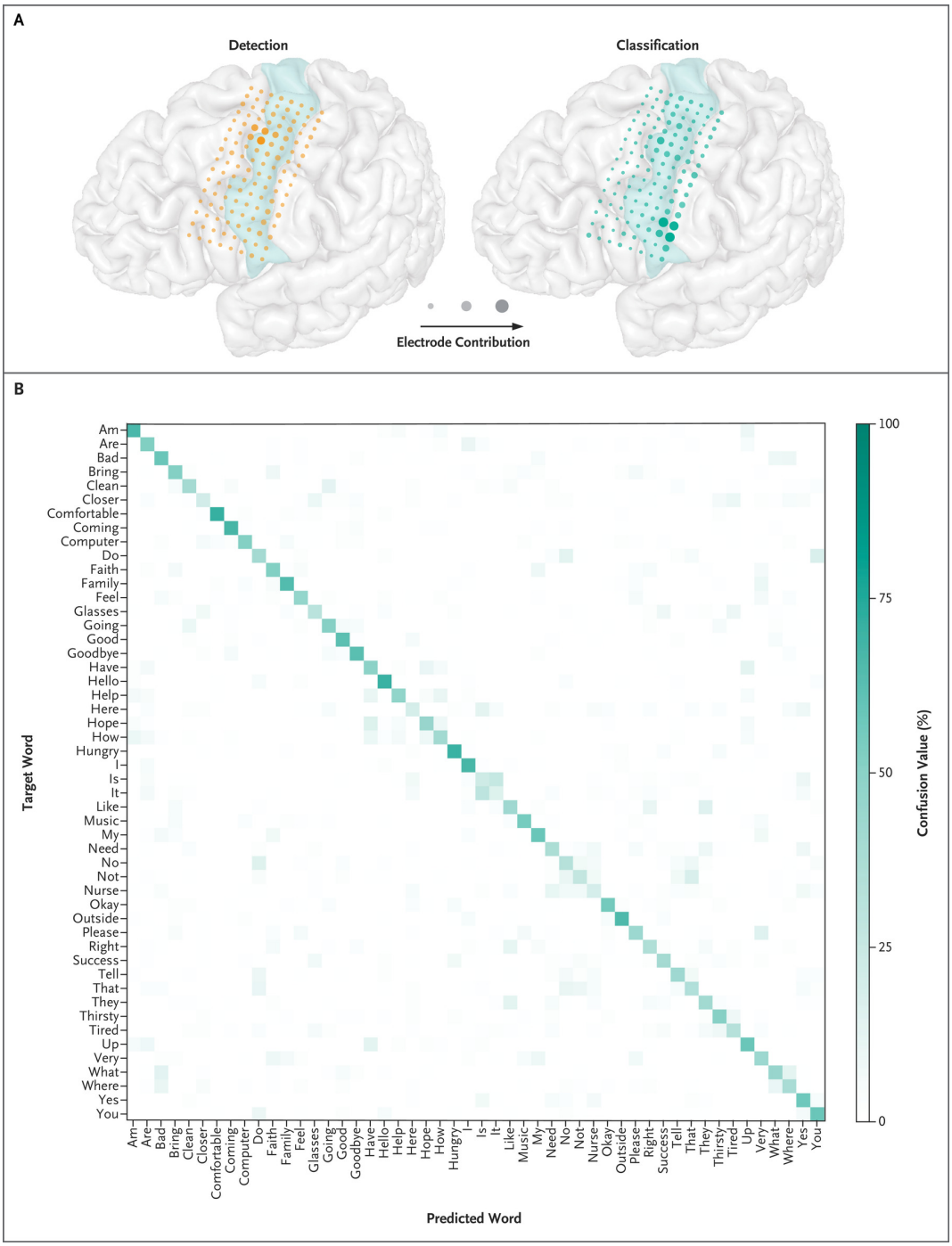


Figure 1.3. Distinct Neural Activity Patterns during Word-Production Attempts. (continued on next page).

(Previous page.) **Figure 1.3. Distinct Neural Activity Patterns during Word-Production Attempts.** Panel A shows the participant's brain reconstruction overlaid with the locations of the implanted electrodes and their contributions to the speech-detection and word-classification models. Plotted electrode size (area) and opacity are scaled by relative contribution (important electrodes appear larger and more opaque than other electrodes). Each set of contributions is normalized to sum to 1. For anatomical reference, the precentral gyrus is highlighted in light green. Panel B shows word confusion values computed with the use of the isolated-word data. For each target word (each row), the confusion value measures how often the word classifier predicted (regardless of whether the prediction was correct) each of the 50 possible words (each column) while the participant was attempting to say that target word. The confusion value is computed as a percentage relative to the total number of isolated-word trials for each target word, with the values in each row summing to 100%. Values along the diagonal correspond to correct classifications, and off-diagonal values correspond to incorrect classifications. The natural-language model was not used in this analysis.

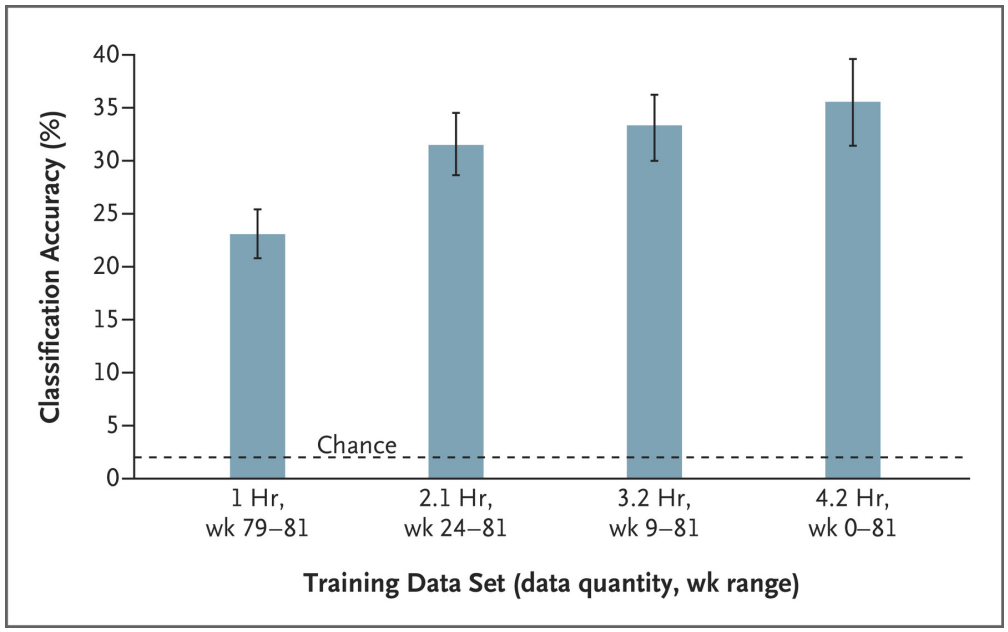


Figure 1.4. Signal Stability and Long-Term Accumulation of Training Data to Improve Decoder Performance. Each bar depicts the mean classification accuracy (the percentage of trials in which the target word was correctly predicted) from isolated-word data sampled from the final weeks of the study period (weeks 79 through 81) after speech-detection and word-classification models were trained on different samples of the isolated-word data from various week ranges. Each result was computed with the use of a 10-fold cross-validation evaluation approach. In this approach, the available data were partitioned into 10 equally sized, nonoverlapping subsets. In the first cross-validation "fold," one of these data subsets is used as the testing set, and the remaining 9 are used for model training. This was repeated 9 more times until each subset was used for testing (after training on the other subsets). This approach ensures that models were never evaluated on the data used during training (Sections S6 and S14). Error bars indicate the 95% confidence interval of the mean, each computed across the 10 cross-validation folds. The data quantities specify the average amount of data used to train the word-classification models across cross-validation folds. Week 0 denotes the first week during which data for this study was collected, which occurred 9 weeks after surgical implantation of the study device. Accuracy of chance performance was calculated as 1 divided by the number of possible words and is indicated by a horizontal dashed line.

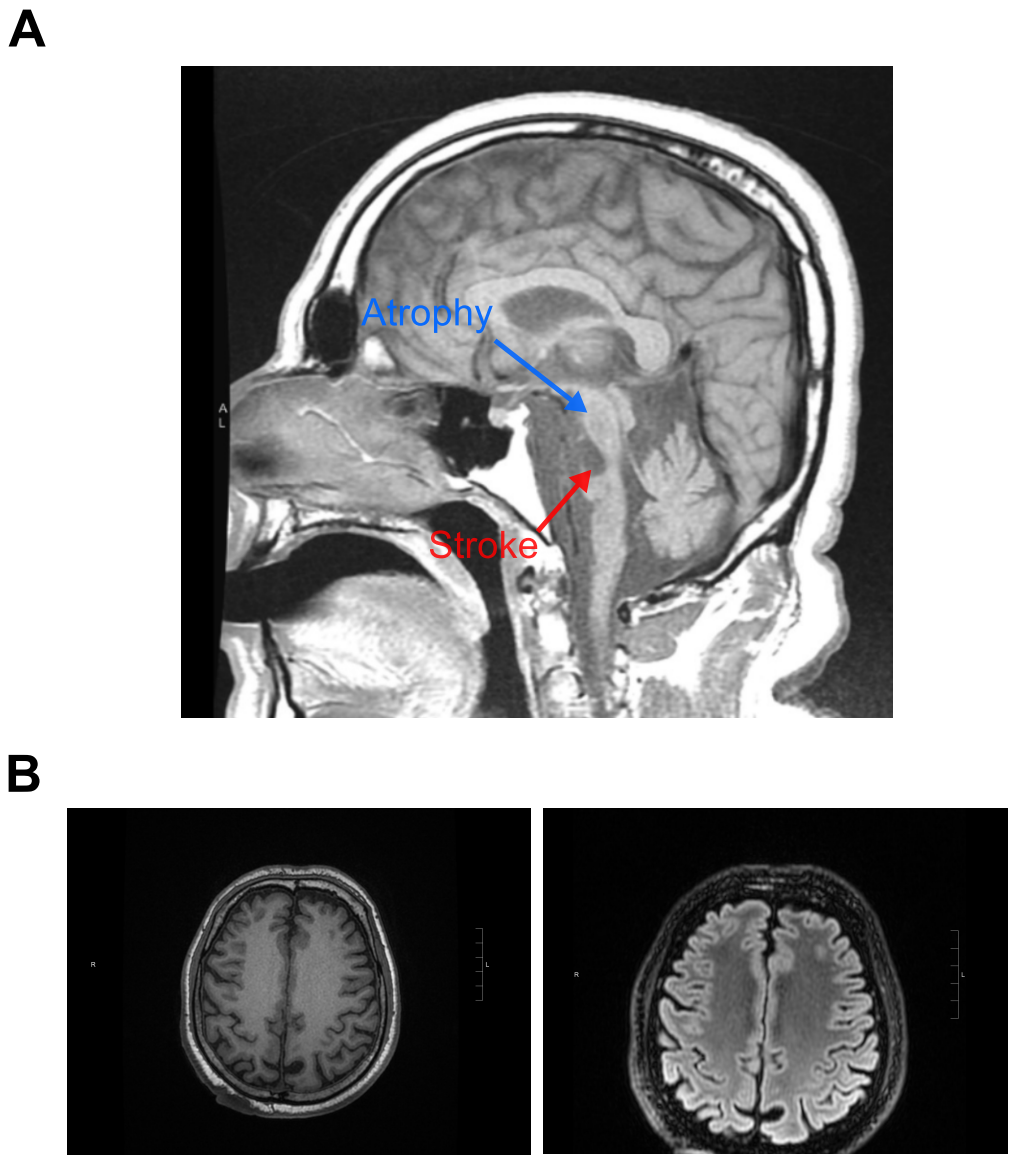


Figure 1.5. MRI results for the participant Panel A shows a sagittal MRI for the participant, who has encephalomalacia and brain-stem atrophy (labeled in blue) caused by pontine stroke (labeled in red). Panel B shows two additional MRI scans that indicate the absence of cerebral atrophy, suggesting that cortical neuron populations (including those recorded from in this study) should be relatively unaffected by the participant's pathology.

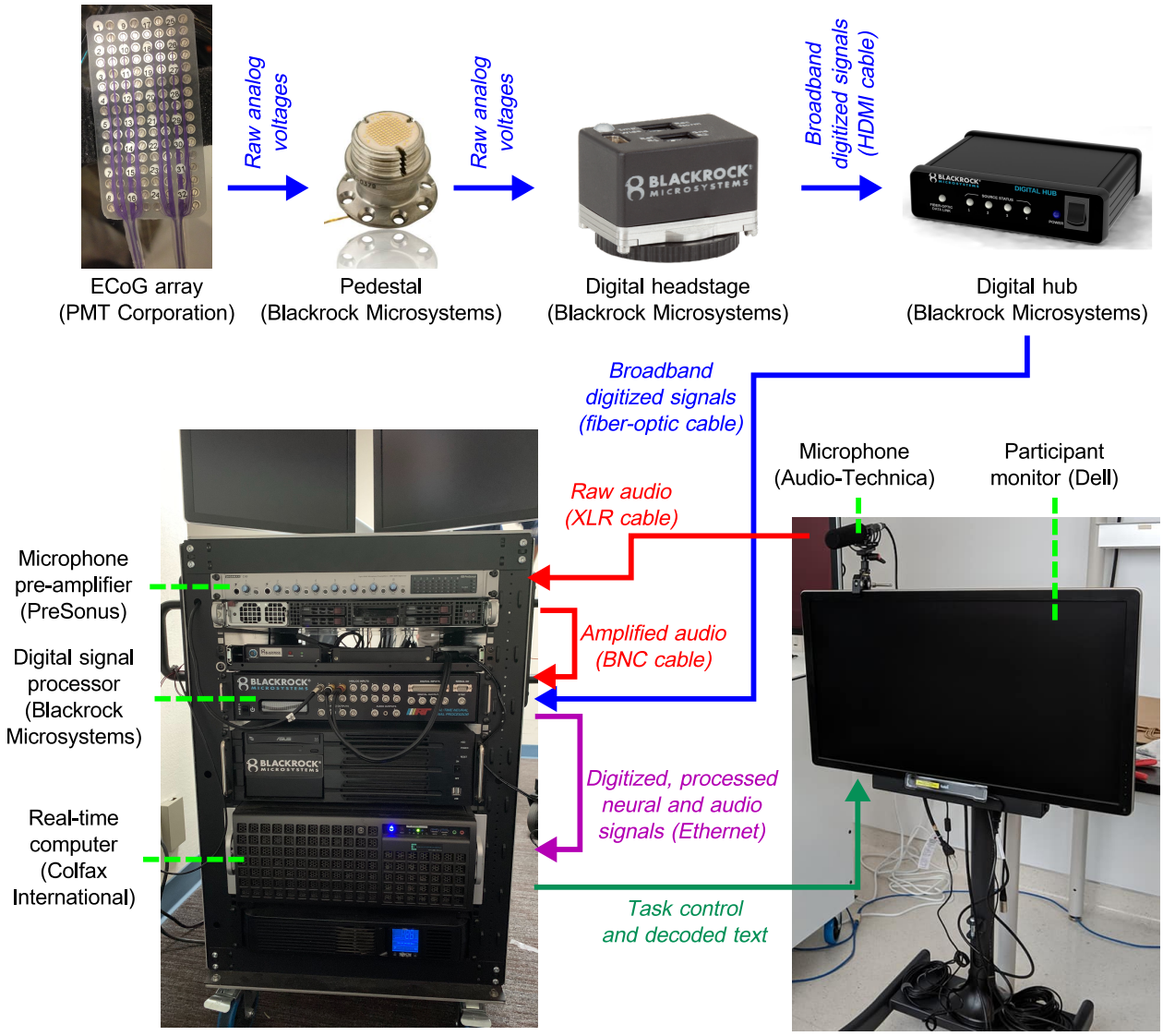


Figure 1.6. Real-time neural data acquisition hardware infrastructure Electrocorticography (ECoG) data acquired from the implanted array and percutaneous pedestal connector are processed and transmitted to the Neuroport digital signal processor (DSP). Simultaneously, microphone data are acquired, amplified, and transmitted to the DSP. Signals from the DSP are transmitted to the real-time computer. The real-time computer controls the task displayed to the participant, including any decoded sentences that are provided in real time as feedback. Speaker data (output from the real-time computer) are also sent to the DSP and synchronized with the neural signals (not depicted). During earlier sessions, a human patient cable connected to the pedestal acquired the ECoG signals, which were then processed by a front-end amplifier before being transmitted to the DSP (the human patient cable and front-end amplifier, manufactured by Blackrock Microsystems, are not depicted here, but they replaced the digital headstage and digital hub in this pipeline when they were used).

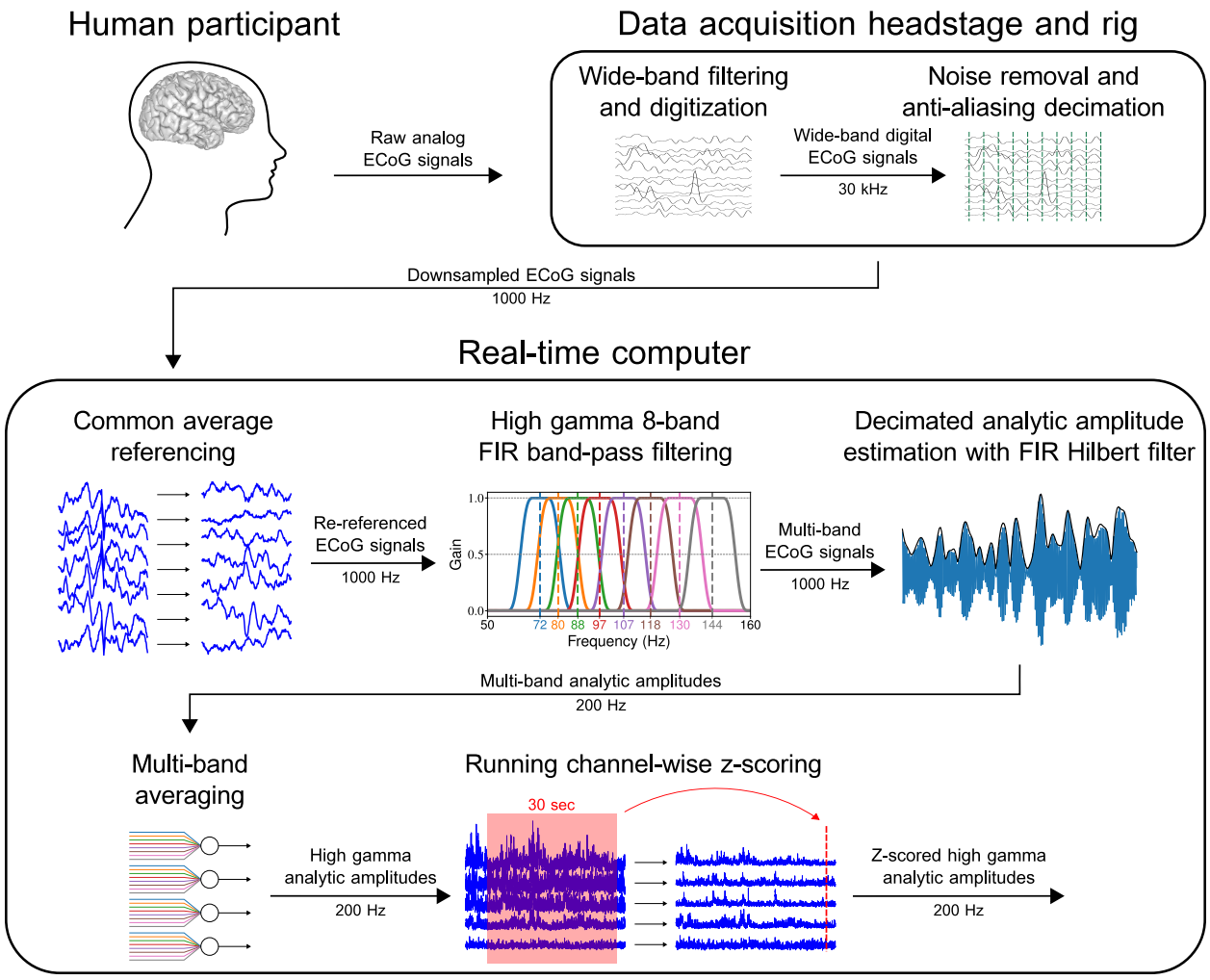


Figure 1.7. Real-time neural signal processing pipeline Using the data acquisition headstage and rig, the participant's electrocorticography (ECoG) signals were acquired at 30 kHz, filtered with a wide-band filter, conditioned with a software-based line noise cancellation technique, low-pass filtered at 500 Hz, and streamed to the real-time computer at 1 kHz. On the real-time computer, custom software was used to perform common average referencing, multi-band high gamma band-pass filtering, analytic amplitude estimation, multi-band averaging, and running z-scoring on the ECoG signals. The resulting signals were then used as the measure of high gamma activity for the remaining analyses. This figure was adapted from our previous work (David A. Moses, Leonard, et al. 2019), which implemented a similar neural signal preprocessing pipeline.

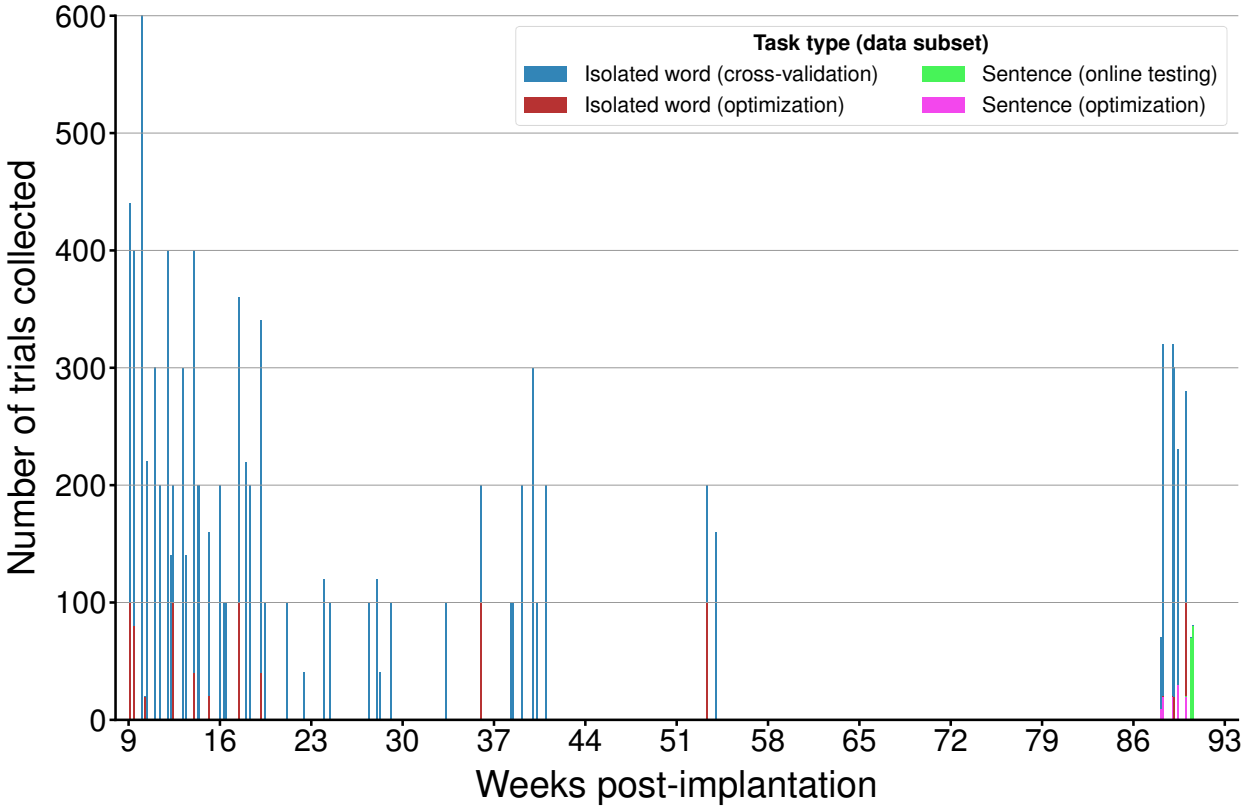


Figure 1.8. Data collection timeline Bars are stacked vertically if more than one data type was collected in a day (the height of the stacked bars for any given day is equal to the total number of trials collected that day). The irregularity of the data collection schedule was influenced by external and clinical time constraints unrelated to the implanted device. The gap from 55–88 weeks was due to clinical guidelines concerning the COVID-19 pandemic.

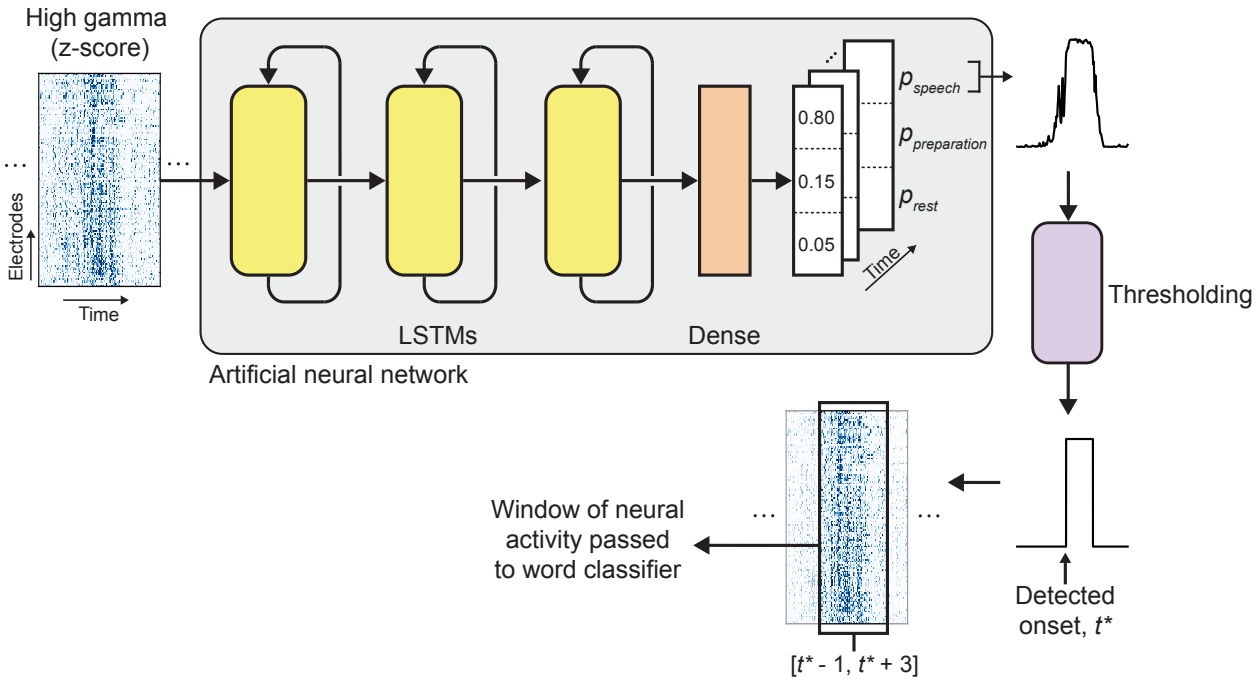


Figure 1.9. Speech detection model schematic The z-scored high gamma activity across all electrodes is processed time point by time point by an artificial neural network consisting of a stack of three long short-term memory layers (LSTMs) and a single dense (fully connected) layer. The dense layer projects the latent dimensions of the last LSTM layer into probability space for three event classes: speech, preparation, and rest. The predicted speech event probability time series is smoothed and then thresholded with probability and time thresholds to yield onset (t^*) and offset times of detected speech events. During sentence decoding, each time a speech event was detected, the window of neural activity spanning from -1 to $+3$ seconds relative to the detected onset (t^*) was passed to the word classifier. The neural activity, predicted speech probability time series (upper right), and detected speech event (lower right) shown are the actual neural data and detection results across a 7-second time window for an isolated word trial in which the participant attempted to produce the word "family".

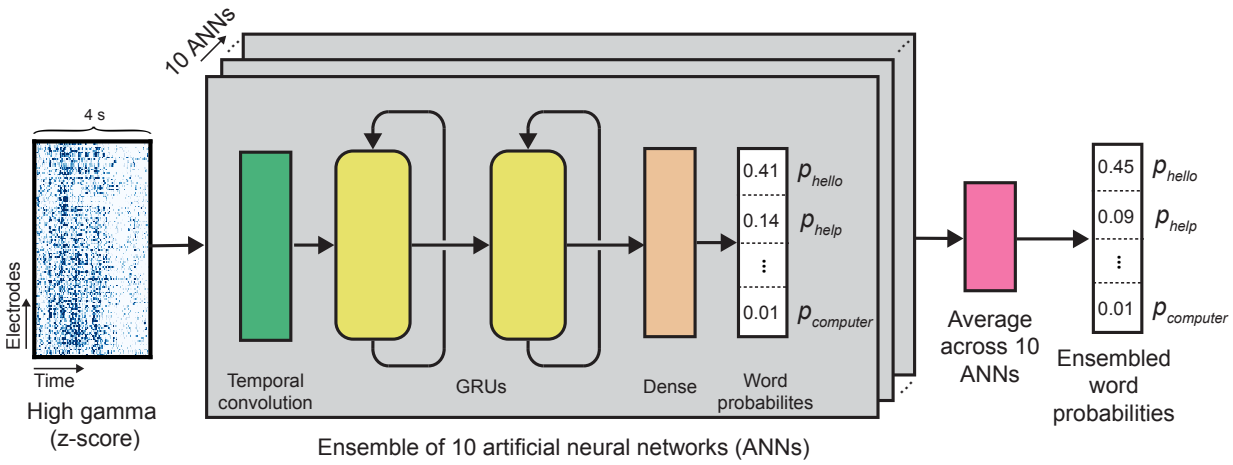


Figure 1.10. Word classification model schematic For each classification, a 4-second time window of high gamma activity is processed by an ensemble of 10 artificial neural network (ANN) models. Within each ANN, the high gamma activity is processed by a temporal convolution followed by two bidirectional gated recurrent unit (GRU) layers. A dense layer projects the latent dimension from the final GRU layer into probability space, which contains the probability of each of the words from the 50-word set being the target word during the speech production attempt associated with the neural time window. The 10 probability distributions from the ensembled ANN models are averaged together to obtain the final vector of predicted word probabilities.

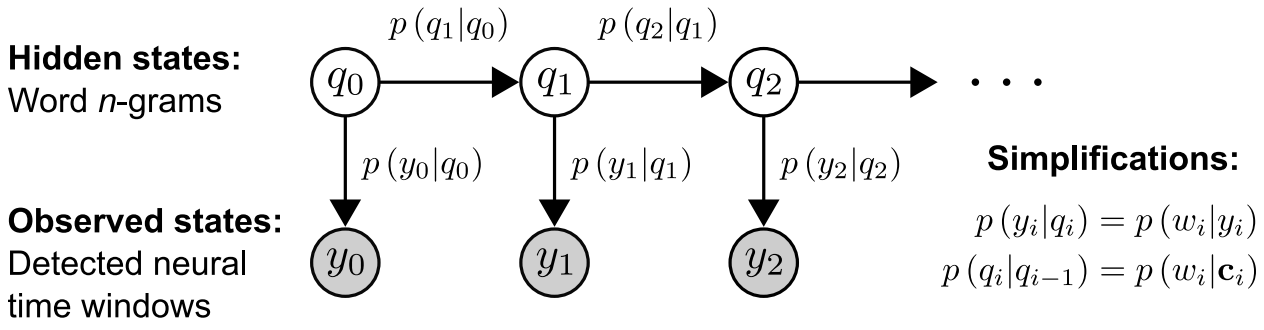


Figure 1.11. Sentence decoding hidden Markov model This hidden Markov model (HMM) describes the relationship between the words that the participant attempts to produce (the hidden states q_i) and the associated detected time windows of neural activity (the observed states y_i). The HMM emission probabilities $py_0|q_0$ can be simplified to $pw_i|y_i$ (the word likelihoods provided by the word classifier), and the HMM transition probabilities $pq_i|q_{i-1}$ can be simplified to $pw_i|i$ (the word-sequence prior probabilities provided by the language model).

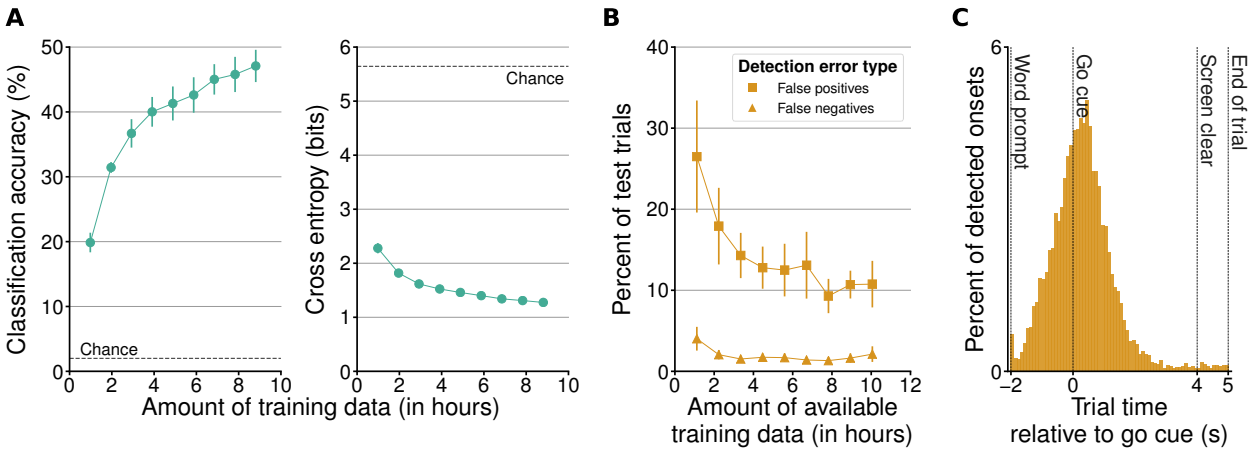


Figure 1.12. Auxiliary modeling results with isolated word data Panel A shows the effect of the amount of training data on word classification accuracy (left) and cross-entropy loss (right) using cortical activity recorded during the participant's isolated word production attempts. Lower cross entropy indicates better performance. Each point depicts mean ± standard deviation across 10 cross-validation folds (the error bars in the cross-entropy plot were typically too small to be seen alongside the circular markers). Chance performance is depicted as a horizontal dashed line in each plot (chance cross-entropy loss is computed as the negative log (base 2) of the reciprocal of the number of word targets). Performance improved more rapidly for the first four hours of training data and then less rapidly for the next 5 hours, although it did not plateau. When using all available isolated word data, the information transfer rate was 25.1 bits per minute (not depicted), and the target word appeared in the top 5 predictions from the word classifier in 81.7% of trials (standard deviation was 2.1% across cross-validation folds; not depicted). Panel B shows the effect of the amount of training data on the frequency of detection errors during speech detection and detected event curation with the isolated word data. Lower error rates indicate better performance. False positives are detected events that were not associated with a word production attempt and false negatives are word production attempts that were not associated with a detected event. Each point depicts mean ± standard deviation across 10 cross-validation folds. Not all of the available training data were used to fit each speech detection model, but each model always used between 47 and 83 minutes of data (not depicted). Panel C shows the distribution of onsets detected from neural activity across 9000 isolated word trials relative to the go cue (100 ms histogram bin size). This histogram was created using results from the final set of analyses in the learning curve scheme (in which all available trials were included in the cross-validated evaluation). The distribution of detected speech onsets had a mean of 308 ms after the associated go cues and a standard deviation of 1017 ms. This distribution was likely influenced to some degree by behavioral variability in the participant's response times. During detected event curation, 429 trials required curation to choose a detected event from multiple candidates (420 trials had 2 candidates and 9 trials had 3 candidates).

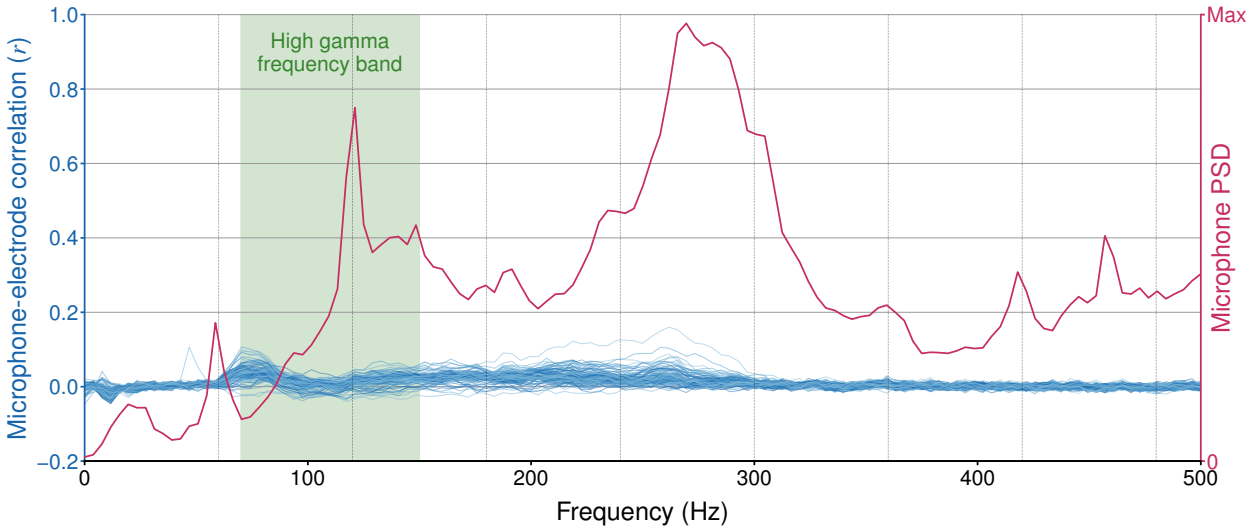


Figure 1.13. Acoustic contamination investigation Each blue curve depicts the average correlations between the spectrograms from a single electrode and the corresponding spectrograms from the time-aligned microphone signal as a function of frequency. The red curve depicts the average power spectral density (PSD) of the microphone signal. Vertical dashed lines mark the 60 Hz line noise frequency and its harmonics. Highlighted in green is the high gamma frequency band (70–150 Hz), which was the frequency band from which we extracted the neural features used during decoding. Across all frequencies, correlations between the electrode and microphone signals are small. There is a slight increase in correlation in the lower end of the high gamma frequency range, but this increase in correlation occurs as the microphone PSD decreases. Because the correlations are low and do not increase or decrease with the microphone PSD, the observed correlations are likely due to factors other than acoustic contamination, such as shared electrical noise. After comparing these results to those observed in the study describing acoustic contamination (which informed the contamination analysis we used here) (Roussel et al. 2020), we conclude that our decoding performance was not artificially improved by acoustic contamination of our electrophysiological recordings.

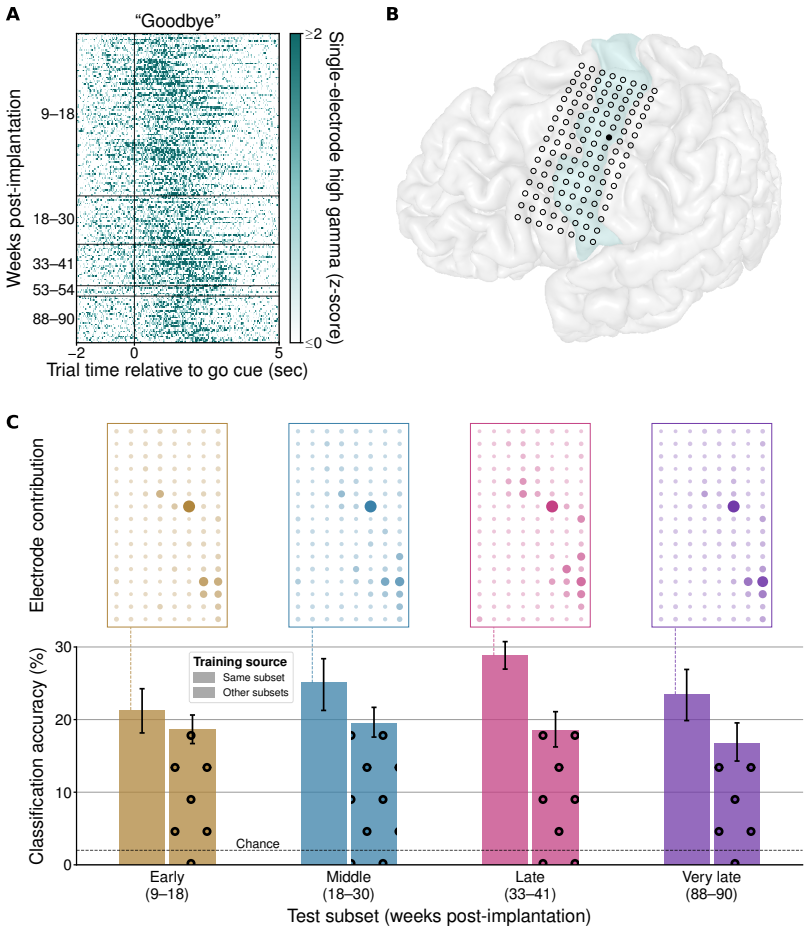


Figure 1.14. Long-term stability of speech-evoked signals Panel A shows neural activity from a single electrode across all of the participant's attempts to say the word "Goodbye" during the isolated word task, spanning 81 weeks of recording. Panel B shows the participant's brain reconstruction overlaid with electrode locations. The electrode shown in Panel A is filled in with black. For anatomical reference, the precentral gyrus is highlighted in light green. Panel C shows word classification outcomes from training and testing the detector and classifier on subsets of isolated word data sampled from four non-overlapping date ranges. Each subset contains data from 20 attempted productions of each word. Each solid bar depicts results from cross-validated evaluation within a single subset, and each dotted bar depicts results from training on data from all of the subsets except for the one that is being evaluated. Each error bar shows the 95% confidence interval of the mean, computed across cross-validation folds. Chance accuracy is depicted as a horizontal dashed line. Electrode contributions computed during cross-validated evaluation within a single subset are shown on top (oriented with the most dorsal and posterior electrode in the upper-right corner). Plotted electrode size (area) and opacity are scaled by relative contribution. Each set of contributions is normalized to sum to 1. These results suggest that speech-evoked cortical responses remained relatively stable throughout the study period, although model recalibration every 2–3 months may still be beneficial for decoding performance.

Table 1.1. Hyperparameter definitions and values.

Model	Hyperparameter description	Search space type	Value range	Optimal values[1]
Speech detector	Smoothing size	Uniform (integer)	$[1, 80]$	$(8, 5, 22)$
	Probability threshold	Uniform	$[0.1, 0.9]$	$(0.297, 0.319, 0.592)$
	Time threshold duration	Uniform (integer)	$[25, 150]$	$(79, 82, 93)$
Word classifier	Number of GRU layers	Uniform (integer)	$[1, 3]$	$(2, 2)$
	Nodes per GRU layer	Uniform (integer)	$[64, 512]$	$(434, 420)$
	Dropout fraction	Uniform	$[0.5, 0.95]$	$(0.704, 0.646)$
	Convolution kernel size and skip	Uniform (integer)	$[1, 2]$	$(2, 2)$
Language model	Initial word smoothing (ψ)	Logarithmically uniform	$[0.001, 1000]$	0.576
Viterbi decoder	Language model scaling factor (L)	Logarithmically uniform	$[0.1, 10]$	0.913

[1] For the speech detection hyperparameters, three values are listed: the first is the optimal value found when optimizing the detector on the isolated word optimization subset (used to detect word production attempts in the cross-validation subsets for evaluation by the word classifier), the second is the optimal value found when optimizing the detector on a subset of the pooled cross-validation subsets (used to detect word production attempts in the isolated word optimization subset for use during hyperparameter optimization of the word classifier), and the third is the optimal value found during hyperparameter optimization of the decoding pipeline with the sentence optimization subset (the value used during online sentence decoding). For the word classification hyperparameters, two values are listed: the first is the optimal value found when optimizing the classifier on the isolated word optimization subset (the value used for all isolated word evaluations) and the second is the optimal value found when optimizing the classifier on a small subset of isolated word trials near the end of the study period (the value used for offline sentence optimization and online sentence decoding). For the language modeling and Viterbi decoding hyperparameters, the optimal value listed was found when optimizing the decoding pipeline with the sentence optimization subset (the value used for online sentence decoding).

References

Angrick, Miguel, Christian Herff, Emily Mugler, et al. (June 1, 2019). "Speech synthesis from ECoG using densely connected 3D convolutional neural networks". *Journal of Neural Engineering* 16.3, p. 036019. ISSN: 1741-2560, 1741-2552. DOI: 10.1088/1741-2552/ab0c59.

Anumanchipalli, Gopala K., Josh Chartier, and Edward F. Chang (Apr. 2019). "Speech synthesis from neural decoding of spoken sentences". *Nature* 568.7753, pp. 493–498. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-019-1119-1.

Beukelman, David R., Susan Fager, Laura Ball, and Aimee Dietz (Jan. 2007). "AAC for adults with acquired neurological conditions: A review". *Augmentative and Alternative Communication* 23.3, pp. 230–242. ISSN: 0743-4618, 1477-3848. DOI: 10.1080/07434610701553668.

Bouchard, Kristofer E., Nima Mesgarani, Keith Johnson, and Edward F. Chang (Mar. 2013). "Functional organization of human sensorimotor cortex for speech articulation". *Nature* 495.7441, pp. 327–332. ISSN: 1476-4687 (Electronic)\r0028-0836 (Linking). DOI: 10.1038/nature11911.

Brumberg, Jonathan S., Kevin M. Pitt, Alana Mantie-Kozlowski, and Jeremy D. Burnison (Feb. 6, 2018). "Brain–Computer Interfaces for Augmentative and Alternative Communication: A Tutorial". *American Journal of Speech-Language Pathology* 27.1, pp. 1–12. ISSN: 1058-0360, 1558-9110. DOI: 10.1044/2017_AJSLP-16-0244.

Brumberg, Jonathan S., E. Joe Wright, Dinal S. Andreasen, et al. (May 12, 2011). "Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech-motor cortex". *Frontiers in Neuroscience* 5, p. 65. ISSN: 1662453X. DOI: 10.3389/fnins.2011.00065.

Chao, Zenas C., Yasuo Nagasaka, and Naotaka Fujii (2010). "Long-term asynchronous decoding of arm motion using electrocorticographic signals in monkey". *Frontiers in Neuroengineering* 3, p. 3. ISSN: 16626443. DOI: 10.3389/fneng.2010.00003.

Chartier, Josh, Gopala K. Anumanchipalli, Keith Johnson, and Edward F. Chang (2018). "Encoding of Articulatory Kinematic Trajectories in Human Speech Sensorimotor Cortex". *Neuron* 98.5, 1042–1054.e4. DOI: 10.1016/j.neuron.2018.04.031.

Chen, Stanley F. and Joshua Goodman (Oct. 1999). "An empirical study of smoothing techniques for language modeling". *Computer Speech & Language* 13.4, pp. 359–393. ISSN: 08852308. DOI: 10.1006/csla.1999.0128.

Dash, Debadatta, Paul Ferrari, and Jun Wang (2020). "Decoding Imagined and Spoken Phrases From Non-invasive Neural (MEG) Signals". *Frontiers in Neuroscience* 14. ISSN: 1662-453X.

Felgoise, Stephanie H., Vincenzo Zaccheo, Jason Duff, and Zachary Simmons (May 18, 2016). "Verbal communication impacts quality of life in patients with amyotrophic lateral sclerosis". *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* 17.3, pp. 179–183. ISSN: 2167-8421, 2167-9223. DOI: 10.3109/21678421.2015.1125499.

Guenther, Frank H., Jonathan S. Brumberg, E. Joseph Wright, et al. (Dec. 9, 2009). "A Wireless Brain-Machine Interface for Real-Time Speech Synthesis". *PLoS ONE* 4.12. Ed. by Eshel Ben-Jacob, e8218. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0008218.

Guenther, Frank H. and Gregory Hickok (2016). "Neural Models of Motor Speech Control". *Neurobiology of Language*. Elsevier, pp. 725–740. ISBN: 978-0-12-407794-2.

Herff, Christian, Dominic Heger, Adriana de Pesters, et al. (2015). "Brain-to-text: decoding spoken phrases from phone representations in the brain". *Frontiers in Neuroscience* 9 (June), pp. 1–11. DOI: 10.3389/fnins.2015.00217.

Hochberg, Leigh R., Mijail D. Serruya, Gerhard M. Friehs, et al. (July 2006). "Neuronal ensemble control of prosthetic devices by a human with tetraplegia". *Nature* 442.7099, pp. 164–171. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature04970.

Kanas, Vasileios G., Iosif Mporas, Heather L. Benz, et al. (2014). "Real-time voice activity detection for ECoG-based speech brain machine interfaces". *19th International Conference on Digital Signal Processing*. Vol. 2014, pp. 862–865. DOI: 10.1109/ICDSP.2014.6900790.

Kneser, R. and H. Ney (1995). "Improved backing-off for M-gram language modeling". *1995 International Conference on Acoustics, Speech, and Signal Processing*. 1995 International Conference on Acoustics, Speech, and Signal Processing. Vol. 1. Detroit, MI, USA: IEEE, pp. 181–184. ISBN: 978-0-7803-2431-2. DOI: 10.1109/ICASSP.1995.479394.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". *Advances in Neural Information Processing*

Systems 25. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Curran Associates, Inc., pp. 1097–1105.

Linse, Katharina, Elisa Aust, Markus Joos, et al. (2018). "Communication Matters — Pitfalls and Promise of Hightech Communication Devices in Palliative Care of Severely Physically Disabled Patients With Amyotrophic Lateral Sclerosis". 9 (July), pp. 1–18. DOI: 10.3389/fneur.2018.00603.

Lotte, Fabien, Jonathan S. Brumberg, Peter Brunner, et al. (2015). "Electrocorticographic representations of segmental features in continuous speech". *Frontiers in Human Neuroscience* 09 (February), pp. 1–13. ISSN: 1662-5161 (Electronic)\r1662-5161 (Linking). DOI: 10.3389/fnhum.2015.00097.

Makin, Joseph G., David A. Moses, and Edward F. Chang (Apr. 2020). "Machine translation of cortical activity to text with an encoder–decoder framework". *Nature Neuroscience* 23.4, pp. 575–582. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/s41593-020-0608-8.

Martin, Stephanie, Iñaki Iturrate, José del R. Millán, et al. (June 21, 2018). "Decoding Inner Speech Using Electrocorticography: Progress and Challenges Toward a Speech Prosthesis". *Frontiers in Neuroscience* 12, p. 422. ISSN: 1662-453X. DOI: 10.3389/fnins.2018.00422.

Moses, David A, Matthew K Leonard, and Edward F Chang (June 1, 2018). "Real-time classification of auditory sentences using evoked cortical activity in humans". *Journal of Neural Engineering* 15.3, p. 036005. ISSN: 1741-2560, 1741-2552. DOI: 10.1088/1741-2552/aaab6f.

Moses, David A., Matthew K. Leonard, Joseph G. Makin, and Edward F. Chang (Dec. 2019). "Real-time decoding of question-and-answer speech dialogue using human cortical activity". *Nature Communications* 10.1, p. 3096. ISSN: 2041-1723. DOI: 10.1038/s41467-019-10994-4.

Moses, David A., Sean L. Metzger, Jessie R. Liu, et al. (July 15, 2021). "Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria". *New England Journal of Medicine* 385.3, pp. 217–227. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMoa2027540.

Mugler, Emily M, James L Patton, Robert D Flint, et al. (2014). "Direct classification of all American English phonemes using signals from functional speech motor cortex." *Journal of neural engineering* 11.3, pp. 035015–035015. ISSN: 1741-2560. DOI: 10.1088/1741-2560/11/3/035015.

Mugler, Emily M., Matthew C. Tate, Karen Livescu, et al. (2018). "Differential Representation of Articulatory Gestures and Phonemes in Precentral and Inferior Frontal Gyri". *The Journal of Neuroscience* 4653, pp. 1206–18. DOI: 10.1523/JNEUROSCI.1206-18.2018.

Nip, Ignatius and Carole R. Roth (2017). "Anarthria". *Encyclopedia of Clinical Neuropsychology*. Ed. by Jeffrey Kreutzer, John DeLuca, and Bruce Caplan. Cham: Springer International Publishing, pp. 1–1. ISBN: 978-3-319-56782-2. DOI: 10.1007/978-3-319-56782-2_855-4.

Pandarinath, Chethan, Paul Nuyujukian, Christine H. Blabe, et al. (2017). "High performance communication by people with paralysis using an intracortical brain-computer

interface". *eLife* 6, pp. 1–27. ISSN: 2050-084X (Electronic) 2050-084X (Linking). DOI: 10.7554/eLife.18554.

Pels, Elmar G.M., Erik J. Aarnoutse, Sacha Leinders, et al. (Oct. 2019). "Stability of a chronic implanted brain-computer interface in late-stage amyotrophic lateral sclerosis". *Clinical Neurophysiology* 130.10, pp. 1798–1803. ISSN: 13882457. DOI: 10.1016/j.clinph.2019.07.020.

Rao, Vikram R., Matthew K. Leonard, Jonathan K. Kleen, et al. (June 2017). "Chronic ambulatory electrocorticography from human speech cortex". *NeuroImage* 153, pp. 273–282. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2017.04.008.

Roussel, Philémon, Gaël Le Godais, Florent Bocquelet, et al. (Oct. 15, 2020). "Observation and assessment of acoustic contamination of electrophysiological brain signals during speech production and sound perception". *Journal of Neural Engineering* 17.5, p. 056028. ISSN: 1741-2552. DOI: 10.1088/1741-2552/abb25e.

Salari, E., Z. V. Freudenburg, M. P. Branco, et al. (Dec. 2019). "Classification of Articulator Movements and Movement Direction from Sensorimotor Cortex Activity". *Scientific Reports* 9.1, p. 14165. ISSN: 2045-2322. DOI: 10.1038/s41598-019-50834-5.

Sellers, Eric W, David B Ryan, and Christopher K Hauser (Oct. 2014). "Noninvasive brain-computer interface enables communication after brainstem stroke". *Science translational medicine* 6.257, 257re7–257re7. DOI: 10.1126/scitranslmed.3007801.

Shoham, Shy, Eric Halgren, Edwin M. Maynard, and Richard A. Normann (Oct. 2001). "Motor-cortical activity in tetraplegics". *Nature* 413.6858, pp. 793–793. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/35101651.

Silversmith, Daniel B., Reza Abiri, Nicholas F. Hardy, et al. (Sept. 7, 2020). "Plug-and-play control of a brain–computer interface through neural map stabilization". *Nature Biotechnology* 39.3, pp. 326–335. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-020-0662-5.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman (2014). "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps". *Workshop at the International Conference on Learning Representations*. 2014 International Conference on Learning Representations. Ed. by Yoshua Bengio and Yann LeCun. Banff, Canada.

Sollich, Peter and Anders Krogh (1996). "Learning with ensembles: How overfitting can be useful". *Advances in Neural Information Processing Systems 8*. Ed. by D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo. MIT Press, pp. 190–196.

Vansteensel, Mariska J., Elmar G.M. Pels, Martin G. Bleichner, et al. (2016). "Fully Implanted Brain–Computer Interface in a Locked-In Patient with ALS". *New England Journal of Medicine* 375.21, pp. 2060–2066. ISSN: 0028-4793\r1533-4406. DOI: 10.1056/NEJMoa1608085.

Viterbi, Andrew J. (1967). "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm". *IEEE Transactions on Information Theory* 13.2, pp. 260–269. ISSN: 0018-9448. DOI: 10.1109/TIT.1967.1054010.

Watanabe, Shinji, Marc Delcroix, Florian Metze, and John R Hershey (2017). *New era for robust speech recognition: exploiting deep learning*. Berlin, Germany: Springer-Verlag. ISBN: 978-3-319-64680-0.

Wolpaw, Jonathan R., Richard S. Bedlack, Domenic J. Reda, et al. (July 17, 2018). "Independent home use of a brain-computer interface by people with amyotrophic lateral sclerosis". *Neurology* 91.3, e258–e267. ISSN: 0028-3878, 1526-632X. DOI: 10.1212/WNL.0000000000005812.

Chapter 2

Generalizable spelling using a speech neuroprosthesis in an individual with severe limb and vocal paralysis

Disclaimer: This chapter is a direct adaptation of the following article. Supplementary material is not included in this adaptation, but is available online.

Sean L. Metzger*, **Jessie R. Liu***, David A. Moses*, et al. (2022). Generalizable spelling using a speech neuroprosthesis in an individual with severe limb and vocal paralysis. *Nature Communications*, 13(6510). doi: 10.1038/s41467-022-33611-3.

* Denotes equal contribution.

Personal contributions: I designed and trained the real-time speech detection model, performed nearest-class distance and evoked signal analyses, performed statistical assessments, and contributed to the neural-feature analyses. With Sean L. Metzger, I generated figures, with David A. Moses we collected data and designed the spelling process. With Edward F. Chang we wrote the original draft of the manuscript with input from all authors.

2.1 Abstract

Neuroprostheses have the potential to restore communication to people who cannot speak or type due to paralysis. However, it is unclear if silent attempts to speak can be used to control a communication neuroprosthesis. Here, we translated direct cortical signals in a clinical-trial participant (ClinicalTrials.gov; NCT03698149) with severe limb and vocal-tract paralysis into single letters to spell out full sentences in real time. We used deep-learning and language-modeling techniques to decode letter sequences as the participant attempted to silently spell using code words that represented the 26 English letters (e.g. "alpha" for "a"). We leveraged broad electrode coverage beyond speech-motor cortex to include supplemental control signals from hand cortex and complementary information from low- and high-frequency signal components to improve decoding accuracy. We decoded sentences using words from a 1,152-word vocabulary at a median character error rate of 6.13% and speed of 29.4 characters per minute. In offline simulations, we showed that our approach generalized to large vocabularies containing over 9,000 words (median character error rate of 8.23%). These results illustrate the clinical viability of a silently controlled speech neuroprosthesis to generate sentences from a large vocabulary through a spelling-based approach, complementing previous demonstrations of direct full-word decoding.

2.2 Introduction

Devastating neurological conditions such as stroke and amyotrophic lateral sclerosis can lead to anarthria, the loss of ability to communicate through speech (Beukelman et al. 2007). Anarthric patients can have intact language skills and cognition, but paralysis may inhibit their ability to operate assistive devices, severely restricting communication with family, friends, and caregivers and reducing self-reported quality of life (Felgoise et al. 2016).

Brain-computer interfaces (BCIs) have the potential to restore communication to such patients by decoding neural activity into intended messages (Brumberg et al. 2018; Vansteensel et al. 2016). Existing communication BCIs typically rely on decoding imagined arm and hand movements into letters to enable spelling of intended sentences (Pandarinath et al. 2017; Willett et al. 2021). Although implementations of this approach have exhibited promising results, decoding natural attempts to speak directly into speech or text may offer faster and more natural control over a communication BCI. Indeed, a recent survey of prospective BCI users suggests that many patients would prefer speech-driven neuroprostheses over arm- and hand-driven neuroprostheses (Branco et al. 2021). Additionally, there have been several recent advances in the understanding of how the brain represents vocal-tract movements to produce speech (Bouchard et al. 2013; Carey et al. 2017; Chartier et al. 2018; Lotte et al. 2015) and demonstrations of text decoding from the brain activity of able speakers (Herff et al. 2015; Makin et al. 2020; Mugler et al. 2014; Sun et al. 2020; Dash, Ferrari, et al. 2020; Wilson et al. 2020; Cooney et al. 2022; Angrick et al. 2021), suggesting that decod-

ing attempted speech from brain activity could be a viable approach for communication restoration.

To assess this, we recently developed a speech neuroprosthesis to directly decode full words in real time from the cortical activity of a person with anarthria and paralysis as he attempted to speak (David A. Moses, Metzger, et al. 2021). This approach exhibited promising decoding accuracy and speed, but as an initial study focused on a preliminary 50-word vocabulary. While direct word decoding with a limited vocabulary has immediate practical benefit, expanding access to a larger vocabulary of at least 1000 words would cover over 85% of the content in natural English sentences (Adolphs et al. 2003) and enable effective day-to-day use of assistive-communication technology (Tilborg et al. 2016). Hence, a powerful complementary technology could expand current speech-decoding approaches to enable users to spell out intended messages from a large and generalizable vocabulary while still allowing fast, direct word decoding to express frequent and commonly used words. Separately, in this prior work the participant was controlling the neuroprosthesis by attempting to speak aloud, making it unclear if the approach would be viable for potential users who cannot produce any vocal output whatsoever.

Here, we demonstrate that real-time decoding of silent attempts to say 26 alphabetic code words from the NATO phonetic alphabet can enable highly accurate and rapid spelling in a clinical-trial participant (ClinicalTrials.gov; NCT03698149) with paralysis and anarthria. During training sessions, we cued the participant to attempt to produce individual code words and a hand-motor movement, and we used the simultaneously recorded cortical activity

from an implanted 128-channel electrocorticography (ECoG) array to train classification and detection models. After training, the participant performed spelling tasks in which he spelled out sentences in real time with a 1152-word vocabulary using attempts to silently say the corresponding alphabetic code words. A beam-search algorithm used predicted code-word probabilities from a classification model to find the most likely sentence given the neural activity while automatically inserting spaces between decoded words. To initiate spelling, the participant silently attempted to speak, and a speech-detection model identified this start signal directly from ECoG activity. After spelling out the intended sentence, the participant attempted the hand-motor movement to disengage the speller. When the classification model identified this hand-motor command from ECoG activity, a large neural network-based language model rescored the potential sentence candidates from the beam search and finalized the sentence. In post-hoc simulations, our system generalized well across large vocabularies of over 9000 words.

2.3 Results

Overview of the real-time spelling pipeline

We designed a sentence-spelling pipeline that enabled a clinical-trial participant (ClinicalTrials.gov; NCT03698149) with anarthria and paralysis to silently spell out messages using signals acquired from a high-density electrocorticography (ECoG) array implanted over his

sensorimotor cortex (Figure 2.1). We tested the spelling system under copy-typing and conversational task conditions. In each trial of the copy-typing task condition, the participant was presented with a target sentence on a screen and then attempted to replicate that sentence. In the conversational task condition, there were two types of trials: Trials in which the participant spelled out volitionally chosen responses to questions presented to him and trials in which he spelled out arbitrary, unprompted sentences. Prior to real-time testing, no day-of recalibration occured; model parameters and hyperparameters were fit using data exclusively from preceding sessions.

When the participant was ready to begin spelling a sentence, he attempted to silently say an arbitrary word (Figure 2.1a). We define silent-speech attempts as volitional attempts to articulate speech without vocalizing. Meanwhile, the participant's neural activity was recorded from each electrode and processed to simultaneously extract high-gamma activity (HGA; between 70 and 150 Hz) and low-frequency signals (LFS; between 0.3–100 Hz; Figure 2.1b). A speech-detection model processed each time point of data in the combined feature stream (containing HGA+LFS features; Figure 2.1c) to detect this initial silent-speech attempt.

Once an attempt to speak was detected, the paced spelling procedure began (Figure 2.1d). In this procedure, an underline followed by three dots appeared on the screen in white text. The dots disappeared one by one, representing a countdown. After the last dot disappeared, the underline turned green to indicate a go cue, at which time the participant attempted to silently say the NATO code word corresponding to the first letter in the sentence. The

time window of neural features from the combined feature stream obtained during the 2.5-s interval immediately following the go cue was passed to a neural classifier (Figure 2.1e). Shortly after the go cue, the countdown for the next letter automatically started. This procedure was then repeated until the participant volitionally disengaged it (described later in this section).

The neural classifier processed each time window of neural features to predict probabilities across the 26 alphabetic code words (Figure 2.1f). A beam-search algorithm used the sequence of predicted letter probabilities to compute potential sentence candidates, automatically inserting spaces into the letter sequences where appropriate and using a language model to prioritize linguistically plausible sentences. During real-time sentence spelling, the beam search only considered sentences composed of words from a predefined 1152-word vocabulary, which contained common words that are relevant for assistive-communication applications. The most likely sentence at any point in the task was always visible to the participant (Figure 2.1d). We instructed the participant to continue spelling even if there were mistakes in the displayed sentence, since the beam search could correct the mistakes after receiving more predictions. After attempting to silently spell out the entire sentence, the participant was instructed to attempt to squeeze his right hand to disengage the spelling procedure (Figure 2.1h). The neural classifier predicted the probability of this attempted hand-motor movement from each 2.5-s window of neural features, and if this probability was greater than 80%, the spelling procedure was stopped and the decoded sentence was finalized (Figure 2.1i). To finalize the sentence, sentences with incomplete words were first removed

from the list of potential candidates, and then the remaining sentences were rescored with a separate language model. The most likely sentence was then updated on the participant's screen (Figure 2.1g). After a brief delay, the screen was cleared and the task continued to the next trial.

To train the detection and classification models prior to real-time testing, we collected data as the participant performed an isolated-target task. In each trial of this task, a NATO code word appeared on the screen, and the participant was instructed to attempt to silently say the code word at the corresponding go cue. In some trials, an indicator representing the hand-motor command was presented instead of a code word, and the participant was instructed to imagine squeezing his right hand at the go cue for those trials.

Decoding performance

To evaluate the performance of the spelling system, we decoded sentences from the participant's neural activity in real time as he attempted to spell out 150 sentences (two repetitions each of 75 unique sentences selected from an assistive-communication corpus; see Table 2.1) during the copy-typing task. We evaluated the decoded sentences using word error rate (WER), character error rate (CER), words per minute (WPM), and characters per minute (CPM) metrics (Figure 2.2). For characters and words, the error rate is defined as the edit distance, which is the minimum number of character or word deletions, insertions, and substitutions required to convert the predicted sentence to the target sentence that was dis-

played to the participant, divided by the total number of characters or words in the target sentence, respectively. These metrics are commonly used to assess the decoding performance of automatic speech recognition systems (Hannun et al. 2014) and brain-computer interface applications (Willett et al. 2021; David A. Moses, Metzger, et al. 2021).

We observed a median CER of 6.13% and median WER of 10.53% (99% confidence interval (CI) [2.25, 11.6] and [5.76, 24.8]) across the real-time test blocks (each block contained multiple sentence-spelling trials; Figure 2.2a, b). Across 150 sentences, 105 (70%) were decoded without error, and 69 of the 75 sentences (92%) were decoded perfectly at least one of the two times they were attempted. Additionally, across 150 sentences, 139 (92.7%) sentences were decoded with the correct number of letters, enabled by high classification accuracy of the attempted hand squeeze (Figure 2.2e). We also observed a median CPM of 29.41 and median WPM of 6.86 (99% CI [29.1, 29.6] and [6.54, 7.12]) across test blocks, with spelling rates in individual blocks as high as 30.79 CPM and 8.60 WPM (Figure 2.2c, d). These rates are higher than the median rates of 17.37 CPM and 4.16 WPM (99% CI [16.1, 19.3] and [3.33, 5.05]) observed with the participant as he used his commercially available Tobii Dynavox assistive-typing device (as measured in our previous work (David A. Moses, Metzger, et al. 2021)).

To understand the individual contributions of the classifier, beam search, and language model to decoding performance, we performed offline analyses using data collected during these real-time copy-typing task blocks (Figure 2.2a, b). To examine the chance performance of the system, we replaced the model's predictions with randomly generated values while

continuing to use the beam search and language model. This resulted in a CER and WER that was significantly worse than the real-time results ($z = 7.09$, $P = 8.08 \times 10^{-12}$ and $z = 7.09$, $P = 8.08 \times 10^{-12}$ respectively, two-sided Wilcoxon Rank-Sum test with 6-way Holm-Bonferroni correction). This demonstrates that the classification of neural signals was critical to system performance and that system performance was not just relying on a constrained vocabulary and language-modeling techniques.

To assess how well the neural classifier alone could decode the attempted sentences, we compared character sequences composed of the most likely letter for each individual 2.5-second window of neural activity (using only the neural classifier) to the corresponding target character sequences. All whitespace characters were ignored during this comparison (during real-time decoding, these characters were inserted automatically by the beam search). This resulted in a median CER of 35.1% (99% CI [30.6, 38.5]), which is significantly lower than chance ($z = 7.09$, $P = 8.08 \times 10^{-12}$, two-sided Wilcoxon Rank-Sum test with 6-way Holm-Bonferroni correction), and shows that time windows of neural activity during silent code-word production attempts were discriminable. The median WER was 100% (99% CI [100.0, 100.0]) for this condition; without language modeling or automatic insertion of whitespace characters, the predicted character sequences rarely matched the corresponding target character sequences exactly.

To measure how much decoding was improved by the beam search, we passed the neural classifier's predictions into the beam search and constrained character sequences to be composed of only words within the vocabulary without incorporating any language modeling.

This significantly improved CER and WER over only using the most likely letter at each timestep ($z = 4.51$, $P = 6.37 \times 10^{-6}$ and $z = 6.61$, $P = 1.19 \times 10^{-10}$ respectively, two-sided Wilcoxon Rank-Sum test with 6-way Holm-Bonferroni correction). As a result of not using language modeling, which incorporates the likelihood of word sequences, the system would sometimes predict nonsensical sentences, such as "Do no tooth at again" instead of "Do not do that again" (Figure 2.2f). Hence, including language modeling to complete the full real-time spelling pipeline significantly improved median CER to 6.13% and median WER to 10.53% over using the system without any language modeling ($z = 5.53$, $P = 6.34 \times 10^{-8}$ and $z = 6.11$, $P = 2.01 \times 10^{-9}$ respectively, two-sided Wilcoxon Rank-Sum test with 6-way Holm-Bonferroni correction), illustrating the benefits of incorporating the natural structure of English during decoding.

Discriminatory content in high-gamma activity and low-frequency signals

Previous efforts to decode speech from brain activity have typically relied on content in the high-gamma frequency range (between 70 and 170 Hz, but exact boundaries vary) during decoding (Herff et al. 2015; Makin et al. 2020; David A. Moses, Leonard, et al. 2019). However, recent studies have demonstrated that low-frequency content (between 0 and 40 Hz) can also be used for spoken- and imagined-speech decoding (Mugler et al. 2014; Sun et al. 2020; Dash, Paul, et al. 2020; Proix et al. 2022; Anumanchipalli et al. 2019), although

the differences in the discriminatory information contained in each frequency range remain poorly understood.

In this work, we used both high-gamma activity (HGA; between 70 and 150 Hz) and low-frequency signals (LFS; between 0.3 and 16.67 Hz after downsampling with anti-aliasing) as neural features to enable sentence spelling. To characterize the speech content of each feature type, we used the most recent 10,682 trials of the isolated-target task) to train 10-fold cross-validated models using only HGA, only LFS, and both feature types simultaneously (HGA+LFS). In each of these trials, the participant attempted to silently say one of the 26 NATO code words. Models using only LFS demonstrated higher code-word classification accuracy than models using only HGA, and models using HGA+LFS outperformed the other two models ($z = 3.78$, $P = 4.71 \times 10^{-4}$ for all comparisons, two-sided Wilcoxon Rank-Sum test with 3-way Holm-Bonferroni correction; Figure 2.3a, Figure 2.10, Table 2.4), achieving a median classification accuracy of 54.2% (99% CI [51.6, 56.2], Figure 2.3a, Figure 2.11). Confusion matrices depicting the classification results with each model are depicted in Figure 2.11, Figure 2.12, and Figure 2.13.

We then investigated the relative contributions of each electrode and feature type to the neural classification models trained using HGA, LFS, and HGA+LFS. For each model, we first computed each electrode's contribution to classification by measuring the effect that small changes to the electrode's values had on the model's predictions (Simonyan et al. 2014). Electrode contributions for the HGA model were primarily localized to the ventral portion of the grid, corresponding to the ventral aspect of the ventral sensorimotor cortex (vSMC),

pars opercularis, and pars triangularis (Figure 2.3b). Contributions for the LFS model were much more diffuse, covering more dorsal and posterior parts of the grid corresponding to dorsal aspects of the vSMC in the pre- and postcentral gyri (Figure 2.3d). Contributions for the HGA model and the LFS model were moderately correlated with a Spearman rank correlation of 0.501 (n = 128 electrode contributions per feature type, $P < 0.01$). The separate contributions from HGA and LFS in the HGA+LFS model were highly correlated with the contributions for the HGA-only and LFS-only models, respectively (n = 128 electrode contributions per feature type, $P < 0.01$ for both Spearman rank correlations of 0.922 and 0.963, respectively; Figure 2.3c, e). These findings indicate that the information contained in the two feature types that was most useful during decoding was not redundant and was recorded from relatively distinct cortical areas.

To further characterize HGA and LFS features, we investigated whether LFS had increased feature or temporal dimensionality, which could have contributed to increased decoding accuracy. First, we performed principal component analysis (PCA) on the feature dimension for the HGA, LFS, and HGA+LFS feature sets. The resulting principal components (PCs) captured the spatial variability (across electrode channels) for the HGA and LFS feature sets and the spatial and spectral variabilities (across electrode channels and feature types, respectively) for the HGA + LFS feature set. To explain more than 80% of the variance, LFS required significantly more feature PCs than HGA ($z = 12.2$, $P = 7.57 \times 10^{-34}$, two-sided Wilcoxon Rank-Sum test with 3-way Holm-Bonferroni correction; Figure 2.3f) and the combined HGA+LFS feature set required significantly more feature

PCs than the individual HGA or LFS feature sets (z = 12.2, P = 7.57×10^{-34} and z = 11.6, P = 2.66×10^{-33}, respectively, two-sided Wilcoxon Rank-Sum test with 3-way Holm-Bonferroni correction; Figure 2.3f). This suggests that LFS did not simply replicate HGA at each electrode but instead added unique feature variance.

To assess the temporal content of the features, we first used a similar PCA approach to measure temporal dimensionality. We observed that the LFS features required significantly more temporal PCs than both the HGA and HGA+LFS feature sets to explain more than 80% of the variance (z = 12.2, P = 7.57×10^{-34} and z = 12.2, P = 7.57×10^{-34}, respectively, Figure 2.3g; two-sided Wilcoxon Rank-Sum test with 3-way Holm-Bonferroni correction). Because the inherent temporal dimensionality for each feature type remained the same within the HGA+LFS feature set, the required number of temporal PCs to explain this much variance for the HGA+LFS features was in between the corresponding numbers for the individual feature types. Then, to assess how the temporal resolution of each feature type affected decoding performance, we temporally smoothed each feature time series with Gaussian filters of varying widths. A wider Gaussian filter causes a greater amount of temporal smoothing, effectively temporally blurring the signal and hence lowering temporal resolution. Temporally smoothing the LFS features decreased the classification accuracy significantly more than smoothing the HGA or HGA+LFS features (Wilcoxon signed-rank statistic = 737.0, P = 4.57×10^{-5} and statistic = 391.0, P = 1.13×10^{-8}, two-sided Wilcoxon signed-rank test with 3-way Holm-Bonferroni correction; Figure 2.3h). The effects of temporal smoothing were not significantly different between HGA and HGA+LFS (Wilcoxon

signed-rank statistic = 1460.0, P = 0.443). This is largely consistent with the outcomes of the temporal-PCA comparisons. Together, these results indicate that the temporal content of LFS had higher variability and contained more speech-related discriminatory information than HGA.

Differences in neural discriminability between NATO code words and letters

During control of our system, the participant attempted to silently say NATO code words to represent each letter ("alpha" instead of "a", "beta" instead of "b", and so forth) rather than simply saying the letters themselves. We hypothesized that neural activity associated with attempts to produce code words would be more discriminable than letters due to increased phonetic variability and longer utterance lengths. To test this, we first collected data using a modified version of the isolated-target task in which the participant attempted to say each of the 26 English letters instead of the NATO code words that represented them. Afterwards, we trained and tested classification models using HGA+LFS features from the most recent 29 attempts to silently say each code word and each letter in 10-fold cross-validated analyses. Indeed, code words were classified with significantly higher accuracy than the letters (z = 3.78, P = 1.57×10^{-4}, two-sided Wilcoxon Rank-Sum test; Figure 2.4a).

To perform a model-agnostic comparison between the neural discriminability of each type of utterance (either code words or letters), we computed nearest-class distances for each

utterance using the HGA+LFS feature set. Here, each utterance represented a single class, and distances were only computed between utterances of the same type. A larger nearest-class distance for a code word or letter indicates that that utterance is more discriminable in neural feature space because the neural activation patterns associated with silent attempts to produce it are more distinct from other code words or letters, respectively. We found that nearest-class distances for code words were significantly higher overall than for letters ($z = 2.98$, $P = 2.85 \times 10^{-3}$, two-sided Wilcoxon Rank-Sum test; Figure 2.4b), although not all code words had a higher nearest-class distance than its corresponding letter (Figure 2.4c).

Distinctions in evoked neural activity between silent- and overt-speech attempts

The spelling system was controlled by silent-speech attempts, differing from our previous work in which the same participant used overt-speech attempts (attempts to speak aloud) to control a similar speech-decoding system (David A. Moses, Metzger, et al. 2021). To assess differences in neural activity and decoding performance between the two types of speech attempts, we collected a version of the isolated-target task in which the participant was instructed to attempt to say the code words aloud (overtly instead of silently). The spatial patterns of evoked neural activity for the two types of speech attempts exhibited similarities (Figure 2.14), and inspections of evoked HGA for two electrodes suggest that some neural populations respond similarly for each speech type while others do not (Figure 2.5a–c).

To compare the discriminatory neural content between silent- and overt-speech attempts, we performed 10-fold cross-validated classification analyses using HGA+LFS features associated with the speech attempts (Figure 2.5d). First, for each type of attempted speech (silent or overt), we trained a classification model using data collected with that speech type. To determine if the classification models could leverage similarities in the neural representations associated with each speech type to improve performance, we also created models by pre-training on one speech type and then fine-tuning on the other speech type. We then tested each classification model on held-out data associated with each speech type and compared all 28 combinations of pairs of results (all statistical results detailed in Table 2.7). Models trained solely on silent data but tested on overt data and vice versa resulted in classification accuracies that were above chance (median accuracies of 36.3%, 99% CI [35.0, 37.5] and 33.5%, 99% CI [31.0, 35.0], respectively; chance accuracy is 3.85%). However, for both speech types, training and testing on the same type resulted in significantly higher performance ($P < 0.01$, two-sided Wilcoxon Rank-Sum test, 28-way Holm-Bonferroni correction). Pre-training models using the other speech type led to increases in classification accuracy, though the increase was more modest and not significant for the overt speech type (median accuracy increasing by 2.33%, $z = 2.65$, $P = 0.033$ for overt, median accuracy increasing by 10.4%, $z = 3.78$, $P = 4.40 \times 10^{-3}$ for silent, two-sided Wilcoxon Rank-Sum test, 28-way Holm-Bonferroni correction). Together, these results suggest that the neural activation patterns evoked during silent and overt attempts to speak shared some similarities but were not identical.

Generalizability to larger vocabularies and alternative tasks

Although the 1152-word vocabulary enabled communication of a wide variety of common sentences, we also assessed how well our approach can scale to larger vocabulary sizes. Specifically, we simulated the copy-typing spelling results using three larger vocabularies composed of 3303, 5249, and 9170 words that we selected based on their words' frequencies in large-scale English corpora. For each vocabulary, we retrained the language model used during the beam search to incorporate the new words. The large language model used when finalizing sentences was not altered for these analyses because it was designed to generalize to any English text.

High performance was maintained with each of the new vocabularies, with median character error rates (CERs) of 7.18% (99% CI [2.25, 11.6]), 7.93% (99% CI [1.75, 12.1]), and 8.23% (99% CI [2.25, 13.5]) for the 3303-, 5249-, and 9170-word vocabularies, respectively (Figure 2.6a; median real-time CER was 6.13% (99% CI [2.25, 11.6]) with the original vocabulary containing 1,152 words). Median word error rates (WERs) were 12.4% (99% CI [8.01, 22.7]), 11.1% (99% CI [8.01, 23.1]), and 13.3% (99% CI [7.69, 28.3]), respectively (Figure 2.6b; WER was 10.53% (99% CI [5.76, 24.8]) for the original vocabulary). Overall, no significant differences were found between the CERs or WERs with any two vocabularies ($P > 0.01$ for all comparisons, two-sided Wilcoxon Rank-Sum test with 6-way Holm-Bonferroni correction), illustrating the generalizability of our spelling approach to larger vocabulary sizes that enable fluent communication.

Finally, to assess the generalizability of our spelling approach to behavioral contexts beyond the copy-typing task structure, we measured performance as the participant engaged in a conversational task condition. In each trial of this condition, the participant was either presented with a question (as text on a screen) or was not presented with any stimuli. He then attempted to spell out a volitionally chosen response to the presented question or any arbitrary sentence if no stimulus was presented. To measure the accuracy of each decoded sentence, we asked the participant to nod his head to indicate if the sentence matched his intended sentence exactly. If the sentence was not perfectly decoded, the participant used his commercially available assistive-communication device to spell out his intended message. Across 28 trials of this real-time conversational task condition, the median CER was 14.8% (99% CI [0.00, 29.7]) and the median WER was 16.7% (99% CI [0.00, 44.4]) (Figure 2.6c, d). We observed a slight increase in decoding error rates compared to the copy-typing task, potentially due to the participant responding using incomplete sentences (such as "going out" and "summer time") that would not be well represented by the language models. Nevertheless, these results demonstrate that our spelling approach can enable a user to generate responses to questions as well as unprompted, volitionally chosen messages.

2.4 Discussion

Here, we demonstrated that a paralyzed clinical-trial participant (ClinicalTrials.gov; NCT03698149) with anarthria could control a neuroprosthesis to spell out intended mes-

sages in real time using attempts to silently speak. With phonetically rich code words to represent individual letters and an attempted hand movement to indicate an end-of-sentence command, we used deep-learning and language-modeling techniques to decode sentences from electrocorticographic (ECoG) signals. These results significantly expand our previous word-decoding findings with the same participant (David A. Moses, Metzger, et al. 2021) by enabling completely silent control, leveraging both high- and low-frequency ECoG features, including a non-speech motor command to finalize sentences, facilitating large-vocabulary sentence decoding through spelling, and demonstrating continued stability of the relevant cortical activity beyond 128 weeks since device implantation.

Previous implementations of spelling brain-computer interfaces (BCIs) have demonstrated that users can type out intended messages by visually attending to letters on a screen (Rezeika et al. 2018; Sellers et al. 2014) or by using motor imagery to control a two-dimensional computer cursor (Vansteensel et al. 2016; Pandarinath et al. 2017) or attempt to handwrite letters (Willett et al. 2021). BCI performance using penetrating microelectrode arrays in motor cortex has steadily improved over the past 20 years (Gilja et al. 2012; Kawala-Sterniuk et al. 2021; Serruya et al. 2002), recently achieving spelling rates as high as 90 characters per minute with a single participant (Willett et al. 2021), although this participant was able to speak normally. Our results extend the list of immediately practical and clinically viable control modalities for spelling-BCI applications to include silently attempted speech with an implanted ECoG array, which may be preferred for daily use by some patients due to the relative naturalness of speech (Branco et al. 2021) and may be more chronically ro-

bust across patients through the use of less invasive, non-penetrating electrode arrays with broader cortical coverage.

In post-hoc analyses, we showed that decoding performance improved as more linguistic information was incorporated into the spelling pipeline. This information helped facilitate real-time decoding with a 1152-word vocabulary, allowing for a wide variety of general and clinically relevant sentences as possible outputs. Furthermore, through offline simulations, we validated this spelling approach with vocabularies containing over 9000 common English words, which exceeds the estimated lexical-size threshold for basic fluency and enables general communication (Laufer 1989; Webb et al. 2009). These results add to consistent findings that language modeling can significantly improve neural-based speech decoding (Herff et al. 2015; Sun et al. 2020; David A. Moses, Metzger, et al. 2021) and demonstrates the immediate viability of speech-based spelling approaches for a general-purpose assistive-communication system.

In this study, we showed that neural signals recorded during silent-speech attempts by an anarthric person can be effectively used to drive a speech neuroprosthesis. Supporting the hypothesis that these signals contained similar speech-motor representations to signals recorded during overt-speech attempts, we showed that a model trained solely to classify overt-speech attempts can achieve above-chance classification of silent-speech attempts, and vice versa. Additionally, the spatial localization of electrodes contributing most to classification performance was similar for both overt and silent speech, with many of these electrodes located in the ventral sensorimotor cortex, a brain area that is heavily implicated in artic-

ulatory speech-motor processing (Bouchard et al. 2013; Carey et al. 2017; Chartier et al. 2018; Conant et al. 2018).

Overall, these results further validate silently attempted speech as an effective alternative behavioral strategy to imagined speech and expand findings from our previous work involving the decoding of overt-speech attempts with the same participant (David A. Moses, Metzger, et al. 2021), indicating that the production of residual vocalizations during speech attempts is not necessary to control a speech neuroprosthesis. These findings illustrate the viability of attempted-speech control for individuals with complete vocal-tract paralysis (such as those with locked-in syndrome), although future studies with these individuals are required to further our understanding of the neural differences between overt-speech attempts, silent-speech attempts, and purely imagined speech as well as how specific medical conditions might affect these differences. We expect that the approaches described here, including recording methodology, task design, and modeling techniques, would be appropriate for both speech-related neuroscientific investigations and BCI development with patients regardless of the severity of their vocal-tract paralysis, assuming that their speech-motor cortices are still intact and that they are mentally capable of attempting to speak.

In addition to enabling spatial coverage over the lateral speech-motor cortical brain regions, the implanted ECoG array also provided simultaneous access to neural populations in the hand-motor (hand knob) cortical area that is typically implicated during executed or attempted hand movements (Gerardin et al. 2000). Our approach is the first to combine the two cortical areas to control a BCI. This ultimately enabled our participant to use an

attempted hand movement, which was reliably detectable and highly discriminable from silent-speech attempts with 98.43% classification accuracy (99% CI [95.31, 99.22]), to indicate when he was finished spelling any particular sentence. This may be a preferred stopping mechanism compared to previous spelling BCI implementations that terminated spelling for a sentence after a pre-specified time interval had elapsed or extraneously when the sentence was completed (Pandarinath et al. 2017) or required a head movement to terminate the sentence (Willett et al. 2021). By also allowing a silent-speech attempt to initiate spelling, the system could be volitionally engaged and disengaged by the participant, which is an important design feature for a practical communication BCI. Although attempted hand movement was only used for a single purpose in this first demonstration of a multimodal communication BCI, separate work with the same participant suggests that non-speech motor imagery could be used to indicate several distinct commands (Silversmith et al. 2021).

One drawback of the current approach is that it relies on code words instead of letters during spelling. Although the use of these longer code words improved neural discriminability, they are less natural to use. Separately, the participant had to attempt to produce code words at a pre-defined pace during spelling, which enabled straightforward parcellation of the neural activity into separate time windows for classification but reduced flexibility for the user. Future work can focus on improving letter decoding and implementing flexible, user-controlled pacing (for example, through augmented speech-attempt detection) to facilitate more naturalistic spelling. Additionally, the present results are limited to only one participant; to fully assess the clinical viability of this spelling system as a neuroprosthesis,

it will need to be validated with more participants.

In future communication neuroprostheses, it may be possible to use a combined approach that enables rapid decoding of full words or phrases from a limited, frequently used vocabulary (David A. Moses, Metzger, et al. 2021) as well as slower, generalizable spelling for out-of-vocabulary items. Transfer-learning methods could be used to cross-train differently purposed speech models using data aggregated across multiple tasks and vocabularies, as validated in previous speech-decoding work (Makin et al. 2020). Although clinical and regulatory guidelines concerning the implanted percutaneous connector prevented the participant from being able to use the current spelling system independently, development of a fully implantable ECoG array and a software application to integrate the decoding pipeline with an operating system's accessibility features could allow for autonomous usage. Facilitated by deep-learning techniques, language modeling, and the signal stability and spatial coverage afforded by ECoG recordings, future communication neuroprostheses could enable users with severe paralysis and anarthria to control assistive technology and personal devices using naturalistic silent-speech attempts to generate intended messages and attempted non-speech motor movements to issue high-level, interactive commands.

2.5 Methods

Clinical trial overview

This study was conducted as part of the BCI Restoration of Arm and Voice (BRAVO) clinical trial (ClinicalTrials.gov; NCT03698149). The goal of this single-institution clinical trial is to assess the incidence of treatment-emergent adverse events associated with the ECoG-based neural interface and to determine if ECoG and custom decoding methods can enable long-term assistive neurotechnology to restore communication and mobility. The data presented here and the present work do not support or inform any conclusions about the primary outcomes of this trial. The clinical trial began in November 2018. The Food and Drug Administration approved an investigational device exemption for the neural implant used in this study. The study protocol was approved by the Committee on Human Research at the University of California, San Francisco. The data safety monitoring board agreed to the release of the results of this work prior to the completion of the trial. The participant gave his informed consent to participate in this study after the details concerning the neural implant, experimental protocols, and medical risks were thoroughly explained to him.

Participant

The participant, who was 36 years old at the start of the study, was diagnosed with severe spastic quadriparesis and anarthria by neurologists and a speech-language pathologist after experiencing an extensive pontine stroke. He is fully cognitively intact. Although he retains

the ability to vocalize grunts and moans, he is unable to produce intelligible speech, and his attempts to speak aloud are abnormally effortful due to his condition (according to self-reported descriptions). He typically relies on assistive computer-based interfaces that he controls with residual head movements to communicate. This participant has participated in previous studies as part of this clinical trial (David A. Moses, Metzger, et al. 2021; Silversmith et al. 2021), although neural data from those studies were not used in the present study. He provided verbal consent (using his assistive computer-based interface) to participate in the study and allow his image to appear in material accompanying this chapter. He also provided verbal consent (again using this interface) to have a designated third-party individual physically sign the consent forms on his behalf.

Neural implant

The neural implant device consisted of a high-density electrocorticography (ECoG) array (PMT) and a percutaneous connector (Blackrock Microsystems) (David A. Moses, Metzger, et al. 2021). The ECoG array contained 128 disk-shaped electrodes arranged in a lattice formation with 4-mm center-to-center spacing. The array was surgically implanted on the pial surface of the left hemisphere of the brain over cortical regions associated with speech production, including the dorsal posterior aspect of the inferior frontal gyrus, the posterior aspect of the middle frontal gyrus, the precentral gyrus, and the anterior aspect of the postcentral gyrus (Bouchard et al. 2013; Chartier et al. 2018; Guenther et al. 2016). The

percutaneous connector was implanted in the skull to conduct electrical signals from the ECoG array to a detachable digital headstage and cable (NeuroPlex E; Blackrock Microsystems), minimally processing and digitizing the acquired brain activity and transmitting the data to a computer. The device was implanted in February 2019 at UCSF Medical Center without any surgical complications.

Data acquisition and preprocessing

We acquired neural features from the implanted ECoG array using a pipeline involving several hardware components and processing steps (Figure 2.8). We connected a headstage (a detachable digital connector; NeuroPlex E, Blackrock Microsystems) to the percutaneous pedestal connector, which digitized neural signals from the ECoG array and transmitted them through an HDMI connection to a digital hub (Blackrock Microsystems). The digital hub then transmitted the digitized signals through an optical fiber cable to a Neuroport system (Blackrock Microsystems), which applied noise cancellation and an anti-aliasing filter to the signals before streaming them at 1 kHz through an Ethernet connection to a separate real-time computer (Colfax International). The Neuroport system was controlled using the NeuroPort Central Suite software package (version 7.0.4; Blackrock Microsystems).

On the real-time processing computer, we used a custom Python software package (rtNSR) to process and analyze the ECoG signals, execute the real-time tasks, perform real-time decoding, and store the data and task metadata (David A. Moses, Metzger, et al. 2021; David

A. Moses, Leonard, et al. 2019; David A Moses et al. 2018). Using this software package, we first applied a common average reference (across all electrode channels) to each time sample of the ECoG data. Common average referencing is commonly applied to multi-channel datasets to reduce shared noise (Ludwig et al. 2009; Williams et al. 2018). These re-referenced signals were then processed in two parallel processing streams to extract high-gamma activity (HGA) and low-frequency signal (LFS) features using digital finite impulse response (FIR) filters designed using the Parks-McClellan algorithm (Parks et al. 1972) (Figure 2.8; filters were designed using the SciPy Python package (version 1.5.4)). Briefly, we used these FIR filters to compute the analytic amplitude of the signals in the high-gamma frequency band (70–150 Hz) and an anti-aliased version of the signals (with a cutoff frequency at 100 Hz). We combined the time-synchronized high-gamma analytic amplitudes and downsampled signals into a single feature stream at 200 Hz. Next, we z-scored the values for each channel and each feature type using a 30-s sliding window to compute running statistics. Finally, we implemented an artifact-rejection approach that identified neural time points containing at least 32 features with z-score magnitudes greater than 10, replacing each of these time points with the z-score values from the preceding time point and ignoring these time points when updating the running z-score statistics. During real-time decoding and in offline analyses, we used the z-scored high-gamma analytic amplitudes as the HGA features and the z-scored downsampled signals as the LFS features (and the combination of the two as the HGA+LFS feature set). The neural classifier further downsampled these feature streams by a factor of 6 before using them for inference (using an anti-aliasing filter with a cutoff frequency at 16.67

Hz), but the speech detector did not.

We performed all data collection and real-time decoding tasks in the participant's bedroom or a small office room nearby. We uploaded data to our lab's server infrastructure and trained the decoding models using NVIDIA V100 GPUs hosted on this infrastructure.

Task design

We recorded neural data with the participant during two general types of tasks: an isolated-target task and a sentence-spelling task (Figure 2.7). In each trial of the isolated-target task, a text target appeared on the screen along with 4 dots on either side. Dots on both sides disappeared one by one until no dots remained, at which point the text target turned green to represent a go cue. At this go cue, the participant either attempted to say the target (silently or aloud, depending on the current task instructions) if it was either a NATO code word or an English letter. If the target was a text string containing the word "Right" and an arrow pointing right, the participant instead attempted to squeeze his right hand. We used the neural data collected during the isolated-target task to train and optimize the detection and classification models and to evaluate classifier performance.

The sentence-spelling task is described in the start of the Results section and in Figure 2.1. Briefly, the participant used the full spelling pipeline (described in the following sub-section) to either spell sentences presented to him as targets in a copy-typing task condition or to spell arbitrary sentences in a conversational task condition. We did not implement functionality

to allow the participant to retroactively alter the predicted sentence, although the language model could alter previously predicted words in a sentence after receiving additional character predictions. Data collected during the sentence-spelling task were used to optimize beam-search hyperparameters and evaluate the full spelling pipeline.

Modeling

We fit detection and classification models using data collected during the isolated-target task as the participant attempted to produce code words and the hand-motor command. After fitting these models offline, we saved the trained models to the real-time computer for use during real-time testing. We implemented these models using the PyTorch Python package (version 1.6.0). In addition to these two models, we also used language models to enable sentence spelling. We used hyperparameter optimization procedures on held-out validation datasets to choose values for model hyperparameters (see Table 2.8). We used the Python software packages NumPy (version 1.19.1), scikit-learn (version 0.24.2), and pandas (version 0.25.3) during modeling and data analysis.

Speech detection

To determine when the participant was attempting to engage the spelling system, we developed a real-time silent-speech detection model. This model used long short-term memory layers, a type of recurrent neural network layer, to process neural activity in real time and detect attempts to silently speak (David A. Moses, Metzger, et al. 2021). This model used both

LFS and HGA features (a total of 256 individual features) at 200 Hz. The speech-detection model was trained using supervised learning and truncated backpropagation through time. For training, we labeled each time point in the neural data as one of four classes depending on the current state of the task at that time: 'rest', 'speech preparation', 'motor', and 'speech.' Though only the speech probabilities were used during real-time evaluation to engage the spelling system, the other labels were included during training to help the detection model disambiguate attempts to speak from other behavior. See Figure 2.9 for further details about the speech-detection model.

Classification

We trained an artificial neural network (ANN) to classify the attempted code word or hand-motor command y_i from the time window of neural activity x_i associated with an isolated-target trial or 2.5-s letter-decoding cycle i. The training procedure was a form of maximum likelihood estimation, where given an ANN classifier parameterized by θ and conditioned on the neural activity x_i, our goal during model fitting was to find the parameters θ^* that maximized the probability of the training labels. This can be written as the following optimization problem:

$$\theta^* = \arg\max_{\theta} \prod_i p_\theta(y_i|x_i) = \arg\min_{\theta} -\sum_i \log p_\theta(y_i|x_i) \tag{2.1}$$

We approximated the optimal parameters θ^* using stochastic gradient descent and the Adam optimizer (Kingma et al. 2017).

To model the temporal dynamics of the neural time-series data, we used an ANN with a one-dimensional temporal convolution on the input layer followed by two layers of bidirectional gated recurrent units (GRUs) (Cho et al. 2014), for a total of three layers. We multiplied the final output of the last GRU layer by an output matrix and then applied a softmax function to yield the estimated probability of each of the 27 labels $\hat{y}_i$ given x_i.

Classifier ensembling for sentence spelling

During sentence spelling, we used model ensembling to improve classification performance by reducing overfitting and unwanted modeling variance caused by random parameter initializations (Fort et al. 2020). Specifically, we trained 10 separate classification models using the same training dataset and model architecture but with different random parameter initializations. Then, for each time window of neural activity x_i, we averaged the predictions from these 10 different models together to produce the final prediction $\hat{y}_i$.

Incremental classifier recalibration for sentence spelling

To improve sentence-spelling performance, we trained the classifiers used during sentence spelling on data recorded during sentence-spelling tasks from preceding sessions (in addition to data from the isolated-target task). In an effort to only include high-quality sentence-spelling data when training these classifiers, we only used data from sentences that were decoded with a character error rate of 0.

Beam search

During sentence spelling, our goal was to compute the most likely sentence text s^* given the neural data X. We used the formulation from Hannun et al. (Hannun et al. 2014) to find s^* given its likelihood from the neural data and its likelihood under an adjusted language-model prior, which allowed us to incorporate word-sequence probabilities with predictions from the neural classifier. This can be expressed formulaically as:

$$s^* = \arg\max_s p_{nc}(s|X)p_{lm}(s^\alpha)|s|^\beta \tag{2.2}$$

Here, $p_{nc}(s|X)$ is the probability of s under the neural classifier given each window of neural activity, which is equal to the product of the probability of each letter in s given by the neural classifier for each window of neural activity x_i. $p_{lm}(s)$ is the probability of the sentence s under a language-model prior. Here, we used an n-gram language model to approximate $p_{lm}(s)$. Our n-gram language model, with n = 3, provides the probability of each word given the preceding two words in a sentence. We implemented this language model using custom code as well as utility functions from the NLTK Python package (version 3.6.2). The probability under the language model of a sentence is then taken as the product of the probability of each word given the two words that precede it.

As in Hannun et al. (Hannun et al. 2014), we assumed that the n-gram language-model prior was too strong and downweighted it using a hyperparameter α. We also included a word-insertion bonus β to encourage the language model to favor sentences containing

more words, counteracting an implicit consequence of the language model that causes the probability of a sentence under it $p_{lm}(s)$ to decrease as the number of words in s increases. $|s|$ denotes the cardinality of s, which is equal to the number of words in s. If a sentence s was partially completed, only the words preceding the final whitespace character in s were considered when computing $p_{lm}(s)$ and $|s|$.

We then used an iterative beam-search algorithm as in Hannun et al. (Hannun et al. 2014) to approximate s^* at each timepoint $t = \tau$. We used a list of the B most likely sentences from $t = \tau - 1$ (or a list containing a single empty-string element if t = 1) as a set of candidate prefixes, where B is the beam width. Then, for each candidate prefix l and each English letter c with $p_{nc}(c|x_\tau) > 0.001$, we constructed new candidate sentences by considering l followed by c. Additionally, for each candidate prefix l and each text string c^+, composed of an English letter followed by the whitespace character, with $p_{nc}(c^+|x_\tau) > 0.001$, we constructed more new candidate sentences by considering l followed by c^+. Here and throughout the beam search, we considered $p_{nc}(c^+|x_\tau) = p_{nc}(c|x_\tau)$ for each c and corresponding c^+. Next, we discarded any resulting candidate sentences that contained words or partially completed words that were not valid given our constrained vocabulary. Then, we rescored each remaining candidate sentence $\tilde{l}$ with $p(\tilde{l}) = p_{nc}(\tilde{l}|X_{1:\tau})p_{lm}(\tilde{l})^\alpha|\tilde{l}|^\beta$. The most likely candidate sentence, s^*, was then displayed as feedback to the participant

We chose values for α, β, and B using hyperparameter optimization.

If at any time point t the probability of the attempted hand-motor command (the sentence-finalization command) was $> 80\%$, the B most likely sentences from the previ-

ous iteration of the beam search were processed to remove any sentence with incomplete or out-of-vocabulary words. The probability of each remaining sentence $\hat{l}$ was then recomputed as:

$$p(\hat{l}) = p_{nc}(\hat{l}|X_{1:t-1})p_{lm}(\hat{l})^{\alpha}|\hat{l}|^{\beta}p_{gpt2}(\hat{l})^{\alpha_{gpt2}} \tag{2.3}$$

Here, $p_{gpt2}(\hat{l})$ denotes the probability of $\tilde{l}$ under the DistilGPT-2 language model, a low-parameter variant of GPT-2 (implemented using the lm-scorer Python package (version 0.4.2)), and α_{gpt2} represents a scaling hyperparameter that was set through hyperparameter optimization. The most likely sentence $\tilde{l}$ given this formulation was then displayed to the participant and stored as the finalized sentence.

Performance evaluation

Character error rate and word error rate

Because CER and WER are overly influenced by short sentences, as in previous studies (Willett et al. 2021; David A. Moses, Metzger, et al. 2021) we reported CER and WER as the sum of the character or word edit distances between each of the predicted and target sentences in a sentence-spelling block and then divided this number by the total number of characters or words across all target sentences in the block. Each block contained between two to five sentence trials.

Assessing performance during the conversational task condition

To obtain ground-truth sentences to calculate CERs and WERs for the conversational condition of the sentence-spelling task, after completing each block we reminded the participant of the questions and the decoded sentences from that block, and then, for each decoded sentence, he either confirmed that the decoded sentence was correct or typed out the intended sentence using his commercially available assistive-communication device. Each block contained between two to four sentence trials.

Characters and words per minute

We calculated the characters per minute and words per minute rates for each sentence-spelling (copy-typing) block as follows:

$$\text{rate} = \frac{\sum_i N_i}{\sum_i D_i} \tag{2.4}$$

Here, i indexes each trial, N_i denotes the number of words or characters (including whitespace characters) decoded for trial i, and D_i denotes the duration of trial i (in minutes; computed as the difference between the time at which the window of neural activity corresponding to the final code word in trial i ended and the time of the go cue of the first code word in trial i).

Electrode contributions

To compute electrode contributions using data recorded during the isolated-target task, we computed the derivative of the classifier's loss function with respect to the input features across time as in Simonyan et al. (Simonyan et al. 2014), yielding a measure of how much the predicted model outputs were affected by small changes to the input feature values for each electrode and feature type (HGA or LFS) at each time point. Then, we calculated the L2-norm of these values across time and averaged the resulting values across all isolated-target trials, yielding a single contribution value for each electrode and feature type for that classifier.

Cross-validation

For each fold, we used stratified cross-validation folds of the isolated-target task. We split each fold into a training set containing 90% of the data and a held-out testing set containing the remaining 10%. In all, 10% of the training dataset was then randomly selected (with stratification) as a validation set.

Analyzing neural-feature principal components

To characterize the HGA and LFS neural features, we used bootstrapped principal component analyses. First, for each NATO code word, we randomly sampled (with replacement) cue-aligned time windows of neural activity (spanning from the go cue to 2.5 s after the go cue) from the first 318 silently attempted isolated-target trials for that code word. To clearly

understand the role of each feature stream for classification, we downsampled the signals by a factor of 6 to obtain the signals used by the classifier. Then, we trial averaged the data for each code word, yielding 26 trial averages across time for each electrode and feature set (HGA, LFS, and HGA+LFS). We then arranged this into a matrix with dimensionality N $\times$ TC, where N is the number of features (128 for HGA and for LFS; 256 for HGA+LFS), T is the number of time points in each 2.5-s window, and C is the number of NATO code words (26), by concatenating the trial-averaged activity for each feature. We then performed principal component analysis along the feature dimension of this matrix. Additionally, we arranged the trial-averaged data for each code word into a matrix with dimensionality T $\times$ NC. We then performed principal component analysis along the temporal dimension. For each analysis, we performed the measurement procedure 100 times to obtain a representative distribution of the minimum number of principal components required to explain more than 80% of the variance.

Nearest-class distance comparison

To compare nearest-class distances for the code words and letters, we first calculated averages across 1000 bootstrap iterations of the combined HGA+LFS feature set across 47 silently attempted isolated-target trials for each code word and letter. We then computed the Frobenius norm of the difference between each pairwise combination. For each code word, we used the smallest computed distance between that code word and any other code word as the nearest-class distance. We then repeated this process for the letters.

Generalizability to larger vocabularies

During real-time sentence spelling, the participant created sentences composed of words from a 1152-word vocabulary that contained common words and words relevant to clinical caregiving. To assess the generalizability of our system, we tested the sentence-spelling approach in offline simulations using three larger vocabularies. The first of these vocabularies was based on the 'Oxford 3000' word list, which is composed of 3000 core words chosen based on their frequency in the Oxford English Corpus and relevance to English speakers (*About the Oxford 3000 and 5000 word lists at Oxford Learner's Dictionaries* 2021). The second was based on the 'Oxford 5000' word list, which is the 'Oxford 3000' list augmented with an additional 2,000 frequent and relevant words. The third was a vocabulary based on the most frequent 10,000 words in Google's Trillion Word Corpus, a corpus that is over 1 trillion words in length (Brants et al. 2006). To eliminate non-words that were included in this list (such as "f", "gp", and "ooo"), we excluded words composed of 3 or fewer characters if they did not appear in the 'Oxford 5000' list. After supplementing each of these three vocabularies with the words from the original 1152-word vocabulary that were not already included, the three finalized vocabularies contained 3303, 5249, and 9170 words (these sizes are given in the same order that the vocabularies were introduced).

For each vocabulary, we retrained the n-gram language model used during the beam-search procedure with n-grams that were valid under the new vocabulary and used the larger vocabulary during the beam search. We then simulated the sentence-spelling experiments

offline using the same hyperparameters that were used during real-time testing.

Statistics and reproducibility

Statistical analyses

The statistical tests used in this work are all described in the figure captions and text. In brief, we used two-sided Wilcoxon Rank-Sum tests to compare any two groups of observations. When the observations were paired, we instead used a two-sided Wilcoxon signed-rank test. We used Holm-Bonferroni correction for comparisons in which the underlying neural data were not independent of each other. We considered P-values < 0.01 as significant. We computed P-values for Spearman rank correlations using permutation testing. For each permutation, we randomly shuffled one group of observations and then determined the correlation. We computed the P-value as the fraction of permutations that had a correlation value with a larger magnitude than the Spearman rank correlation computed on the non-shuffled observations. For any confidence intervals around a reported metric, we used a bootstrap approach to estimate the 99% confidence interval. On each iteration (of a total of 2000 iterations), we randomly sampled the data (such as accuracy per cross-validation fold) with replacement and calculated the desired metric (such as the median). The confidence interval was then computed on this distribution of the bootstrapped metric. We used SciPy (version 1.5.4) during statistical testing.

Reproducibility of experiments

Because this is a pilot study with a single participant, further work is required to definitively determine if the current approach is reproducible with other participants.

Data exclusions

During the copy-typing condition of the sentence-spelling task, the participant was instructed to attempt to silently spell each intended sentence regardless of how accurate the decoded sentence displayed as feedback was. However, during a small number of trials, the participant self-reported making a mistake (for example, by using the wrong code word or forgetting his place in the sentence) and sometimes stopped his attempt. This mostly occurred during initial sentence-spelling sessions while he was still getting accustomed to the interface. To focus on evaluating the performance of our system rather than the participant's performance, we excluded these trials (13 trials out of 163 total trials) from performance-evaluation analyses, and we had the participant attempt to spell the sentences in these trials again in subsequent sessions to maintain the desired amount of trials during performance evaluation (2 trials for each of the 75 unique sentences). Including these rejected sentences when evaluating performance metrics only modestly increased the median CER and WER observed during real-time spelling to 8.52% (99% CI [3.20, 15.1]) and 13.75% (99% CI [8.71, 29.9]), respectively.

During the conversational condition of the sentence-spelling task, trials were rejected if the participant self-reported making a mistake (as in the copy-typing condition) or if an

intended word was outside of the 1152 word vocabulary. For some blocks, the participant indicated that he forgot one of his intended responses when we asked him to report the intended response after the block concluded. Because there was no ground truth for this conversational task condition, we were unable to use the trial for analysis. Of 39 original conversational sentence-spelling trials, the participant got lost on 2 trials, tried to use an out-of-vocabulary word during 6 trials, and forgot the ground-truth sentence during 3 trials (leaving 28 trials for performance evaluation). Incorporating blocks where the participant used intended words outside of the vocabulary only modestly raised CER and WER to median values of 15.7% (99% CI [6.25, 30.4]) and 17.6%, (99% CI [12.5, 45.5]) respectively.

2.6 Acknowledgements

We are indebted to our participant Bravo-1 for his tireless dedication to the research project. We also thank members of Karunesh Ganguly's lab for help with the clinical study, Todd Dubnicoff for video editing, Kenneth Probst for illustrations, Nick Halper and Kian Torab for hardware support, members of the Chang Lab for feedback, Viv Her and Clarence Pang for administrative support, and the participant's caregivers for logistic support. For this work, the National Institutes of Health (grant NIH U01 DC018671-01A1) and William K. Bowes, Jr. Foundation supported authors S.L.M., J.R.L., D.A.M., M.E.D., M.P.S., K.T.L., J.C., G.K.A., and E.F.C. Authors A.T.C. and K.G. did not have relevant funding for this work.

2.7 Author contributions

S.L.M. designed and trained the neural classifier, developed real-time classification, language-modeling, and beam-search approaches and software, and developed the offline classification, spelling-simulation, and neural-feature analyses. J.R.L. designed and trained the real-time speech detection model, performed nearest-class distance and evoked-signal analyses, performed statistical assessments, and contributed to the neural-feature analyses. D.A.M. managed and coordinated the research project and designed and implemented the real-time software infrastructure used to collect data and enable real-time sentence spelling. SLM. and J.R.L. generated figures. S.L.M., J.R.L., and D.A.M. designed the spelling process. D.A.M. and M.E.D. designed the graphical user interface for the spelling process. S.L.M., J.R.L., D.A.M., and E.F.C. prepared the manuscript with input from other authors. S.L.M., J.R.L., D.A.M., M.E.D., M.P.S., K.T.L., and J.C. helped collect the data, and, along with G.K.A., were involved in methodological design. M.P.S., A.T.C., K.G., and E.F.C. performed regulatory and clinical supervision. E.F.C. conceived, designed, and supervised the clinical trial.

2.8 Competing interests

S.L.M., J.R.L., D.A.M., and E.F.C. are inventors on a pending provisional patent application that is directly relevant to the neural-decoding approach used in this work. G.K.A and E.F.C are inventors on patent application PCT/US2020/028926, D.A.M. and E.F.C. are

inventors on patent application PCT/US2020/043706 and E.F.C. is an inventor on patent US9905239B2 which are broadly relevant to the neural-decoding approach in this work. The remaining authors declare no competing interests.

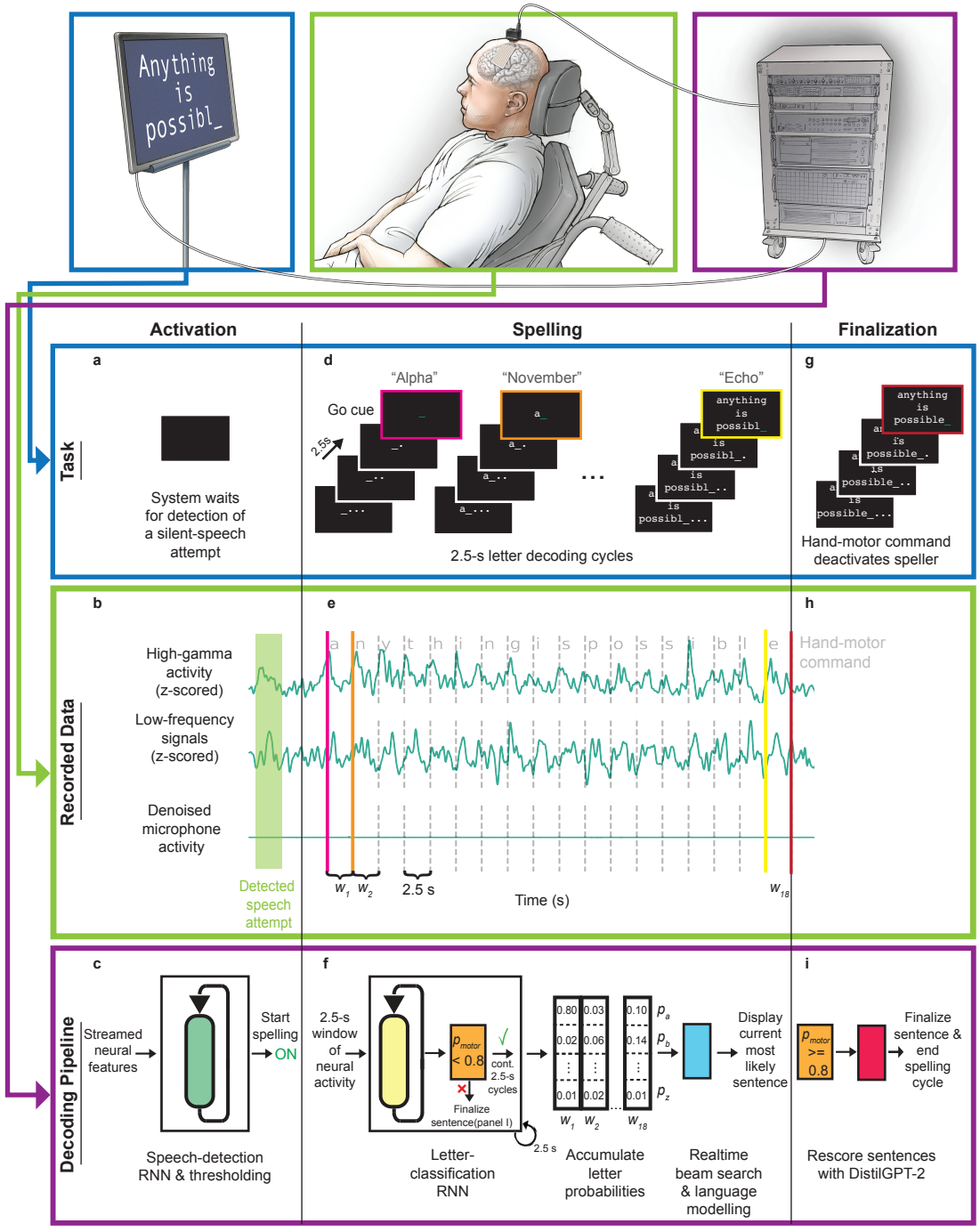


Figure 2.1. Schematic depiction of the spelling pipeline. (continued on next page).

(Previous page.) **Figure 2.1. Schematic depiction of the spelling pipeline.** **a** At the start of a sentence-spelling trial, the participant attempts to silently say a word to volitionally activate the speller. **b** Neural features (high-gamma activity and low-frequency signals) are extracted in real time from the recorded cortical data throughout the task. The features from a single electrode (electrode 0, Figure 2.5a) are depicted. For visualization, the traces were smoothed with a Gaussian kernel with a standard deviation of 150 milliseconds. The microphone signal shows that there is no vocal output during the task. **c** The speech-detection model, consisting of a recurrent neural network (RNN) and thresholding operations, processes the neural features to detect a silent-speech attempt. Once an attempt is detected, the spelling procedure begins. **d** During the spelling procedure, the participant spells out the intended message throughout letter-decoding cycles that occur every 2.5s. Each cycle, the participant is visually presented with a countdown and eventually a go cue. At the go cue, the participant attempts to silently say the code word representing the desired letter. **e** High-gamma activity and low-frequency signals are computed throughout the spelling procedure for all electrode channels and parceled into 2.5-s non-overlapping time windows. **f** An RNN-based letter-classification model processes each of these neural time windows to predict the probability that the participant was attempting to silently say each of the 26 possible code words or attempting to perform a hand-motor command (**g**). Prediction of the hand-motor command with at least 80% probability ends the spelling procedure (**i**). Otherwise, the predicted letter probabilities are processed by a beam-search algorithm in real time and the most likely sentence is displayed to the participant. **g** After the participant spells out his intended message, he attempts to squeeze his right hand to end the spelling procedure and finalize the sentence. **h** The neural time window associated with the hand-motor command is passed to the classification model. **i** If the classifier confirms that the participant attempted the hand-motor command, a neural network-based language model (DistilGPT-2) rescores valid sentences. The most likely sentence after rescoring is used as the final prediction.

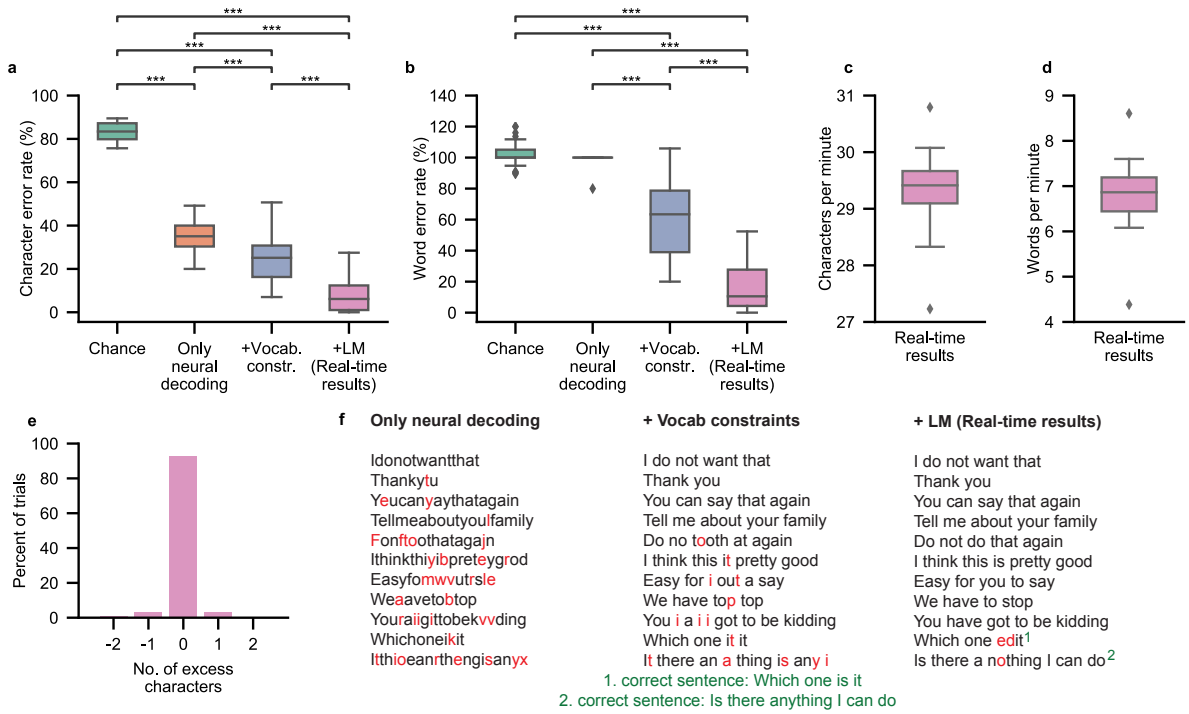


Figure 2.2. Performance summary of the spelling system during the copy-typing task. **a** Character error rates (CERs) observed during real-time sentence spelling with a language model (LM), denoted as '+LM (Real-time results)', and offline simulations in which portions of the system were omitted. In the 'Chance' condition, sentences were created by replacing the outputs from the neural classifier with randomly generated letter probabilities without altering the remainder of the pipeline. In the 'Only neural decoding' condition, sentences were created by concatenating together the most likely character from each of the classifier's predictions during a sentence trial (no whitespace characters were included). In the '+Vocab. constraints' condition, the predicted letter probabilities from the neural classifier were used with a beam search that constrained the predicted character sequences to form words within the 1152-word vocabulary. The final condition '+ LM (Real-time results)' incorporates language modeling. The sentences decoded with the full system in real time exhibited lower CERs than sentences decoded in the other conditions (***P< 0.0001, P-values provided in Table 2.2, two-sided Wilcoxon Rank-Sum test with 6-way Holm-Bonferroni correction). **b** Word error rates (WERs) for real-time results and corresponding offline omission simulations from A (***P< 0.0001, P-values provided in Table 2.3, two-sided Wilcoxon Rank-Sum test with 6-way Holm-Bonferroni correction). **c** The decoded characters per minute during real-time testing. **d** The decoded words per minute during real-time testing. In **a**–**d**, the distribution depicted in each boxplot was computed across n=34 real-time blocks (in each block, the participant attempted to spell between 2 and 5 sentences), and each boxplot depicts the median as a center line, quartiles as bottom and top box edges, and the minimum and maximum values as whiskers (except for data points that are 1.5 times the interquartile range, which are individually plotted). **e** Number of excess characters in each decoded sentence. **f** Example sentence-spelling trials with decoded sentences from each non-chance condition. Incorrect letters are colored red. Superscripts 1 and 2 denote the correct target sentence for the two decoded sentences with errors. All other example sentences did not contain any errors.

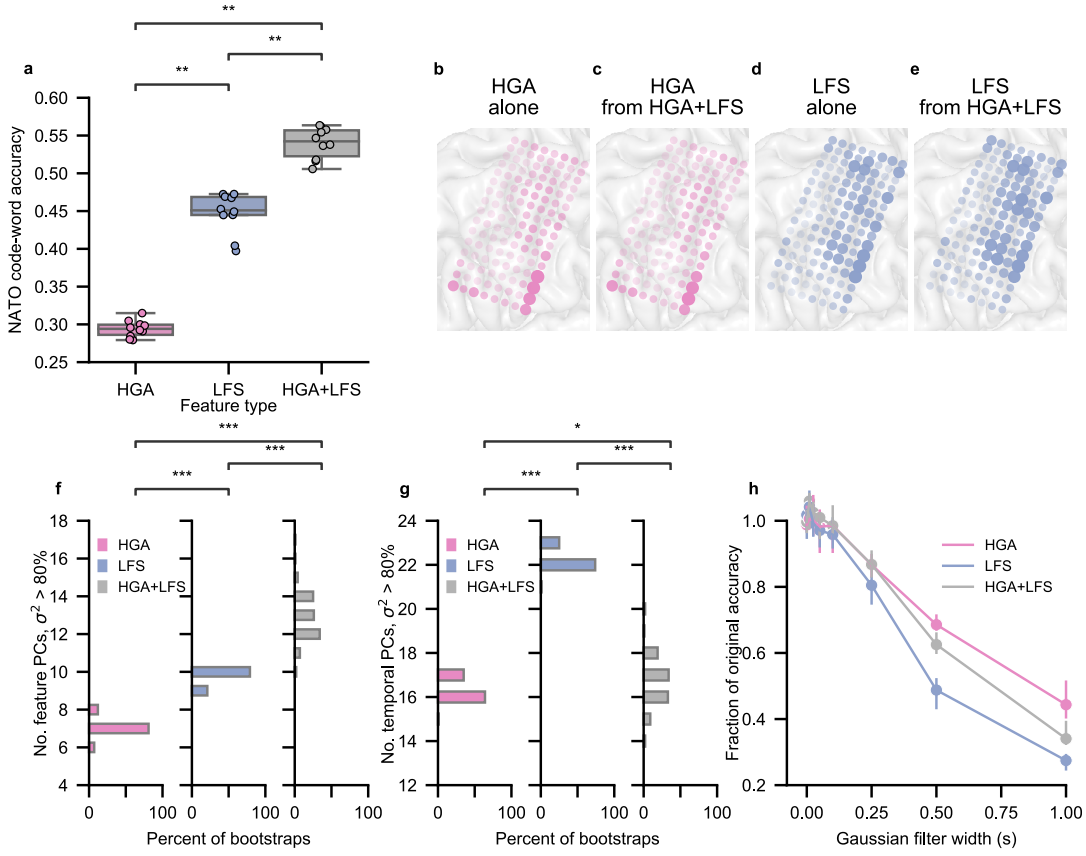


Figure 2.3. Characterization of high-gamma activity (HGA) and low-frequency signals (LFS) during silent-speech attempts. (continued on next page).

(Previous page.) **Figure 2.3. Characterization of high-gamma activity (HGA) and low-frequency signals (LFS) during silent-speech attempts. a** 10-fold cross-validated classification accuracy on silently attempted NATO code words when using HGA alone, LFS alone, and both HGA+LFS simultaneously. Classification accuracy using only LFS is significantly higher than using only HGA, and using both HGA+LFS results in significantly higher accuracy than either feature type alone (**P=4.71×10^{-4}, z=3.78 for each comparison, two-sided Wilcoxon Rank-Sum test with 3-way Holm-Bonferroni correction). Chance accuracy is 3.7%. Each boxplot corresponds to n = 10 cross-validation folds (which are also plotted as dots) and depicts the median as a center line, quartiles as bottom and top box edges, and the minimum and maximum values as whiskers (except for data points that are 1.5 times the interquartile range). **b**–**e** Electrode contributions. Electrodes that appear larger and more opaque provide more important features to the classification model. **b**, **c** Show contributions associated with HGA features using a model trained on HGA alone (**b**) vs using the combined LFS+HGA feature set (**c**). **d**, **e** depict contributions associated with LFS features using a model trained on LFS alone (**d**) vs the combined LFS+HGA feature set (**e**). **f** Histogram of the minimum number of principal components (PCs) required to explain more than 80% of the total variance, denoted as σ^2, in the spatial dimension for each feature set over 100 bootstrap iterations. The number of PCs required were significantly different for each feature set (***P< 0.0001, P-values provided in Table 2.5, two-sided Wilcoxon Rank-Sum test with 3-way Holm-Bonferroni correction). **g** Histogram of the minimum number of PCs required to explain more than 80% of the variance in the temporal dimension for each feature set over 100 bootstrap iterations (***P< 0.0001, P-values provided in Table 2.6, two-sided Wilcoxon Rank-Sum test with 3-way Holm-Bonferroni correction, *P< 0.01 two-sided Wilcoxon Rank-Sum test with 3-way Holm-Bonferroni correction). **h** Effect of temporal smoothing on classification accuracy. Each point represents the median, and error bars represent the 99% confidence interval around bootstrapped estimations of the median.

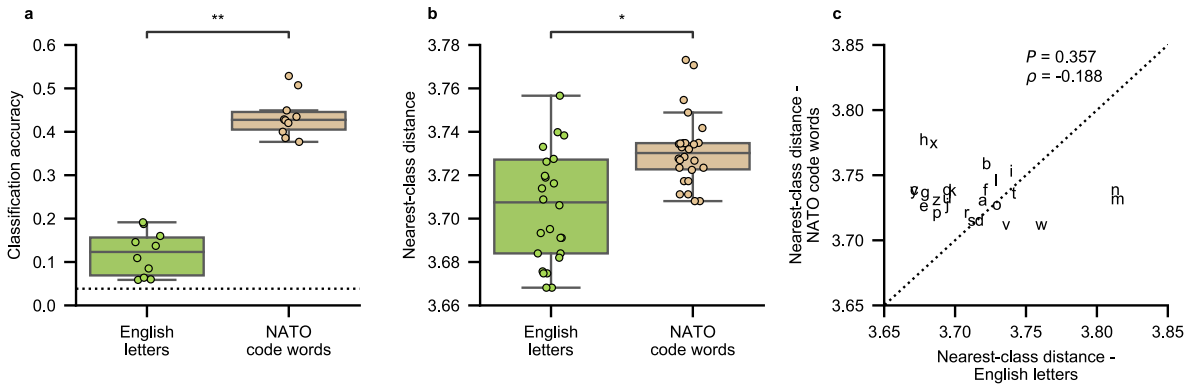


Figure 2.4. Comparison of neural signals during attempts to silently say English letters and NATO code words. **a** Classification accuracy (across n=10 cross-validation folds) using models trained with HGA+LFS features is significantly higher for NATO code words than for English letters (**P=1.57 $\times$ 10^{-4}, z=3.78, two-sided Wilcoxon Rank-Sum test). The dotted horizontal line represents chance accuracy. **b** Nearest-class distance is significantly larger for NATO code words than for letters (boxplots show values across the n = 26 code words or letters; *P=2.85 $\times$ 10^{-3}, z=2.98, two-sided Wilcoxon Rank-Sum test). In **a**, **b**, each data point is plotted as a dot, and each boxplot depicts the median as a center line, quartiles as bottom and top box edges, and the minimum and maximum values as whiskers (except for data points that are 1.5 times the interquartile range). **c** The nearest-class distance is greater for the majority of code words than for the corresponding letters. In **b** and **c**, nearest-class distances are computed as the Frobenius norm between trial-averaged HGA+LFS features.

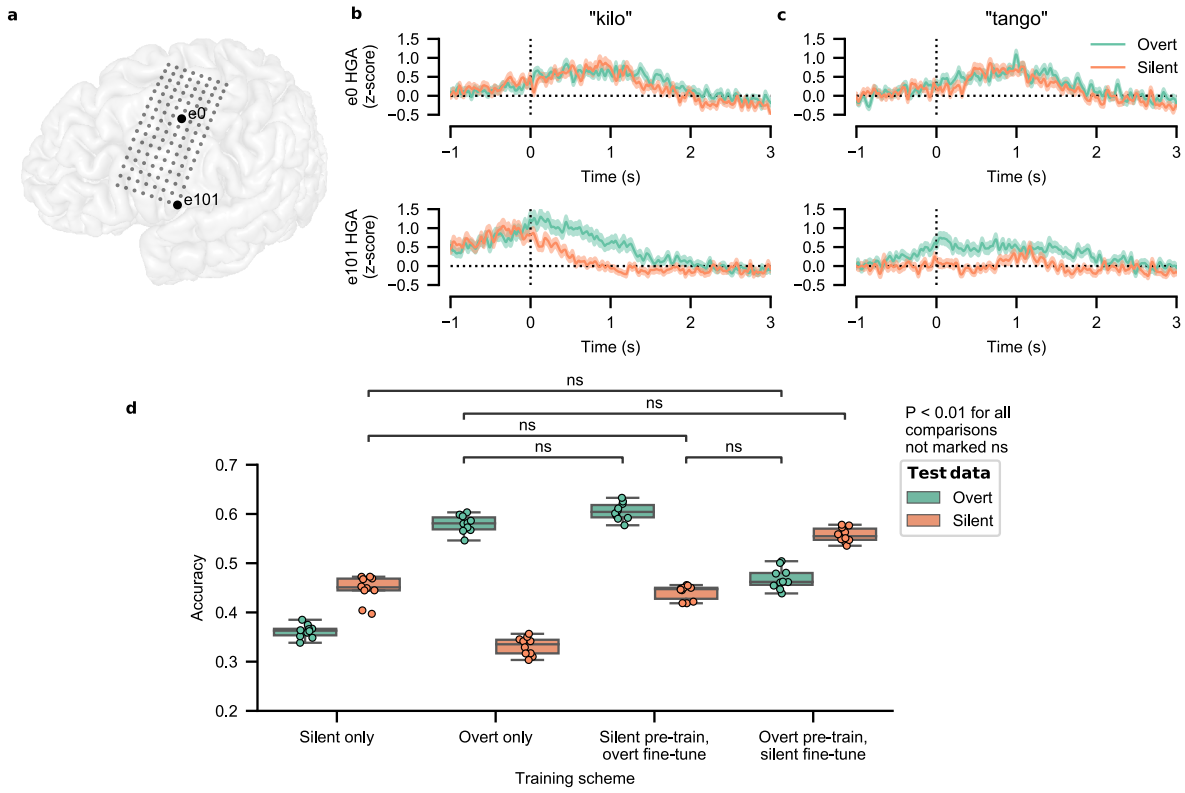


Figure 2.5. Differences in neural signals and classification performance between overt- and silent-speech attempts. a MRI reconstruction of the participant's brain overlaid with implanted electrode locations. The locations of the electrodes used in **b** and **c** are bolded and numbered in the overlay. **b** Evoked high-gamma activity (HGA) during silent (orange) and overt (green) attempts to say the NATO code word kilo. **c** Evoked high-gamma activity (HGA) during silent (orange) and overt (green) attempts to say the NATO code word tango. Evoked responses in **b** and **c** are aligned to the go cue, which is marked as a vertical dashed line at time 0. Each curve depicts the mean±standard error across n=100 speech attempts. **d** Codeword classification accuracy for silent- and overt-speech attempts with various model-training schemes. All comparisons revealed significant differences between the result pairs ($P < 0.01$, two-sided Wilcoxon Rank-Sum test with 28-way Holm-Bonferroni correction) except for those marked as 'ns'. Each boxplot corresponds to n = 10 cross-validation folds (which are also plotted as dots) and depicts the median as a center line, quartiles as bottom and top box edges, and the minimum and maximum values as whiskers (except for data points that are 1.5 times the interquartile range). Chance accuracy is 3.84%.

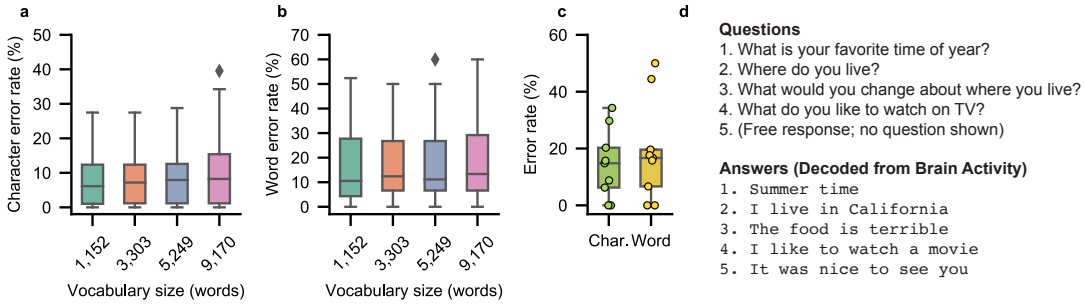


Figure 2.6. The spelling approach can generalize to larger vocabularies and conversational settings. **a** Simulated character error rates from the copy-typing task with different vocabularies, including the original vocabulary used during real-time decoding. **b** Word error rates from the corresponding simulations in **a**. In **a** and **b**, each boxplot corresponds to n=34 blocks (in each of these blocks, the participant attempted to spell between two to five sentences). **c** Character and word error rates across the volitionally chosen responses and messages decoded in real time during the conversational task condition. Each boxplot corresponds to n=9 blocks (in each of these blocks, the participant attempted to spell between two to four conversational responses; each dot corresponds to a single block). In **a**-**c**, each boxplot depicts the median as a center line, quartiles as bottom and top box edges, and the minimum and maximum values as whiskers (except for data points that are 1.5 times the interquartile range, which are individually plotted). **d** Examples of presented questions from trials of the conversational task condition (left) along with corresponding responses decoded from the participant's brain activity (right). In the final example, the participant spelled out his intended message without being prompted with a question.

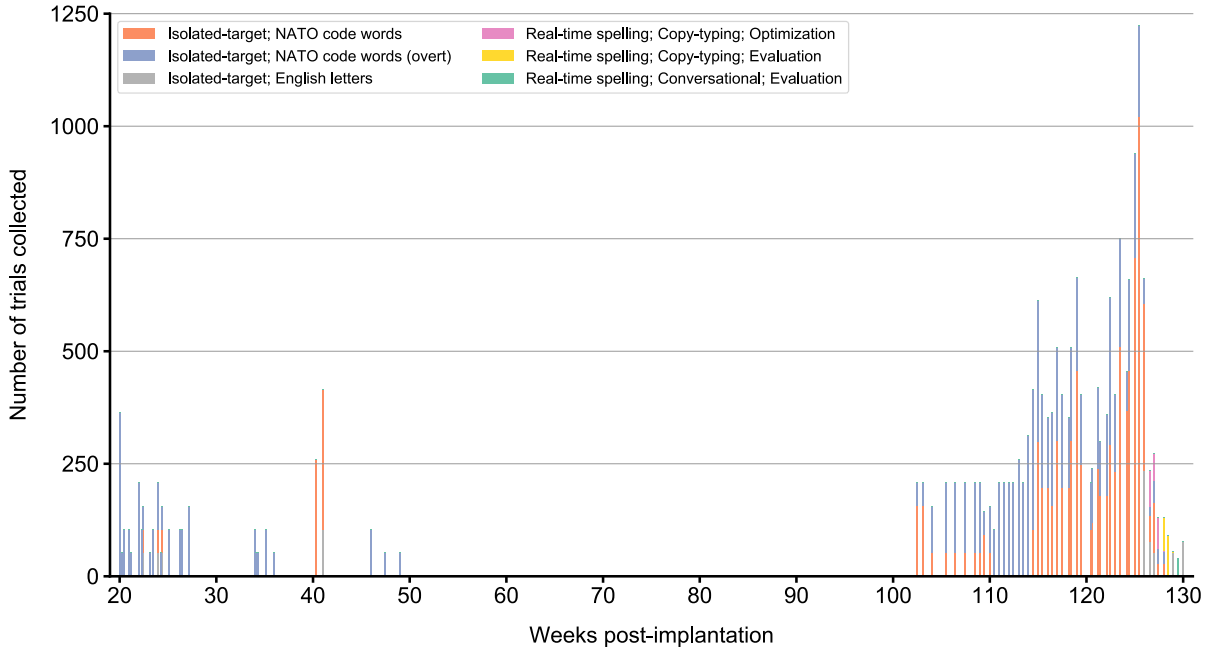


Figure 2.7. Data collection timeline Each bar depicts the total number of trials collected on each day of recording. The participant and implant date are the same as in our previous work (David A. Moses, Metzger, et al. 2021). If more than one type of dataset was collected in a single day, the bar is colored by the proportion of each dataset collected. Each color represents a specific dataset (as specified in the legend). Datasets vary in task type (isolated-target or real-time sentence spelling), utterance set (English letters, NATO code words (which included the attempted hand squeeze), copy-typing sentences, or conversational sentences), and, for the real-time sentence-spelling datasets, the purpose of the data (for hyperparameter optimization or for performance evaluation). All speech-related trials were associated with silent-speech attempts, except for the dataset with "(overt)" in its legend label. Additionally, 3.06% of trials in this overt dataset were actually recorded during a version of the copy-typing sentence-spelling task in which the participant attempted to overtly produce the code words. Datasets were collected on an irregular schedule due to external and clinical time constraints that were unrelated to the neural implant. The gap from 55–88 weeks was specifically due to clinical guidelines during the start of the COVID-19 pandemic that limited or prevented in-person recording sessions.

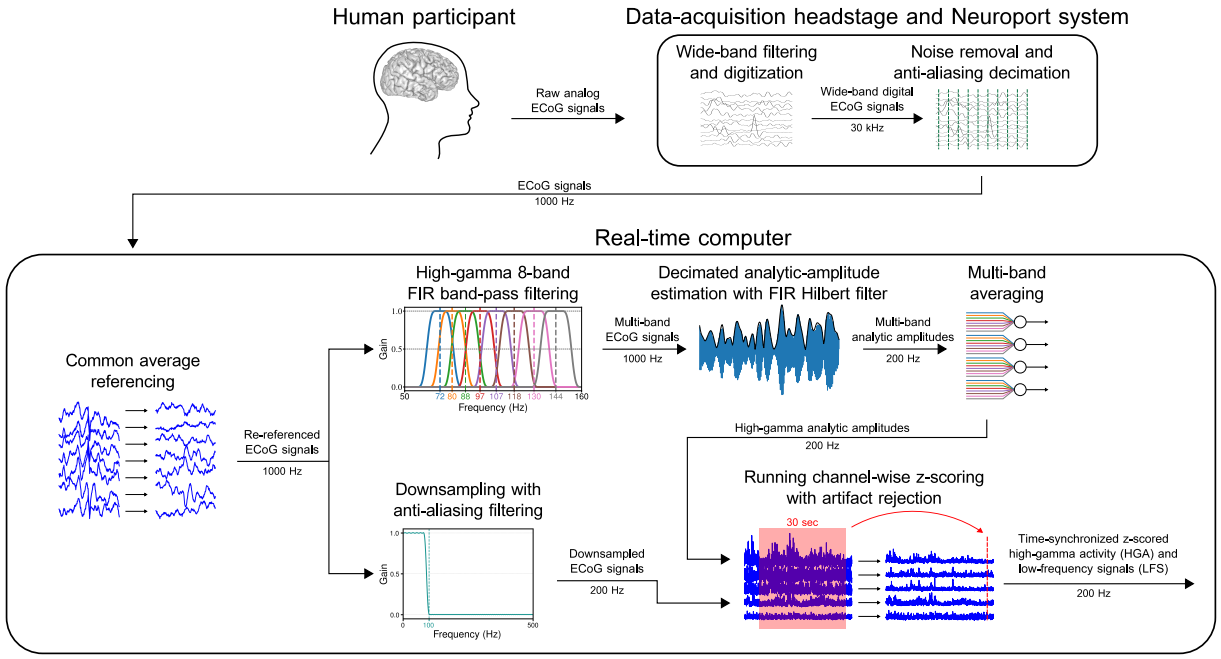


Figure 2.8. Real-time signal-processing pipeline A detachable data-acquisition headstage (NeuroPlex E, Blackrock Microsystems) attached to the percutaneous pedestal connector applied a hardware-based wide-band Butterworth filter (between 0.3 Hz and 7.5 kHz) to the ECoG signals, digitized them with 16-bit, 250-nV per bit resolution, and transmitted them at 30 kHz through additional connections to a Neuroport system (Blackrock Microsystems), which processed the signals using software-based line noise cancellation and an anti-aliasing low-pass filter (at 500 Hz). Afterwards, the processed signals were streamed at 1 kHz to a separate computer for further real-time processing and analysis, where we applied a common average reference (across all electrode channels) to each time sample of the ECoG data. The re-referenced signals were then processed in two parallel streams to extract high-gamma activity (HGA) and low-frequency signal (LFS) features. To compute the HGA features, we applied eight 390$^{\text{th}}$-order band-pass finite impulse response (FIR) filters to the re-referenced signals (filter center frequencies were within the high-gamma band at 72.0, 79.5, 87.8, 96.9, 107.0, 118.1, 130.4, and 144.0 Hz). Then, for each channel and band, we used a 170$^{\text{th}}$-order FIR filter to approximate the Hilbert transform. Specifically, for each channel and band, we set the real component of the analytic signal equal to the original signal delayed by 85 samples (half of the filter order) and set the imaginary component equal to the Hilbert transform of the original signal (approximated by this FIR filter) (Romero et al. 2012). We then computed the magnitude of each analytic signal at every fifth time sample, yielding analytic amplitude signals at 200 Hz. For each channel, we averaged the analytic amplitude values across the eight bands at each time point to obtain a single high-gamma analytic amplitude measure for that channel. To compute the LFS features, we downsampled the re-referenced signals to 200 Hz after applying a 130$^{\text{th}}$-order anti-aliasing low-pass FIR filter with a cutoff frequency of 100 Hz. We then combined the time-synchronized values from the two feature streams (high-gamma analytic amplitudes and downsampled signals) into a single feature stream. Next, we z-scored the values for each channel and each feature type using Welford's method with a 30-second sliding window (Welford 1962). Finally, we implemented a simple artifact-rejection approach to prevent samples with uncommonly large z-score magnitudes from interfering with the running z-score statistics or downstream decoding processes. We adapted this figure from our previous works (David A. Moses, Leonard, et al. 2019; David A. Moses, Metzger, et al. 2021), which implemented similar preprocessing pipelines to compute high-gamma features.

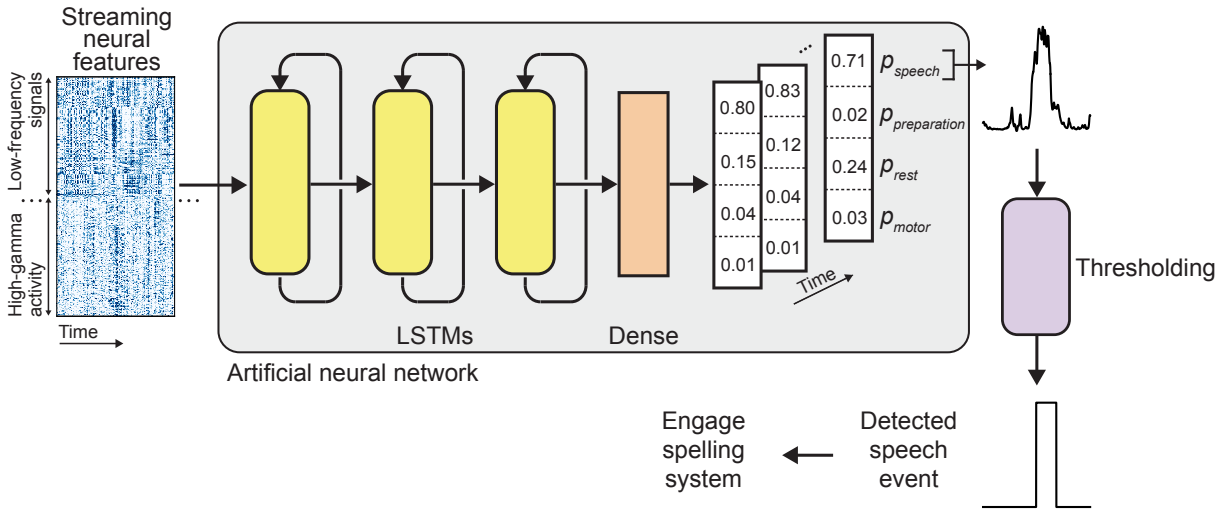


Figure 2.9. Speech-detection model schematic To detect silent-speech attempts from the participant's neural activity during real-time sentence spelling, first the z-scored low-frequency signals (LFS) and high-gamma activity (HGA) for each electrode are processed continuously by a stack of 3 long short-term memory (LSTM) layers. Next, a single dense (fully connected) layer projects the latent dimensions of the final LSTM onto the 4 possible classes: speech, speech preparation, rest, and motor. The stream of speech probabilities is then temporally smoothed, probability thresholded, and time thresholded to yield onsets and offsets of full speech events. Once the participant attempts to silently say something and that speech attempt is detected, the spelling system is engaged and the paced spelling procedure begins. The depicted neural features, predicted speech-probability time series (upper right), and detected speech event (lower right) are the actual neural data and detection results for a 5-second time window at the beginning of a trial of the real-time sentence copy-typing task. This figure was adapted from our previous work (David A. Moses, Metzger, et al. 2021), which implemented a similar speech-detection architecture.

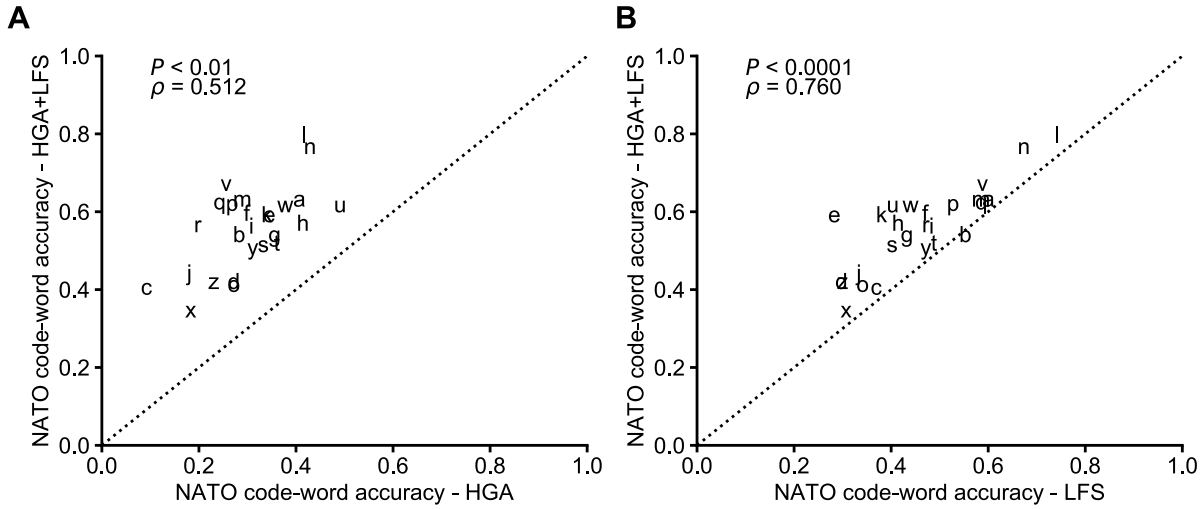


Figure 2.10. Effects of feature selection on code-word classification accuracy **A.** Classification accuracy improves for each code word when using high-gamma activity (HGA) and low-frequency signals (LFS) together (the combined HGA+LFS feature set) instead of only HGA features. The accuracies are significantly correlated with a Spearman rank correlation of 0.512 ($P = 0.0085$, permutation testing with 2000 iterations). **B.** Classification accuracy improves for almost every code word when using HGA+LFS instead of LFS alone. The accuracies are significantly correlated with a Spearman rank correlation of 0.760 ($P \approx 0.00$, permutation testing with 2000 iterations). Because not all possible permutations were tested (the number of possible permutations for 26 elements is 4.03×10^{26}, so we approximate this test with 2000 iterations), the P-value is approximately 0.00 in this case. In both **A** and **B**, code words are represented as lower-case letters and the Spearman rank correlations are shown. The associated P-value was computed via permutation testing. In permutation testing, one group of observations (code-word accuracies for either HGA, LFS, or HGA+LFS) was shuffled before re-computing the correlation between that group of observations and the other group. 2000 iterations were used during permutation testing for each of the two comparisons. The P-value was computed as the proportion of the distribution of correlations computed during permutation testing that were greater in magnitude than the correlation computed on non-shuffled data.

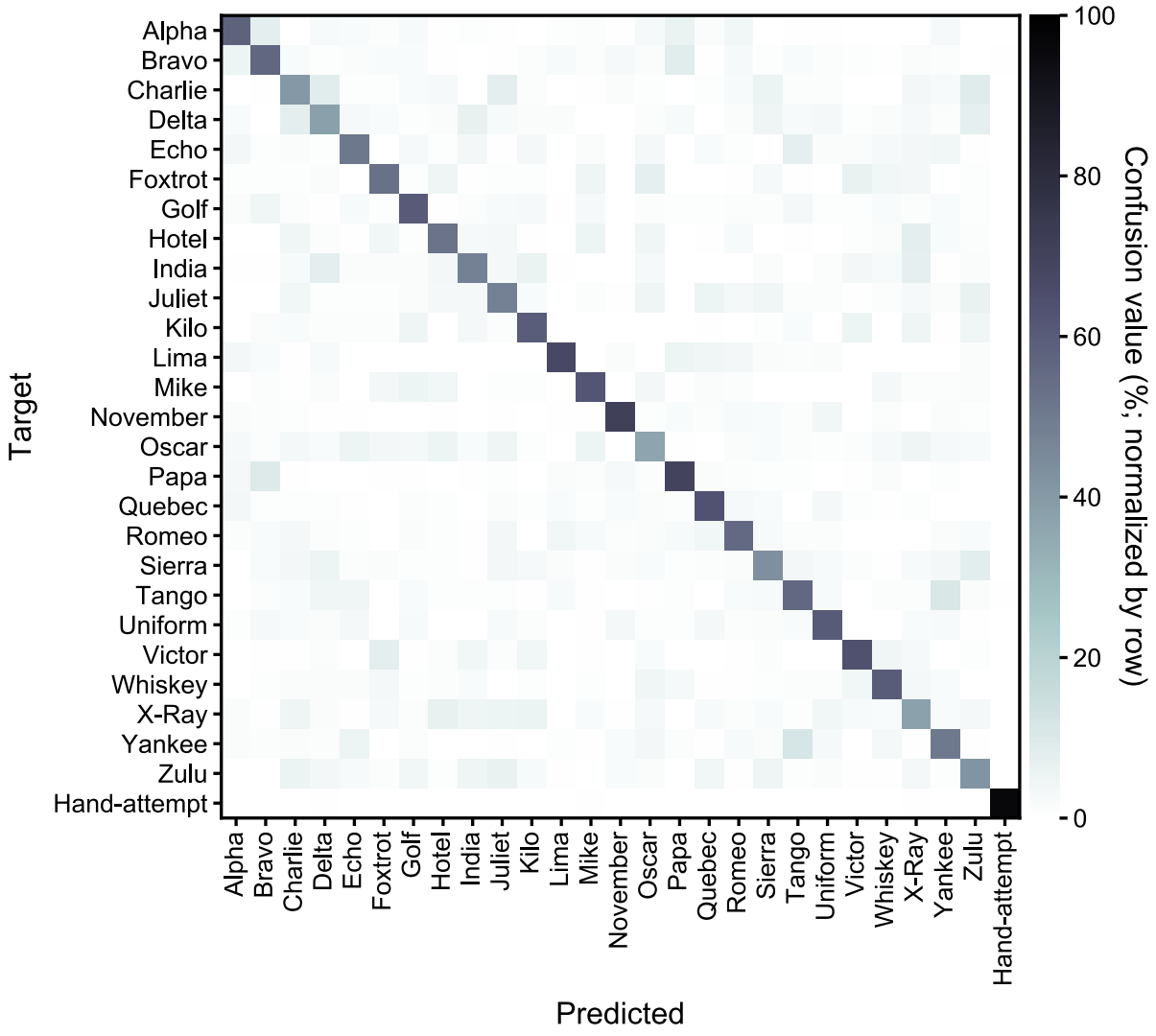


Figure 2.11. Confusion matrix from isolated-target trial classification using HGA and LFS Confusion values, computed during offline classification of neural data (using both high-gamma activity and low-frequency signals) recorded during isolated-target trials, are shown for each NATO code word and the attempted hand squeeze. Each row corresponds to a target code word or the attempted hand squeeze, and the value in each column for that row corresponds to the percent of isolated-target task trials that were correctly classified as the target (if the value is along the diagonal) or misclassified ("confused") as another potential target (if the value is not along the diagonal). The values in each row sum to 100%. In general, silent-speech and hand-squeeze attempts were reliably classified. Including both the attempted NATO code word trials and the attempted hand squeeze trials, the 10-fold cross-validated median accuracy was 56.4% with a 99% confidence interval of $[54.3, 58.2]$.

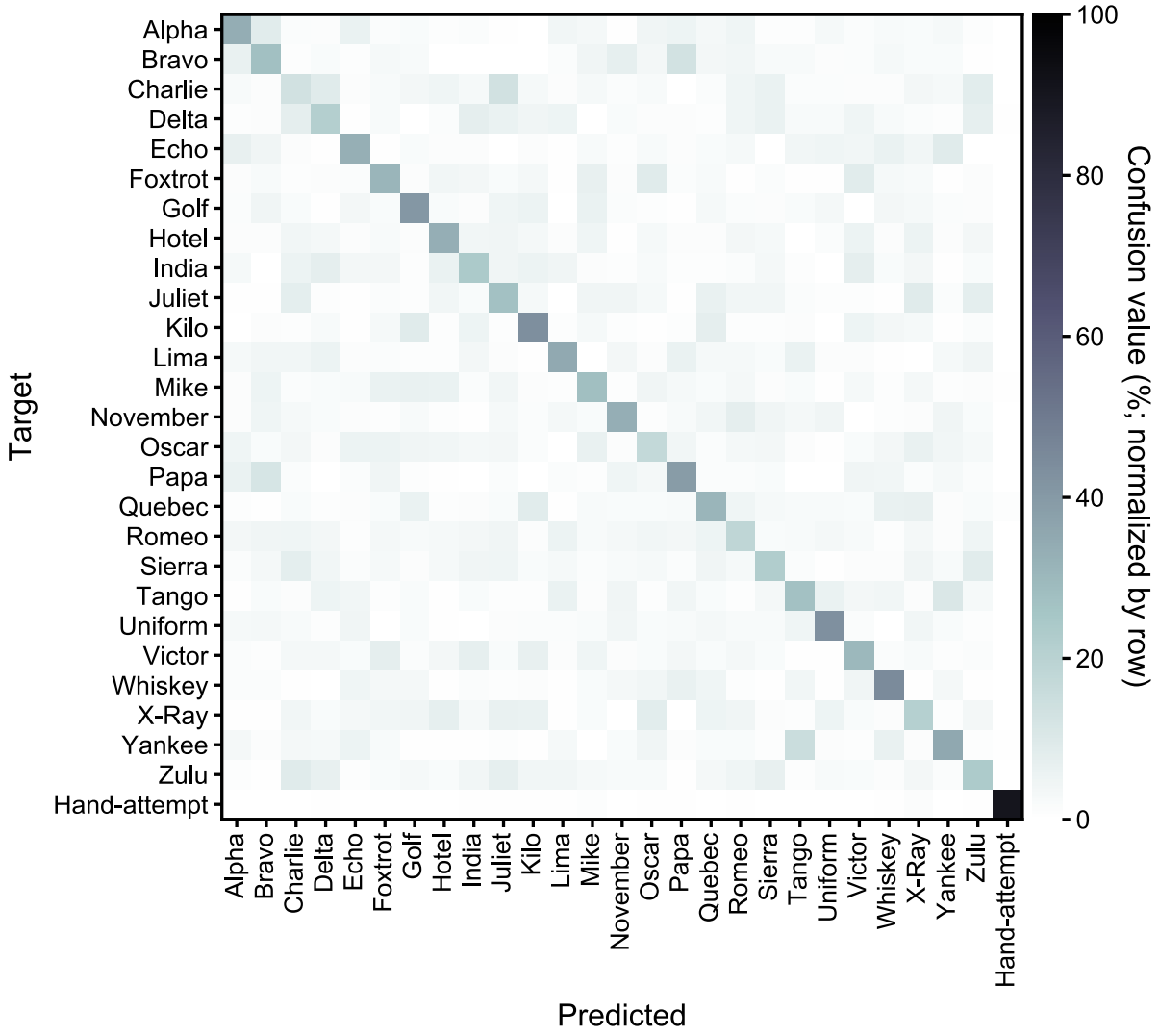


Figure 2.12. Confusion matrix from isolated-target trial classification using only HGA Confusion values, computed during offline classification of neural data (using only high-gamma activity) recorded during isolated-target trials, are shown for each NATO code word and the attempted hand squeeze. Each row corresponds to a target code word or the attempted hand squeeze, and the value in each column for that row corresponds to the percent of isolated-target task trials that were correctly classified as the target (if the value is along the diagonal) or misclassified ("confused") as another potential target (if the value is not along the diagonal). The values in each row sum to 100%. Including both the attempted NATO code word trials and the attempted hand squeeze trials, the 10-fold cross-validated median accuracy was 32.7% with a 99% confidence interval of $[32.0, 33.6]$.

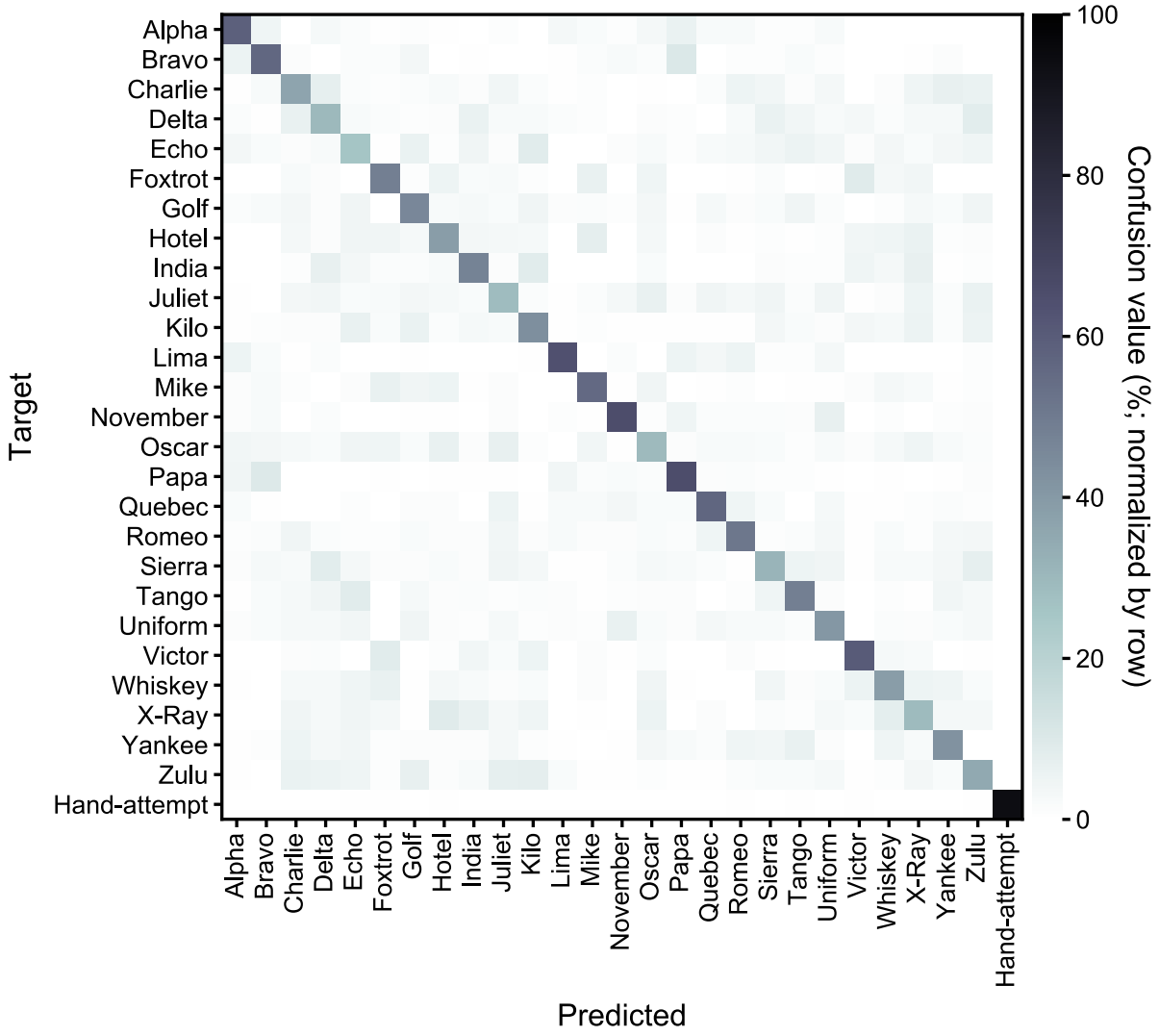


Figure 2.13. Confusion matrix from isolated-target trial classification using only LFS Confusion values, computed during offline classification of neural data (using only low-frequency signals) recorded during isolated-target trials, are shown for each NATO code word and the attempted hand squeeze. Each row corresponds to a target code word or the attempted hand squeeze, and the value in each column for that row corresponds to the percent of isolated-target task trials that were correctly classified as the target (if the value is along the diagonal) or misclassified ("confused") as another potential target (if the value is not along the diagonal). The values in each row sum to 100%. Including both the attempted NATO code word trials and the attempted hand squeeze trials, the 10-fold cross-validated median accuracy was 48.2% with a 99% confidence interval of $[42.9, 49.7]$.

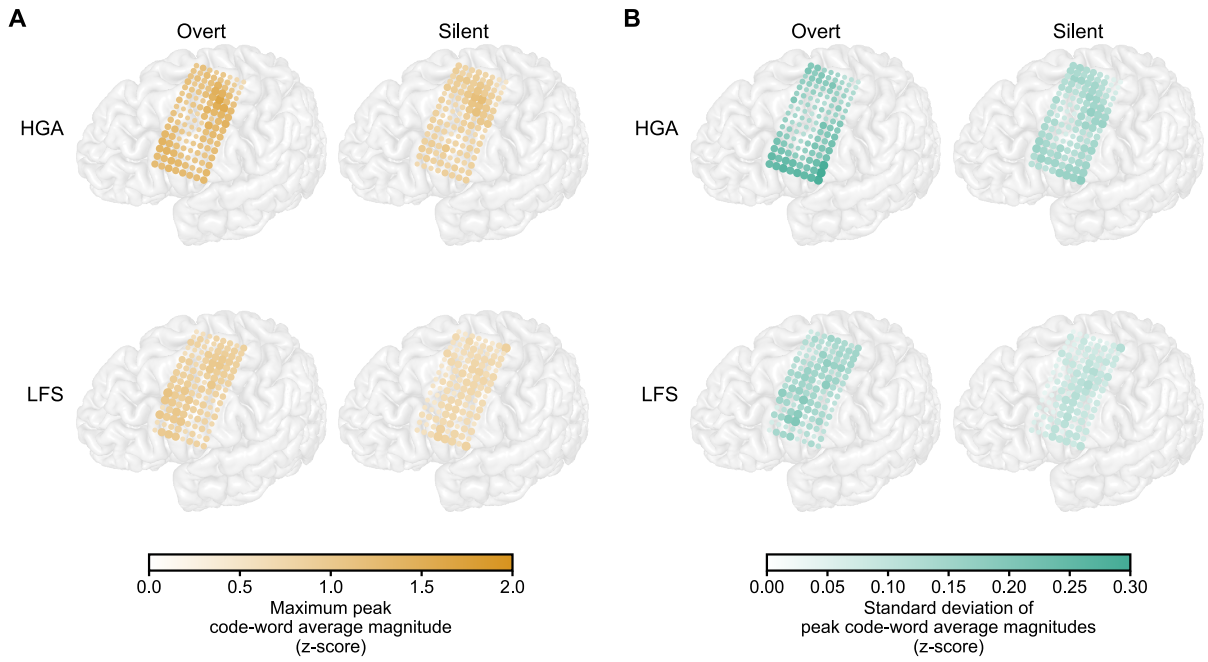


Figure 2.14. Neural-activation statistics during overt- and silent-speech attempts **A.** Each image shows an MRI reconstruction of the participant's brain overlaid with electrode locations and the maximum neural activations for each electrode, type of speech attempt (overt or silent), and feature type (high-gamma activity (HGA) or low-frequency signals (LFS)), measured as maximum peak code-word average magnitudes. To calculate these values, the trial-averaged neural-feature time series was computed for each code word, electrode, type of speech attempt, and feature type using the isolated-target dataset (for each trial, the 2.5-second time window after the go cue was used). Then, the peak magnitude (maximum of the absolute value) of each of these trial-averaged time series was determined. The maximum peak code-word average magnitude for each electrode, type of speech attempt, and feature type was then computed as the maximum value of these peak magnitudes across code words for each combination. The two columns show the values for each type of speech attempt (overt then silent), and the two rows show the values for each feature type (HGA then LFS). **B.** The standard deviation of peak code-word average magnitudes. Here, the standard deviation (instead of the maximum used in **A**) of the peak average magnitudes across the code words for each electrode, type of speech attempt, and feature type is computed and plotted, depicting how much the magnitudes varied across speech targets for that combination. For **A** and **B**, the color of each plotted electrode indicates the true associated value for that electrode, and the size of each electrode depicts the associated value for that electrode relative to the values for the other electrodes (for a given type of speech attempt and feature type).

Table 2.1. Copy-typing task sentences.

Target sentence	Decoded sentence in first trial	Decoded sentence in second trial
good morning	good morning	good for legs
you have got to be kidding	you have got to be kidding a	you have got to be kidding
what do you mean	what do you mean	what do you mean
good to see you	i do i leave you	good to see you
i think this is pretty good	i think this is pretty good	i think they is pretty good
i will check	i will check	i will the it
thank you	thank you	thank you
please sit down	please sit down	please believe
we have to stop	we have to stop	we have to stop
hand that to me please	hand that time please	have that time always
i know what you mean	i know what you mean	i know what you mean
what time is it	what time is it	what time is it
sit over here with me	sit over here with me	sit over here with me
no thanks	no thanks	not happen
you never know	you never know	you never know
great to see you again	great to show my case in	great to stay in town
forget about it	forget about it	forget about it
could you repeat what you said	dog lie on repeat what you said	could you repeat what you said
where do you live	where do you live	where do you live
do not be afraid to ask me questions	do not be afraid to ask me questions	do not be afraid to ask me questions
i cannot believe it	i can not believe it	i can not believe it
thanks for telling me	thank for reading me	thanks for telling me
i do not want that	i do not want that	i do not want that
that is wonderful	that is work from a	that is wonderful
what do you think about that	what do you think about that	what do you think about that
thank you very much	though it very much	thank you very much
i am glad you are here	i am glad you are here	i am glad you are here
how are you doing	how are you doing	how are you doing
i agree	i agree	i agree
i am okay	i am okay	i am okay
tell me what you are doing	tell me what your telling	tell me what you are doing
how long did it take	how long did it take	how long did it take
is there anything i can do	is there a nothing i can do	is there anything i can do
how are things going for you	how are things gives for you	how are things going for you
do you know what he did	do you know on the ice	do you know what he did
was there something else	was there to be a high else	was there something else
where are you going	while are you doing	where are you going
who is that	who is that	why is that
tell me about your family	tell me about your family	tell me about your family
i could probably do better	i could probably do better	i could probably do better
you can say that again	you can say that again	you can say that open
i am sorry to hear that	i am to get to hear that	i am sorry to hear that
will i see you later	will i see you later	well i keep by later
i am doing well	i am doing well	i am doing fine
can that wait until another time	can that wait until another time	can that wait until another time
how much more is there	how much more is there	how much were in there
come talk with me	come talk with me	some take with me
that will be fun	that will be fun	that will be fun
how often do you do this	how often do you do this	how often do you do this
how much will it cost	how much will it cost	how much will it cost
bring that over here	clinic hat for hat	bring that ever here
turn it off	turn it off	turn it off

(continued on next page).

(continued from previous page). Table 2.1. Copy-typing task sentences.

Target sentence	Decoded sentence in first trial	Decoded sentence in second trial
i remember the last time i did that	i remember the last time i did that	i remember to plan new me i did that
i was just kidding	i was mike kidding	i was just kidding
i will meet you there	i will meet you there	i will meet you to eat
i do not really remember	i do not really remember	ddonoyrballyrlrefbhrh
i feel cold	i feel weird	i feel cold
excuse me for interrupting	excuse me for interrupt any	excuse me for interrupting
you are not going to believe this	you plan to go in on a bit love this	ypuaranpdggingloavlinesoeb
do you understand what i mean	do you understand what i mean	do you understand what i mean
what are you talking about	what are you talking about	what are you talking about
which one is it	which one edit	which one is it
would you like to go with me	a all i was like the white me	would you like to go with me
i do not understand	i do not understand	i do not understand
of course i do	of course its	of course him
anything is possible	anything is possible	anything is possible
do not do that again	do not do that again	do not do that again
let me see that	let me see that	let me see that
what have you been doing	what have you been doing	what have you been doing
i had a great time	i had a great time	what a great time
easy for you to say	easy for you to say	easy for you to say
i want to go	i want to go	i want to go
how do you feel	how do you feel	how do you feel
that is all right	that is all right	that is all right
i told you i do not know	i told you i do not know	i told you i do not know

Table 2.2. Statistical comparisons of character error rates across decoding-framework conditions.

Statistical comparison[1]	$\lvert z\text{-value} \rvert$	P-value (corrected)[2]
Chance vs. Only Neural Decoding	7.09	8.08×10^{-12}
Chance vs. + Vocab. Constraints	7.09	8.08×10^{-12}
Chance vs. + LM (Real-time results)	7.09	8.08×10^{-12}
Only Neural Decoding vs. + LM (Real-time results)	6.94	1.21×10^{-11}
+ Vocab. Constraints vs. + LM (Real-time results)	5.53	6.34×10^{-8}
Only Neural Decoding vs. + Vocab. Constraints	4.51	6.37×10^{-6}

[1] Each comparison is a two-sided Wilcoxon Rank-Sum test across 34 real-time spelling blocks.
[2] 6-way Holm-Bonferroni correction for multiple comparisons.

Table 2.3. Statistical comparisons of word error rates across decoding-framework conditions.

Statistical comparison[1]	$\lvert z\text{-value} \rvert$	P-value (corrected)[2]
Chance vs. + LM (Real-time results)	7.09	8.08×10^{-12}
Only Neural Decoding vs. + LM (Real-time results)	7.09	8.08×10^{-12}
Chance vs. + Vocab. Constraints	6.70	8.16×10^{-11}
Only Neural Decoding vs. + Vocab. Constraints	6.61	1.19×10^{-10}
+ Vocab. Constraints vs. + LM (Real-time results)	6.11	2.01×10^{-9}

[1] Each comparison is a two-sided Wilcoxon Rank-Sum test across 34 real-time spelling blocks.
[2] 6-way Holm-Bonferroni correction for multiple comparisons.

Table 2.4. Statistical comparisons of classification accuracy across neural-feature types.

Statistical comparison[1]	$\lvert z\text{-value} \rvert$	P-value (corrected)[2]
HGA vs. LFS	3.78	4.71×10^{-4}
HGA vs. HGA+LFS	3.78	4.71×10^{-4}
LFS vs. HGA+LFS	3.78	4.71×10^{-4}

[1] Each comparison is a two-sided Wilcoxon Rank-Sum test across 10 cross-validation folds.
[2] 6-way Holm-Bonferroni correction for multiple comparisons.

Table 2.5. Statistical comparisons of the number of principal components required to explain more than 80% of the variance in the spatial dimension across neural-feature types.

Statistical comparison[1]	$\lvert z\text{-value} \rvert$	P-value (corrected)[2]
HGA vs. LFS	12.22	7.57×10^{-34}
HGA vs. HGA+LFS	12.22	7.57×10^{-34}
LFS vs. HGA+LFS	12.02	2.66×10^{-33}

[1] Each comparison is a two-sided Wilcoxon Rank-Sum test across 100 bootstrap iterations.
[2] 3-way Holm-Bonferroni correction for multiple comparisons.

Table 2.6. Statistical comparisons of the number of principal components required to explain more than 80% of the variance in the temporal dimension across neural-feature types.

Statistical comparison[1]	$\lvert z\text{-value} \rvert$	P-value (corrected)[2]
HGA vs. LFS	12.22	7.57×10^{-34}
LFS vs. HGA+LFS	12.22	7.57×10^{-34}
HGA vs. HGA+LFS	2.68	0.007 27

[1] Each comparison is a two-sided Wilcoxon Rank-Sum test across 100 bootstrap iterations.
[2] 3-way Holm-Bonferroni correction for multiple comparisons.

Table 2.7. Statistical comparisons of classification accuracy across attempted-speech types with various training schemes.

Group 1 Train	Test	Group 2 Train	Test	| z-value |	P-value (corrected[2])
Silent	Silent	Silent	Overt	3.78	4.4×10^{-3}
Silent	Silent	Overt	Overt	3.78	4.4×10^{-3}
Silent	Silent	Overt	Silent	3.78	4.4×10^{-3}
Silent	Silent	Overt pre-train, silent fine-tune	Silent	3.78	4.4×10^{-3}
Silent	Silent	Silent pre-train, overt fine-tune	Overt	3.78	4.4×10^{-3}
Silent	Overt	Overt	Overt	3.78	4.4×10^{-3}
Silent	Overt	Overt pre-train, silent fine-tune	Silent	3.78	4.4×10^{-3}
Silent	Overt	Overt pre-train, silent fine-tune	Overt	3.78	4.4×10^{-3}
Silent	Overt	Silent pre-train, overt fine-tune	Silent	3.78	4.4×10^{-3}
Silent	Overt	Silent pre-train, overt fine-tune	Overt	3.78	4.4×10^{-3}
Overt	Overt	Overt	Silent	3.78	4.4×10^{-3}
Overt	Overt	Overt pre-train, silent fine-tune	Overt	3.78	4.4×10^{-3}
Overt	Overt	Silent pre-train, overt fine-tune	Silent	3.78	4.4×10^{-3}
Overt	Silent	Overt pre-train, silent fine-tune	Silent	3.78	4.4×10^{-3}
Overt	Silent	Overt pre-train, silent fine-tune	Overt	3.78	4.4×10^{-3}
Overt	Silent	Silent pre-train, overt fine-tune	Silent	3.78	4.4×10^{-3}
Overt	Silent	Silent pre-train, overt fine-tune	Overt	3.78	4.4×10^{-3}
Overt pre-train, silent fine-tune	Silent	Overt pre-train, silent fine-tune	Overt	3.78	4.4×10^{-3}
Overt pre-train, silent fine-tune	Silent	Silent pre-train, overt fine-tune	Silent	3.78	4.4×10^{-3}
Overt pre-train, silent fine-tune	Overt	Silent pre-train, overt fine-tune	Overt	3.78	4.4×10^{-3}
Silent pre-train, overt fine-tune	Silent	Silent pre-train, overt fine-tune	Overt	3.78	4.4×10^{-3}
Overt pre-train, silent fine-tune	Silent	Silent pre-train, overt fine-tune	Overt	3.70	4.4×10^{-3}
Silent	Overt	Overt	Silent	3.17	8.99×10^{-3}
Overt pre-train, silent fine-tune	Overt	Silent pre-train, overt fine-tune	Silent	2.76	2.9×10^{-2}
Overt	Overt	Silent pre-train, overt fine-tune	Overt	2.65	3.26×10^{-2}
Overt	Overt	Overt pre-train, silent fine-tune	Silent	2.57	3.26×10^{-2}
Silent	Silent	Overt pre-train, silent fine-tune	Overt	1.51	2.61×10^{-1}
Silent	Silent	Silent pre-train, overt fine-tune	Silent	0.76	4.5×10^{-1}

[1] Each comparison is a two-sided Wilcoxon Rank-Sum test across 10 cross-validation folds.
[2] 28-way Holm-Bonferroni correction for multiple comparisons.

Table 2.8. Hyperparameter definitions and values.

Model	Hyperparameter description	Search-space type[1]	Value range	Optimal values[2]
Speech detector	Smoothing size	Uniform (int)	$[1, 80]$	78
	Probability threshold	Uniform	$[0.1, 0.9]$	0.304
	Time threshold duration	Uniform (int)	$[25, 150]$	105
Word classifier	Number of GRU layers	Uniform (int)	$[1, 4]$	2
	Nodes per GRU layer	Uniform (int)	$[128, 512]$	274
	Dropout fraction	Uniform	$[0.3, 0.8]$	0.545
	Convolution kernel size and skip	Uniform (int)	$[1, 10]$	4
	Jitter amount (seconds), j	Uniform	$[0.0, 2.0]$	0.474
	Additive noise level, σ_n	Uniform	$[0.0, 1.0]$	0.0027
	Scale min., α_{min}	Uniform	$[0.8, 1.0]$	0.955
	Scale max., α_{max}	Uniform	$[1.0, 1.2]$	1.07
	Max. temporal-masking length (seconds), b	Uniform	$[0.00, 1.35]$	0.871
	Temporal masking probability, p	Uniform	$[0.0, 0.5]$	0.0478
	Channel-wise noise, σ_c	Uniform	$[0.0, 1.0]$	0.0283
Beam search	Language-model scaling factor, α	Uniform	$[0.01, 1.0]$	$(0.642, 0.744)$
	Word-insertion weight, β	Uniform	$[0.0, 30.0]$	$(4.03, 10.5)$
	Number of beams maintained, B	Uniform (int)	$[0, 750]$	$(457, 739)$
	Distil-GPT2 scaling factor, α_{gpt2}	Uniform	$[0.0, 100.0]$	$(1.53, 1.13)$

[1] "Uniform (int)" indicates that hyperparameter values were forced to be integers.

[2] For the language modeling and beam-search hyperparameters, two values are listed: the first is the optimal value found when optimizing on the copy-typing sentence-spelling trials prior to the first day of sentence-spelling evaluations (used during this first day), and the second is the optimal value found when optimizing on the copy-typing sentence-spelling trials from the first day of sentence-spelling evaluations (used for the second day and all subsequent days).

References

About the Oxford 3000 and 5000 word lists at Oxford Learner's Dictionaries (2021). URL: https://www.oxfordlearnersdictionaries.com/us/about/wordlists/oxford3000-5000 (visited on 10/19/2021).

Adolphs, Svenja and Schmitt (Dec. 1, 2003). "Lexical Coverage of Spoken Discourse". *Applied Linguistics* 24.4, pp. 425–438. ISSN: 0142-6001, 1477-450X. DOI: 10.1093/applin/24.4.425.

Angrick, Miguel, Maarten Ottenhoff, Sophocles Goulis, et al. (Nov. 2021). "Speech Synthesis from Stereotactic EEG using an Electrode Shaft Dependent Multi-Input Convolutional Neural Network Approach". *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*. 2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC). ISSN: 2694-0604, pp. 6045–6048. DOI: 10.1109/EMBC46164.2021.9629711.

Anumanchipalli, Gopala K., Josh Chartier, and Edward F. Chang (Apr. 2019). "Speech synthesis from neural decoding of spoken sentences". *Nature* 568.7753, pp. 493–498. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-019-1119-1.

Beukelman, David R., Susan Fager, Laura Ball, and Aimee Dietz (Jan. 2007). "AAC for adults with acquired neurological conditions: A review". *Augmentative and Alternative Communication* 23.3, pp. 230–242. ISSN: 0743-4618, 1477-3848. DOI: 10.1080/07434610701553668.

Bouchard, Kristofer E., Nima Mesgarani, Keith Johnson, and Edward F. Chang (Mar. 2013). "Functional organization of human sensorimotor cortex for speech articulation". *Nature* 495.7441, pp. 327–332. ISSN: 1476-4687 (Electronic)\r0028-0836 (Linking). DOI: 10.1038/nature11911.

Branco, Mariana P., Elmar G. M. Pels, Ruben H. Sars, et al. (Mar. 1, 2021). "Brain-Computer Interfaces for Communication: Preferences of Individuals With Locked-in Syndrome". *Neurorehabilitation and Neural Repair* 35.3. Publisher: SAGE Publications Inc STM, pp. 267–279. ISSN: 1545-9683. DOI: 10.1177/1545968321989331.

Brants, Thorsten and Alex Franz (Sept. 19, 2006). *Web 1T 5-gram Version 1*. Artwork Size: 20971520 KB Pages: 20971520 KB Type: dataset. DOI: 10.35111/CQPA-A498.

Brumberg, Jonathan S., Kevin M. Pitt, Alana Mantie-Kozlowski, and Jeremy D. Burnison (Feb. 6, 2018). "Brain–Computer Interfaces for Augmentative and Alternative Communication: A Tutorial". *American Journal of Speech-Language Pathology* 27.1, pp. 1–12. ISSN: 1058-0360, 1558-9110. DOI: 10.1044/2017_AJSLP-16-0244.

Carey, Daniel, Saloni Krishnan, Martina F. Callaghan, et al. (2017). "Functional and Quantitative MRI Mapping of Somatomotor Representations of Human Supralaryngeal Vocal Tract". *Cerebral cortex* 27.1, pp. 265–278. ISSN: 2076792171. DOI: 10.1093/cercor/bhw393.

Chartier, Josh, Gopala K. Anumanchipalli, Keith Johnson, and Edward F. Chang (2018). "Encoding of Articulatory Kinematic Trajectories in Human Speech Sensorimotor Cortex". *Neuron* 98.5, 1042–1054.e4. DOI: 10.1016/j.neuron.2018.04.031.

Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, et al. (Sept. 25, 2014). "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". Proceedings of the Empirical Methods in Natural Language Processing (EMNLP), pp. 1724–1734. DOI: http://dx.doi.org/10.3115/v1/D14-1179.

Conant, David F., Kristofer E. Bouchard, Matthew K. Leonard, and Edward F. Chang (2018). "Human sensorimotor cortex control of directly-measured vocal tract movements during vowel production". *The Journal of Neuroscience* 38.12, pp. 2382–17. ISSN: 1529-2401 (Electronic) 0270-6474 (Linking). DOI: 10.1523/JNEUROSCI.2382-17.2018.

Cooney, Ciaran, Raffaella Folli, and Damien Coyle (June 2022). "A Bimodal Deep Learning Architecture for EEG-fNIRS Decoding of Overt and Imagined Speech". *IEEE Transactions on Biomedical Engineering* 69.6, pp. 1983–1994. ISSN: 0018-9294, 1558-2531. DOI: 10.1109/TBME.2021.3132861.

Dash, Debadatta, Paul Ferrari, and Jun Wang (2020). "Decoding Imagined and Spoken Phrases From Non-invasive Neural (MEG) Signals". *Frontiers in Neuroscience* 14. ISSN: 1662-453X.

Dash, Debadatta, Ferrari Paul, Angel Hernandez, et al. (Oct. 5, 2020). *Neural Speech Decoding for Amyotrophic Lateral Sclerosis*. DOI: 10.21437/Interspeech.2020-3071.

Felgoise, Stephanie H., Vincenzo Zaccheo, Jason Duff, and Zachary Simmons (May 18, 2016). "Verbal communication impacts quality of life in patients with amyotrophic lateral sclerosis". *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration* 17.3, pp. 179–183. ISSN: 2167-8421, 2167-9223. DOI: 10.3109/21678421.2015.1125499.

Fort, Stanislav, Huiyi Hu, and Balaji Lakshminarayanan (June 24, 2020). "Deep Ensembles: A Loss Landscape Perspective". *arXiv:1912.02757 [cs, stat]*. arXiv: 1912.02757.

Gerardin, Emmanuel, Angela Sirigu, Stéphane Lehéricy, et al. (Nov. 2000). "Partially Overlapping Neural Networks for Real and Imagined Hand Movements". *Cerebral Cortex* 10.11. _eprint: https://academic.oup.com/cercor/article-pdf/10/11/1093/9751012/1001093.pdf, pp. 1093–1104. ISSN: 1047-3211. DOI: 10.1093/cercor/10.11.1093.

Gilja, Vikash, Paul Nuyujukian, Cindy A Chestek, et al. (Dec. 2012). "A high-performance neural prosthesis enabled by control algorithm design". *Nature Neuroscience* 15.12, pp. 1752–1757. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/nn.3265.

Guenther, Frank H. and Gregory Hickok (2016). "Neural Models of Motor Speech Control". *Neurobiology of Language*. Elsevier, pp. 725–740. ISBN: 978-0-12-407794-2.

Hannun, Awni Y., Andrew L. Maas, Daniel Jurafsky, and Andrew Y. Ng (Dec. 8, 2014). "First-Pass Large Vocabulary Continuous Speech Recognition using Bi-Directional Recurrent DNNs". *arXiv:1408.2873 [cs]*. arXiv: 1408.2873.

Herff, Christian, Dominic Heger, Adriana de Pesters, et al. (2015). "Brain-to-text: decoding spoken phrases from phone representations in the brain". *Frontiers in Neuroscience* 9 (June), pp. 1–11. DOI: 10.3389/fnins.2015.00217.

Kawala-Sterniuk, Aleksandra, Natalia Browarska, Amir Al-Bakri, et al. (Jan. 3, 2021). "Summary of over Fifty Years with Brain-Computer Interfaces—A Review". *Brain Sciences* 11.1, p. 43. ISSN: 2076-3425. DOI: 10.3390/brainsci11010043.

Kingma, Diederik P. and Jimmy Ba (Jan. 29, 2017). "Adam: A Method for Stochastic Optimization". *arXiv:1412.6980 [cs]*. arXiv: 1412.6980.

Laufer, Batia (1989). "What percentage of text-lexis is essential for comprehension". *Special language: From humans thinking to thinking machines* 316323.

Lotte, Fabien, Jonathan S. Brumberg, Peter Brunner, et al. (2015). "Electrocorticographic representations of segmental features in continuous speech". *Frontiers in Human Neuroscience* 09 (February), pp. 1–13. ISSN: 1662-5161 (Electronic)\r1662-5161 (Linking). DOI: 10.3389/fnhum.2015.00097.

Ludwig, Kip A, Rachel M Miriani, Nicholas B Langhals, et al. (Mar. 2009). "Using a common average reference to improve cortical neuron recordings from microelectrode arrays". *Journal of neurophysiology* 101.3, pp. 1679–89. DOI: 10.1152/jn.90989.2008.

Makin, Joseph G., David A. Moses, and Edward F. Chang (Apr. 2020). "Machine translation of cortical activity to text with an encoder–decoder framework". *Nature Neuroscience* 23.4, pp. 575–582. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/s41593-020-0608-8.

Moses, David A, Matthew K Leonard, and Edward F Chang (June 1, 2018). "Real-time classification of auditory sentences using evoked cortical activity in humans". *Journal of Neural Engineering* 15.3, p. 036005. ISSN: 1741-2560, 1741-2552. DOI: 10.1088/1741-2552/aaab6f.

Moses, David A., Matthew K. Leonard, Joseph G. Makin, and Edward F. Chang (Dec. 2019). "Real-time decoding of question-and-answer speech dialogue using human cortical

activity". *Nature Communications* 10.1, p. 3096. ISSN: 2041-1723. DOI: 10.1038/s41467-019-10994-4.

Moses, David A., Sean L. Metzger, Jessie R. Liu, et al. (July 15, 2021). "Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria". *New England Journal of Medicine* 385.3, pp. 217–227. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMoa2027540.

Mugler, Emily M, James L Patton, Robert D Flint, et al. (2014). "Direct classification of all American English phonemes using signals from functional speech motor cortex." *Journal of neural engineering* 11.3, pp. 035015–035015. ISSN: 1741-2560. DOI: 10.1088/1741-2560/11/3/035015.

Pandarinath, Chethan, Paul Nuyujukian, Christine H. Blabe, et al. (2017). "High performance communication by people with paralysis using an intracortical brain-computer interface". *eLife* 6, pp. 1–27. ISSN: 2050-084X (Electronic) 2050-084X (Linking). DOI: 10.7554/eLife.18554.

Parks, Thomas W. and James H. McClellan (1972). "Chebyshev Approximation for Nonrecursive Digital Filters with Linear Phase". *IEEE Transactions on Circuit Theory* 19.2, pp. 189–194. ISSN: 0018-9324 VO - 19. DOI: 10.1109/TCT.1972.1083419.

Proix, Timothée, Jaime Delgado Saa, Andy Christen, et al. (Jan. 10, 2022). "Imagined speech can be decoded from low- and cross-frequency intracranial EEG features". *Nature Communications* 13.1, p. 48. ISSN: 2041-1723. DOI: 10.1038/s41467-021-27725-3.

Rezeika, Aya, Mihaly Benda, Piotr Stawicki, et al. (Mar. 30, 2018). "Brain–Computer Interface Spellers: A Review". *Brain Sciences* 8.4, p. 57. ISSN: 2076-3425. DOI: 10.3390/brainsci8040057.

Romero, David Ernesto Troncoso and Gordana Jovanovic (2012). "Digital FIR Hilbert Transformers: Fundamentals and Efficient Design Methods". *MATLAB - A Fundamental Tool for Scientific Computing and Engineering Applications - Volume 1*, pp. 445–482.

Sellers, Eric W, David B Ryan, and Christopher K Hauser (Oct. 2014). "Noninvasive brain-computer interface enables communication after brainstem stroke". *Science translational medicine* 6.257, 257re7–257re7. DOI: 10.1126/scitranslmed.3007801.

Serruya, Mijail D., Nicholas G. Hatsopoulos, Liam Paninski, et al. (Mar. 2002). "Instant neural control of a movement signal". *Nature* 416.6877, pp. 141–142. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/416141a.

Silversmith, Daniel B., Reza Abiri, Nicholas F. Hardy, et al. (Mar. 2021). "Plug-and-play control of a brain–computer interface through neural map stabilization". *Nature Biotechnology* 39.3. Number: 3 Publisher: Nature Publishing Group, pp. 326–335. ISSN: 1546-1696. DOI: 10.1038/s41587-020-0662-5.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman (2014). "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps". *Workshop at the International Conference on Learning Representations*. 2014 International Conference on Learning Representations. Ed. by Yoshua Bengio and Yann LeCun. Banff, Canada.

Sun, Pengfei, Gopala K Anumanchipalli, and Edward F Chang (Dec. 1, 2020). "Brain2Char: a deep architecture for decoding text from brain recordings". *Journal of Neural Engineering* 17.6, p. 066015. ISSN: 1741-2560, 1741-2552. DOI: 10.1088/1741-2552/abc742.

Tilborg, Arjan van and Stijn R. J. M. Deckers (Mar. 31, 2016). "Vocabulary Selection in AAC: Application of Core Vocabulary in Atypical Populations". *Perspectives of the ASHA Special Interest Groups* 1.12, pp. 125–138. ISSN: 2381-4764, 2381-473X. DOI: 10.1044/persp1.SIG12.125.

Vansteensel, Mariska J., Elmar G.M. Pels, Martin G. Bleichner, et al. (2016). "Fully Implanted Brain–Computer Interface in a Locked-In Patient with ALS". *New England Journal of Medicine* 375.21, pp. 2060–2066. ISSN: 0028-4793\r1533-4406. DOI: 10.1056/NEJMoa1608085.

Webb, Stuart and Michael P. H. Rodgers (June 2009). "Vocabulary Demands of Television Programs". *Language Learning* 59.2, pp. 335–366. ISSN: 00238333, 14679922. DOI: 10.1111/j.1467-9922.2009.00509.x.

Welford, B. P. (1962). "Note on a Method for Calculating Corrected Sums of Squares and Products". *Technometrics* 4.3, pp. 419–419. ISSN: 00401706. DOI: 10.1080/00401706.1962.10490022.

Willett, Francis R., Donald T. Avansino, Leigh R. Hochberg, et al. (May 13, 2021). "High-performance brain-to-text communication via handwriting". *Nature* 593.7858, pp. 249–254. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-021-03506-2.

Williams, Ashley J., Michael Trumpis, Brinnae Bent, et al. (July 2018). "A Novel µECoG Electrode Interface for Comparison of Local and Common Averaged Referenced Signals". *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Honolulu, HI: IEEE, pp. 5057–5060. ISBN: 978-1-5386-3646-6. DOI: 10.1109/EMBC.2018.8513432.

Wilson, Guy H, Sergey D Stavisky, Francis R Willett, et al. (Nov. 25, 2020). "Decoding spoken English from intracortical electrode arrays in dorsal precentral gyrus". *Journal of Neural Engineering* 17.6, p. 066007. ISSN: 1741-2552. DOI: 10.1088/1741-2552/abbfef.

Chapter 3

The cortical dynamics of planning spoken syllable sequences

Disclaimer: This chapter contains currently unpublished material and is a direct adaptation of a manuscript that I am preparing for submission to a scientific journal. I encourage those reading this chapter to first search for the related publication, as it will contain updated material and interpretations. The published article should have a similar title and will have myself as first author, followed by Lingyun Zhao, Patrick W. Hullett, and Edward F. Chang.

Personal contributions: I conceived the project (with Edward F. Chang), designed the task, and along with other lab members collected the data. I performed all analyses (in brief, including data processing, annotation and behavior quantification, statistics, sequence and articulatory complexity encoding, reaction time prediction, and figure generation) except for NMF clustering, task phase decoding, and calculation of articulatory and auditory controls, which were performed by my co-author Lingyun Zhao. Co-author Patrick W. Hullett and I designed and collected the stimulation data, and I analyzed and quantified all stimulation results. I wrote the current draft of the manuscript.

3.1 Abstract

Speech production requires the fluent sequencing and execution of complex sequences of speech sounds. This process is traditionally conceptualized as a progression of neural processes across distinct cortical regions, each of which underlies speech planning, initiation, and execution of articulation. Although the general process of speech planning has been implicated in many frontal areas, speech-motor sequencing in particular is not well understood. We used direct high-density electrocorticography to record cortical activity while participants performed a delayed go-cue task where we modulated sequencing demands by increasing sequence complexity at the level of the syllable. Instead of only finding neural activity isolated to single task phases, we also found a distributed cortical network characterized by sustained neural activity across the encoding of the target sequence, the delay period, and through the production of the sequence. Sustained population activity reflected the progression of task phases and was strongly modulated by sequence complexity. Importantly, sustained activity in the middle precentral gyrus (mPrCG) most consistently encoded sequence complexity. Sustained activity in the mPrCG additionally predicted reaction time, indicating the specific role of speech-motor sequencing to the mPrCG. Further, electrocortical stimulation of the mPrCG caused transient speech disfluencies, consistent with apraxia of speech. These results suggest that the planning of speech sequences is mediated by a cortical network with parallel processing. In this network, the mPrCG is a critical node of speech-motor sequencing.

3.2 Introduction

Fluent speech production requires the precise planning and coordination of articulatory movements. While progress has been made in understanding how the brain controls articulatory movements, less is known about the processes upstream of continuous articulation. Broadly, traditional models of speech production propose a model of hierarchical processing, where brain areas with particular functions are activated in phasic progression (Levelt 1993). These models typically propose Broca's area (within the inferior frontal gyrus, IFG) for planning and sequencing, the supplementary motor area (SMA) for speech initiation, the precentral gyrus for phoneme level plans and/or articulatory execution, and the superior temporal gyrus (STG) or supramarginal gyrus (SMG) for phonological targets (Hickok 2012; Guenther et al. 2016).

For transient stimuli or behavior, there is indeed cortical phasic activity. For example, neural populations in the STG that are only active in processing perceived speech or neural populations in the primary motor cortex that are only active in sending low level articulatory commands to downstream effectors (Cheung et al. 2016; Chartier et al. 2018). Speech planning, on the other hand, is not necessarily time locked to behavior and can encompass several different processes including higher level functions such as idea formation, syntax, and lexical access to lower level processes such as motor sequencing (Lashley 1951; MacNeilage 1998; MacKay 1970; Levelt 1993; Hickok 2012; Hickok et al. 2022; Guenther et al. 2016). Though some studies have found neural populations that are exclusively active prior

to articulation (Flinker et al. 2015; Castellucci et al. 2022), neural populations with activity during stimulus presentation or perception and prior to and during speech production have also been found (Castellucci et al. 2022; Gregory B Cogan et al. 2017; Gregory B. Cogan et al. 2014; Leonard et al. 2019). The existence of sustained activity challenges the traditional notion that planning is restricted to a period exclusively before the onset of articulation. It may be that whether neural populations exhibit phasic or sustained activity is dependent on cortical location or the specific process of speech planning being supported. In motor neuroscience, it is well known that both phasic and sustained activity supports motor planning (Gnadt et al. 1988; Guo et al. 2017; Zimnik et al. 2021). Further, one study demonstrated that motor planning does not occur only before an initiated action, but continues to overlap with execution for subsequent actions (Zimnik et al. 2021).

In particular, sequencing is essential to fluent speech production, translating abstract speech targets (e.g. the word "production" or the syllables "pro", "duc", "tion") into serially ordered behavior (Lashley 1951). Speech sequencing has been theorized to occur at both the phonemic and syllabic levels (Levelt 1993; MacNeilage 1998; Bohland et al. 2006; Peeva et al. 2010). Importantly, sequencing directly relies on both the correct selection or generation of speech units in the sequence and the successful execution of these sequences. Sequence generation may be reflected in processes facilitating phonological target encoding. This has been hypothesized to occur in regions encompassing the posterior STG and the SMG, while the motoric aspect of sequencing has been hypothesized to occur in Broca's area (Guenther et al. 2016; Hickok et al. 2022). To further contrast these two aspects, errors of

sequence generation would likely resemble paraphasias, where entire segments are incorrect but motoric execution is not impaired (e.g. saying "collection" instead of "connection") (Binder 2017), whereas errors of motor sequencing may resemble apraxia of speech (AOS), where difficulty in producing speech increases with increasing sequence complexity (Strand et al. 2014). Though Broca's area has been implicated in speech sequencing (Bohland et al. 2006; Guenther et al. 2016; Hickok et al. 2022; Peeva et al. 2010), lesion, resection, or stimulation to this area more often causes speech arrest or anomia (a deficit in naming) (Andrews et al. 2022; Lu et al. 2021). Additionally, recent lesion and resection case studies have localized AOS to damage in the precentral gyrus (Itabashi et al. 2016; Chang et al. 2020; Levy et al. 2023), leaving an open question for what brain areas control speech-motor sequencing.

A major challenge in studying the neural correlates of speech-motor sequencing is the lack of simultaneous spatial and temporal resolution. Functional magnetic resonance imaging (fMRI) has enabled the localization of putative speech planning areas (Bohland et al. 2006) but is unable to account for dynamic processes that occur too fast to be grouped into phasic processes. In fact, fMRI localization methods that rely on defining cortical planning areas as the subtraction of production activity from pre-production activity, lessens the chance of localizing areas with sustained activity, which will have both. Both localizing areas with any (sustained or phasic) relevant planning activity and studying low-level speech-motor planning therefore require high temporal resolution with a great amount of spatial coverage. Further, determining whether cortical areas are causally involved in speech-motor planning

has largely been left to stroke lesion studies and rare case studies, where cortical damage is rarely specific and focal and analysis is post hoc. Though these studies have greatly progressed our knowledge of putative causes of disorders of speech-motor planning, like AOS, there remains a lack of a direct link between a neurobiological process in an area and perturbation of that area causing disfluencies.

To overcome these challenges, we used high-density electrocorticography (ECoG) recordings from all cortical areas that have been implicated in speech production. We used a delayed go-cue task where participants are asked to produce various syllable sequences. We vary the sequence and articulatory complexity of these sequences in order to observe the dynamics of cortical responses when low level sequencing demands are modulated at the phoneme and syllable. We observe these dynamics across 4 defined task phases: encoding, delay, pre-speech, and speech. Instead of only phasic activity, where an area is active during one phase, we found prominent sustained activity across several cortical areas, including traditional speech planning areas like Broca's area, the supramarginal gyrus, and the posterior STG, but also strongly in the precentral gyrus and supplementary motor area. We found that primarily an area in the precentral gyrus, which we term the "middle precentral gyrus" (mPrCG) and the posterior STG consistently encode the sequence complexity (complexity of the sequence at the syllable level, not the phoneme level). The mPrCG in particular also correlated with reaction time and speech errors, but did not encode articulatory movements, suggesting that the mPrCG is involved in speech-motor sequencing.

With ECoG, we were able to further directly investigate the causality of sequence com-

plexity encoding using direct electrocortical stimulation. Indeed, stimulation caused speech errors consistent with apraxia of speech, a clinical speech disorder of speech-motor programming, without direct motor or perceptual effects, only in the mPrCG and not in any other sites with sequence complexity encoding, such as Broca's area. Together, these findings challenge traditional models of speech production, first by showing that speech planning is not a purely phasic process and is instead supported in part by sustained activity. Further, we show that the mPrCG is in fact a critical node of speech-motor sequencing and provide the first neurobiological link between the mPrCG, sequencing, and apraxia of speech.

3.3 Results

Task and behavior

Thirteen participants performed a delayed go-cue task where they were prompted to read, wait a short delay, and then repeat syllable sequences (Figure 3.1A, Table 3.1). A target sequence was displayed on the screen for 2.5 seconds (termed the encoding period) before being replaced by a white fixation cross, starting the delay period. The delay period duration was slightly jittered (1 second on average). The white fixation cross would turn green to indicate the go-cue and participants were instructed to say aloud the sequence they saw. These sequences varied in their sequence and articulatory complexity (Figure 3.1B, Table 3.2). Sequence complexity refers to whether the sequence is composed by repeating the same syllable

(simple, e.g. "ba-ba-ba") or whether each element in the sequence is unique (complex, e.g. "ba-da-ga"). Articulatory complexity refers to whether the syllables used in the sequences are consonant-vowel pairs (simple, e.g. "ba") or consonant cluster-vowel pairs (complex, e.g. "gloo").

Participants were instructed to respond as soon as they could after seeing the go-cue and reaction times were 0.5 seconds on average. Participants made almost no errors on isolated syllables and simple sequences, but made more errors on complex sequences, especially complex sequences of complex syllables (Figure 3.17).

Widespread sustained neural activity

We recorded high-density electrocorticography (ECoG) from many cortical areas involved in speech production while participants performed the task (Figure 3.1B, Figure 3.6). We extracted the high-gamma (70-150 Hz) analytic amplitude (HGA) from the raw ECoG signal. This frequency band is known to correlate with local neuronal signaling (Ray et al. 2010; Steinschneider et al. 2008). We found that electrodes had the greatest HGA during the most complex condition–complex sequences of complex syllables.

Using trials where the target sequences were complex sequences of complex syllables, we found many electrodes that had significant HGA neural activity to at least one period of the task, be it encoding, delay, pre-speech, or execution (speech) (Figure 3.1C, D). Strikingly, we found electrodes within almost all of these cortical areas with persistent sustained neural

activity; that is, electrodes active during the delay period were active also during encoding, pre-speech, and execution (Figure 3.1D, E). For electrodes with neural activity significantly above baseline for at least one phase, but not all four, we call this group the "non-sustained" population. Sustained activity was localized to several cortical areas, including the inferior frontal gyrus (IFG), the middle frontal gyrus (MFG), the supramarginal gyrus (SMG), the supplementary motor area (SMA), the posterior superior temporal gyrus (pSTG), and the precentral gyrus extending from the Sylvian fissure to the transverse sulcus. Though we defined sustained activity based on the most complex condition, we also investigated whether there was sustained activity when only using trials where the target sequences were simple sequences of simple syllables. We still found electrodes with sustained activity in the same cortical areas, but this was a subset of the electrodes with sustained activity during the most complex condition (Figure 3.7), suggesting that sustained activity may be differently modulated by each condition but that it is not exclusive to higher complexity.

Sustained activity was heterogeneous with different temporal patterns. For example, some electrodes' neural activity was greater during encoding than during execution. We determined approximately four temporal patterns using convex non-negative matrix-factorization (Figure 3.1F, G, Figure 3.19). In contrast to the aforementioned pattern, two patterns had greater neural activity during execution. The final pattern's neural activity is more evenly distributed across all task phases, with a small increase in magnitude during delay. This pattern was most strongly localized to the SMA and the middle and superior frontal gyri. The precentral gyrus had the greatest amount of sustained neural activity, with the mPrCG

containing a more even distribution of the first three clusters than the vPrCG or other cortical areas (Figure 3.1F).

Sustained activity tracks internal task states

We next asked whether sustained activity contained information about task structure and behavior, according to the four different phases (encoding, delay, pre-speech, and speech production). We examined trials where the target sequence was complex sequences of complex syllables (e.g. complex sequence complexity and complex articulatory complexity) as this condition was used to define sustained activity. For individual electrodes with sustained activity, we found that a majority of the electrodes' average activity was significantly different across the four phases ($P < 0.05$, Friedman test, Figure 3.2A blue circles). Further, a subset of electrodes had significantly different activity between all of the four phases (Figure 3.2A orange triangles). For the non-sustained population, although many electrodes showed significantly different neural activity across the four phases, the overall effect size is smaller than that of the sustained population (Figure 3.2A).

Though only a subset of sustained electrodes' neural activity was significantly different between all four task phases, we asked whether the population as a whole could reliably differentiate task phases. We trained a logistic regression classifier to predict the task phase using population activity from all sustained electrodes. We found the probability of decoded phase reflects the actual task phases over time (Figure 3.2B). We further compared the decoding ac-

curacy of the sustained group to other groups of electrodes, such as the non-sustained group. Since these groups have different numbers of total electrodes, we calculated the decoding accuracy using equal numbered subsets of each group, using random resampling. For all groups of electrodes, decoding accuracy increased when more electrodes were used (Figure 3.2C). Importantly, the average accuracy when using sustained electrodes is higher than that of the non-sustained activity when controlling for the number of electrodes included and becomes significantly greater when including more than 20 electrodes (Figure 3.2C teal circles versus navy triangles). It was possible that though we observe sustained activity in many cortical areas, that only a subset contributed to decoding performance. However, we found that decoder weights were rather distributed across all included electrodes and it was not clear that any one area was dominant. To further probe this, we trained decoders that were limited to electrodes in single anatomical regions that had sustained activity. When subsetting electrodes in mPrCG with sustained activity, the decoding accuracy closely followed that of all sustained activity, despite being a markedly smaller population (56 sustained mPrCG electrodes versus 312 total sustained electrodes). As a point of comparison with a cortical area more traditionally proposed to be involved in planning and sequencing, we also calculated this growth curve for electrodes in IFG (Broca's area) with sustained activity. The decoding accuracy for this group closely followed that of the non-sustained group. Together, these results show that sustained activity spanning several cortical regions could more readily distinguish the progression of task phases than the non-sustained activity. This suggests that sustained activity is not simply static or only reflecting sensory processes, but rather

may reflect an internal state associated with progressing through the task. Further, because the mPrCG so closely reflects the entire sustained population activity, this suggests that the mPrCG may play an important role within this network.

To further understand how task phases are represented by sustained activity, we used principal component (PC) analysis to project the trial-averaged activity from all electrodes showing sustained activity onto a latent space of the first three PCs (which explained 81.8% of the variance). We found that sustained population activity formed a trajectory in PC space with distinct locations for different task phases (Figure 3.2D). It is worth noting that the population activity is modulated such that the distances traveled in the encoding and speech phases are similar (Figure 3.2E). Furthermore, we found that the trajectories across the encoding phase and across the speech production phase each travel across two separate planes (adjusted R^2 of 0.96 and 0.83 for encoding and speech production, respectively, Figure 3.2F). These two planes are neither parallel, nor perfectly orthogonal to each other, but intersect at a 56.2 degree angle. Fitting a PC space on sustained activity from mPrCG results in a similar trajectory, with each of the four phases distinct from each other (Figure 3.2G). The proportional distances traveled in the encoding and speech production phases are also similar to each other and to the total sustained group (Figure 3.2H). We found again that two planes could each be fit to the encoding and speech production phases (adjusted R^2 of 0.94 and 0.95 for encoding and speech production, respectively) and that these planes similarly intersect at a 49.8 degree angle (Figure 3.2I). In contrast, fitting a PC space on non-sustained population activity results in a very different trajectory. The trajectory from

non-sustained activity mostly represented the pre-speech and speech production phases, with the encoding and delay phases compressed in the PC space (Figure 3.2J, K). The trajectories in the encoding and speech production phases are less well described by planes (adjusted R^2 of 0.47 and 0.42 for encoding and speech production, respectively), and the angle between the planes is more acute than what is found in all sustained activity, at 39.2 degrees (Figure 3.2L). These results suggest that sustained activity during the delay and speech production is not a replay of sensory processing during encoding, as not only are task phases distinctly represented but encoding and production trajectories are more orthogonal than parallel. Further, the similarity of mPrCG population dynamics alone to the whole of the sustained network suggests that the mPrCG may be a critical node in this sustained network.

Encoding of sequence and articulatory complexity

Sustained activity prominently encoded sequence and/or articulatory complexity of the produced syllable sequences (Figure 3.3A). We determined encoding of sequence complexity by statistically comparing the average HGA magnitude during simple sequences and during complex sequences with the same articulatory complexity (e.g. “ba-ba-ba” vs “ba-da-ga”) during set time windows (Figure 3.3B). We determined the encoding of articulatory complexity similarly, comparing complex sequences where the articulatory complexity was either simple or complex (e.g. “ba-da-ga” vs “blaa-draa-gloo”).

During encoding, sequence complexity was strongly encoded across most of the sustained network. During the delay period, however, sequence complexity was encoded mostly in the precentral and frontal gyri as well as the pSTG. And finally during pre-speech, sequence complexity was largely localized to the mPrCG, with smaller clusters in the pars opercularis and the pSTG. In comparison, the encoding of articulatory complexity was slightly distributed throughout the SMG and the vSMC (centered and in close proximity to the central sulcus) during encoding, and became more strongly localized to the vSMC during pre-speech (Figure 3.3E).

Though the spatial distribution of the encoding of sequence and articulatory complexity was not static over the phases of each trial, certain areas consistently encoded each of these (Figure 3.3F). When considering only electrodes that encoded sequence complexity at every phase of the task, these were densely localized to the mPrCG and the pSTG, two areas not or not typically described in models of speech production (Figure 3.3F).

The mPrCG and pSTG, however, have been shown to be involved in other speech related processes. The mPrCG has been shown to be involved in auditory perception (Cheung et al. 2016; Venezia et al. 2021) and direct laryngeal control (Dichter et al. 2018; Eichert et al. 2020), and is possibly involved in articulatory control given its proximity to the vSMC (Chartier et al. 2018; Mugler et al. 2018). We sought to determine whether sequence and articulatory complexity effects were driven by these other functions by additionally calculating whether electrodes had significant auditory responses or encoded articulatory kinematic trajectories (AKT) (Chartier et al. 2018) (Table 3.3). We found that statistical effects of

sequence complexity (considering the greatest effect size at any single electrode across the task phases) are not significantly correlated with either auditory responses or AKT encoding performance (Figure 3.3G, H). Considering each task phase individually, we still find no significant correlations (Figure 3.8, Figure 3.9). We in fact find very few electrodes that have acceptable AKT model performance ($r > 0.1$). Of these electrodes, we consider the laryngeal component of the AKT model and also find that this is not significantly correlated with statistical effects of sequence complexity (Figure 3.3I, Figure 3.10).

Together, we find that sequence complexity is not merely derived from other functions in these areas (auditory, laryngeal, or articulatory in nature). In the mPrCG and pSTG, sequence complexity is the dominant effect, with almost no effect of articulatory complexity. This suggests that the encoding of sequence complexity in the mPrCG and pSTG is indeed reflective of their role in phonological sequencing, at the syllable level.

Predicting reaction time

Though we established that sustained neural activity was modulated by task variables (sequence and articulatory complexity), it was unclear to what degree this activity was related to higher-level speech planning versus lower-level speech-motor planning. To resolve this, we investigated whether pre-speech activity was correlated with reaction time of executed syllable sequences.

We used linear regression to predict single trial reaction times from HGA during the

encoding, delay, and pre-speech periods (Figure 3.4A). As expected, we found that pre-speech has the greatest percentage of electrodes that significantly predict reaction time (Figure 3.4B). Further, we found that when excluding electrodes that encoded articulatory movements (AKT model $r > 0.1$), the mPrCG had a greater percentage of electrodes that significantly predicted reaction time than the vSMC (Figure 3.4B). These electrodes were localized to the SMA, the mPrCG, and the vSMC. 20.62% of electrodes in the mPrCG significantly predicted reaction time, 40.00% of which also encoded sequence complexity (Figure 3.4B, C). Overlap between sequence complexity, sustained activity, and the prediction of reaction times was found almost exclusively in the mPrCG, suggesting that mPrCG plays a role in speech-motor planning for the execution of syllable sequences.

Stimulation induced speech errors

Together, the presence of strong sustained activity, the encoding of sequence complexity, the correlation with speech errors, and the prediction of reaction time suggest that the mPrCG is involved in phonological sequence planning for execution. However, it was unknown whether the mPrCG's involvement is causal and/or critical for speech sequence execution. To determine this, we applied transient electrocortical stimulation while participants produced simple and complex syllable sequences, as well as other controls (Table 3.4).

We found sites in the mPrCG that resulted in speech errors only during complex speech sequences, which we refer to as "sequencing errors" (Figure 3.5A). At these sites, because

of their location in motor cortex, high enough current amplitudes can elicit pure motor effects (e.g. jaw pulling). To consider a site positive for sequencing errors, we had to be able to identify a lower current amplitude at which sequencing errors were observed in the absence of direct or passive motor effects. If possible, stimulation sites were identified before the stimulation mapping, based on evoked neural activity during the syllable sequencing task (Figure 3.11, Figure 3.12, Figure 3.13, Figure 3.14). That is, if there was observed sustained activity or possible sequence complexity encoding, we would include this site during stimulation. While we observed normal sensorimotor and expected sensorimotor effects at other stimulation sites, we could not elicit sequencing errors at sites in Broca's area, at any current amplitude (Figure 3.5A, Figure 3.15).

Speech errors occurred only in the context of complex speech sequences, both pseudoword syllable sequences and 4-syllable real words (Figure 3.5B, G). We tested additional controls to rule out other possible causes of the speech errors. No disruption to vocalization or simple syllable sequences (e.g. "bababa") excluded speech arrest (Figure 3.5B, G). No disruption to orofacial motor movements (e.g. repeatedly performing lip pucker) excluded direct muscle effects in the context of active movement. We additionally asked 3 participants to produce complex syllable sequences, but one syllable at a time in isolation (e.g. "ba... da... ga"), with no deficits, excluding working memory as causing speech errors. When participants made speech errors, they were immediately and well aware of this fact. One participant described it as feeling "stuck on the last syllable". Of the 4 participants, only 2 made some speech errors outside of stimulation, in total 7. Of these 7, 6 were all made in the utterance

directly after stimulation ended, which could still be due to stimulation effects.

Speech errors most commonly included increased syllable segmentation (where increased inter-syllable silence is observed) and increased syllable duration (Figure 3.5C, F, Figure 3.16). Distortions were also commonly observed (Figure 3.5C). For example, in trying to say "catastrophe" the participant would produce "catastuhphe", where the /r/ is omitted and the vowel /uh/ is slightly distorted (Figure 3.5F). Other errors included stuttering (including false starts), extended pauses between words or sequences, and arrest (unable to continue speaking until stimulation had ended). Pause and arrest were much less common, with arrest being observed once in one participant and twice in another participant. Additionally, we observed one subtle instance of phonological simplification where a patient shortened "ject" to "jec" during one stimulation pulse, though he did not report anything being different (Figure 3.16)

While some speech error types, like distortions, are hard to quantify beyond their observation, syllable segmentation and syllable duration are easily quantified by manually annotating syllable boundaries. Indeed, we find that syllable segmentation (as measured by the inter-syllable duration) is significantly increased when stimulation is applied during complex sequences ($P < 0.05$ for all participants, one-sided Wilcoxon rank-sum test), and not during simple sequences. For one participant, inter-syllable duration was also significantly greater during isolated syllables of complex sequences (z=2.05, P=0.04, one-sided Wilcoxon rank-sum test), though the effect was greater for normal complex sequences (z=3.21, P=8.80 x 10-4, one-sided Wilcoxon rank-sum test). Similarly, we found that syllable duration was

significantly increased when stimulation was applied during complex sequences for all participants ($P < 0.05$, one-sided Wilcoxon rank-sum test). For one participant (EC276), syllable duration was also increased when stimulation was applied during simple sequences (z=3.42, P=2.48 x 10-3, one-sided Wilcoxon rank-sum test). However, the effect was greater for complex sequences (z=4.24, P=8.78 x 10-5, one-sided Wilcoxon rank-sum test).

In one participant (EC260), we had the opportunity to test whether speech errors were still made in the absence of auditory feedback, by asking the participant to mime. Even when miming, the participant still exhibited increased syllable segmentation and duration (visually observed and self-reported by the participant). With this participant we were able to further test whether they could initiate behavior that was not affected by stimulation (e.g. vocalization or orofacial movements) as soon as they started producing speech errors. For example, the participant would continually cycle through the days of the week and as soon as stimulation was applied and they began producing speech errors, they would transition to vocalization, all while stimulation was still ongoing. The participant was able to switch to these unaffected behaviors "with no delay" (self-reported). Though only in one participant, this further points to these speech errors as resultant from perturbations to speech-motor planning for complex sequences and not disruptions to working memory or auditory processing.

Strikingly, these speech errors caused by stimulation were consistent with pure apraxia of speech (AOS). AOS is a clinically diagnosed disorder of speech-motor planning, where patients have difficulty producing fluent and consistent speech sounds, with typically more

difficulty on multisyllabic utterances (Strand et al. 2014). AOS is distinct from dysarthria, where speech is slurred or effortful due to orofacial muscle weakness and the inability to control those muscles. AOS is also distinct from other speech errors, like anomia, agrammatism, or paraphasias, where produced speech may be incorrect but is speech-like with no motor impairments.

3.4 Discussion

Fluent speech production requires the sequencing of complex articulatory movements. Though much has been studied about how articulatory movements are represented in the speech-motor cortex (Chartier et al. 2018; Mugler et al. 2018), the cortical dynamics involved in speech sequencing are not well understood. We investigated this by directly recording cortical activity with electrocorticography (ECoG) while participants produced syllable sequences of varying sequence and articulatory complexity. We find a prominent network of sustained activity across multiple cortical areas, whose population activity reflects processing across phases of the task. Specifically, we find that the middle precentral gyrus (mPrCG) in this network consistently encodes syllable sequence complexity and correlates with reaction times and errors, suggesting that the mPrCG is involved in speech-motor sequencing for complex syllable sequences. Finally, using direct electrical stimulation to the mPrCG, we confirm that this area is causally involved in speech sequence execution for complex sequences, with speech errors resembling that of apraxia of speech (AOS). Importantly, these findings demonstrate

the need to account for both sustained and phasic temporal dynamics and for the mPrCG's role in speech-motor sequencing in models of speech production. These results provide a clear neurobiological link between AOS, speech sequencing, and the mPrCG.

Speech-motor planning of phonological sequences

Similar to previous studies, we found several cortical areas modulated by sequence and articulatory complexity (Bohland et al. 2006; Peeva et al. 2010; Rong et al. 2018). However, we found the most robust encoding of sequence complexity in the mPrCG, as opposed to Broca's area. By specifically testing the mPrCG for motor planning qualities–evoked neural activity that is before the start of production that is selective for types of movements and is predictive of how the movement is executed (reaction time) (Svoboda et al. 2018)–we differentiated sequence complexity encoding in the mPrCG from the rest of the sustained network as specific to motoric execution. The mPrCG has previously been implicated in many different roles related to speech perception and production, including auditory processing (Cheung et al. 2016; Venezia et al. 2021), reading (Kaestner et al. 2022; Dehaene et al. 2001), and laryngeal control (Dichter et al. 2018; Bouchard et al. 2013; Eichert et al. 2020), though it has been unclear whether many functions are facilitated by the mPrCG or whether a single function subserves all these purposes. In a review, we had proposed that the mPrCG plays a role in phonological sequencing for execution, and that this function is used in processes like reading and listening where phonological sequencing may be needed

(Silva et al. 2022). Our results support that notion and we show that the mPrCG plays a role in phonological sequencing for execution, where the phonological unit is at the syllabic level.

Since Broca's seminal case study, Broca's area, and not the mPrCG, had long been theorized to be involved in speech-motor planning and speech sequencing. However, several studies have challenged the idea that Broca's area is critical for speech production. Resections to Broca's area have been dissociated from Broca's aphasia and resection or stimulation more often causes anomia than other speech disfluencies (Lu et al. 2021; Andrews et al. 2022; Mohr et al. 1978). The "aphemia" that Broca originally described in his case study is now known to be more consistent with a form of AOS.

Though our analyses suggested that the encoding of sequence complexity in the mPrCG was specific to motor sequencing, we did still observe sequence complexity in other cortical areas and we did not have causal evidence to suggest that these other areas, like Broca's area, were not involved in sequencing in some aspect. To resolve this, we used direct electrical stimulation at sites that encoded sequence complexity to observe whether sequencing was affected. At putative sequencing sites in Broca's area, we either observed no deficits or perceptual deficits, rather than any deficits associated with speech production. Rather, it was sites in the mPrCG where stimulation induced sequence production errors, in the absence of direct motor effects or speech arrest. Remarkably, these errors only occurred in complex sequences for both nonword syllable sequences and real words. Participants were also keenly aware of these errors, indicating that it was not altered working memory or perceptual effects

driving these errors. Given these results, we find that the mPrCG indeed plays a critical role in speech-motor sequencing.

Timing of motor sequences has also been shown as a variable that may be facilitated by premotor areas, in songbird syllable sequences (Long et al. 2008) and in humans with sequences of finger movements (Kornysheva, Sierk, et al. 2013). Though perturbation of this variable could explain the slowed speech rate we observed, we did not observe this effect in isolation and only observed it for complex syllable sequences. Distorted substitutions and stuttering were also observed, demonstrating that the effect is not unilaterally driven by altered speech timing. Importantly, when patients make these errors during stimulation they are immediately aware that they are making errors and remember what they were supposed to say. This further suggests that what we are disrupting is not related to higher level aspects of speech planning, like working memory, but rather speech-motor sequencing.

In addition to the mPrCG, the pSTG also exhibited sustained activity that consistently encoded sequence complexity. The pSTG is in a region most would consider to be part of Wernicke's area, which has historically been defined as an area critical to language comprehension (Binder 2015). More recently, Wernicke's area has been suggested to not be critical for language comprehension but instead be critical to speech production (Binder 2015). Lesion studies have shown that damage to the pSTG and SMG result in phonemic paraphasia, suggested to result from impaired phonological retrieval (Binder 2015; Pillay et al. 2014; Quigg et al. 2006). Direct electrical stimulation studies have also shown that stimulation to the pSTG can result in speech arrest, anomia, and phonemic paraphasias without impairing

comprehension (Lu et al. 2021; Leonard et al. 2019; Binder 2015). It may be that sequence complexity encoding in the pSTG reflects its role in phonological sequence generation while in the mPrCG this reflects its role in the execution of phonological sequences. Articulatory complexity was also found, primarily localized to the SMG and vSMC. The mixed encoding of sequence and articulatory complexity in the vSMC likely reflects a role in between higher level sequence execution and low level articulatory movements and coordination.

While we focus on establishing the critical role of the mPrCG in speech-motor sequencing, the mechanism by which speech sequencing is achieved is not yet known. That is, how does the representation of a target sequence, e.g. "badaga" or "catastrophe", become serially ordered and executed in time? Several perspectives on serial order exist and have been studied in non-human primate and human motor neuroscience, such as competitive queuing and dynamic systems perspectives (Averbeck et al. 2002; Kornysheva, Bush, et al. 2019). Some models of speech production include processes that could facilitate sequence encoding, such as frame and content encoding (Guenther et al. 2016; MacNeilage 1998), or demonstrate parallel distributed processing to facilitate serial ordering of articulatory movements (Jordan 1997). However, more detailed studies applied to cortical activity are necessary to fully understand whether these mechanisms occur in areas like the mPrCG.

Sustained activity for speech-motor planning

Classic models of speech production outline a phasic process of speech production, wherein processes occurring in designated anatomical regions serially pass information from one to another until eventually motor commands are sent from the ventral sensorimotor cortex (vSMC) to be executed. Phasic activity was certainly observed during speech production articulation and auditory feedback but we also observed prominent sustained activity across many cortical areas associated with speech production. Importantly, neural activity was sustained at the single trial level and not simply the result of trial averaging across many transient peaks.

But what is the purpose of sustained activity? In cued production tasks, the presentation of the stimulus must be transformed from its external representation (e.g. letters on a screen) to some internal representation (perhaps phonological targets) that is not time locked to behavior. In more natural conversation, this may be from deciding or being prompted to speak and holding that idea to initiating speech. Sustained activity may serve to bridge sensory inputs to actionable outputs. This logic is supported by detailed animal work in motor planning that has identified multi-regional circuits associated with motor planning, where persistent activity across multiple regions, including subcortical structures, facilitates maintenance and fluent coordination of action sequences (Inagaki et al. 2022; Guo et al. 2017). Further, human studies of speech production have suggested that sustained activity might facilitate working memory and sensorimotor transformations (Gregory B Cogan et al.

2017; Gregory B. Cogan et al. 2014).

The sustained activity we observed was not localized to only the mPrCG, but was also observed in multiple cortical areas that have been associated with various aspects of speech production. This included Broca's area which has been linked to syntax and higher level planning, the SMG with phonological working memory, the SMA with motor planning and action initiation, the pSTG with phonological targets for speech production, and the vSMC which has primarily been associated with articulatory control. Though each of these areas exhibited sustained activity, they differently encoded sequence and articulatory complexity. For example, the SMG had mixed selectivity to sequence and articulatory complexity, which may reflect its role in phonological working memory at both phonemic and syllabic scales, while the vSMC had a greater concentration of articulatory complexity encoding which may reflect its role in coordinating lower level sequence execution into continuous articulatory movements around the level of the phoneme. Instead of each area sequentially processing and passing off information, it may be that they are persistently encoding information together as a network, with higher level functions modulating or informing lower level planning.

Further, although we identified roughly four patterns of sustained activity, these patterns did not simply correspond to the four task phases. Namely, it is not the case that each pattern had a peak exclusively during one phase, with each phase having such a pattern. It is also not the case that electrodes with the strongest activity during encoding were found in only one anatomical region. Instead, each cortical area exhibited a mixture of these patterns, suggesting that this network is made up of neural populations that are continuously

modulated in parallel. Population level analyses show that the sustained population as a whole traverses manifolds that are roughly 2D and more orthogonal to each other than parallel.

Clinical implications

Apraxia of speech (AOS) is a clinical disorder of speech-motor programming where patients have an impairment in planning and programming speech-motor sequences. AOS is distinct from aphasia and dysarthria (the inability to articulate due to muscle weakness or an inability to control the vocal tract), and errors are increasingly observed with complex articulatory movements and increasing length (Strand et al. 2014). Pure AOS is rare, as it is often observed after strokes where lesions compromise multiple brain structures at the same time, and is usually observed alongside dysarthria and/or aphasia. Though AOS was originally proposed to occur from damage to the superior precentral gyrus of the insula, more recent studies of stroke lesions instead suggests that AOS may be linked to damage of the ventral and middle PrCG (Graff-Radford et al. 2014; Itabashi et al. 2016). Two recent case studies with rare resection to almost exclusively the mPrCG (and some MFG) showed chronic postoperative AOS (pure AOS in Chang et al. 2020, and AOS with mild alexia and mild agraphia in Levy et al. 2023).

In determining whether sequence complexity encoding in the mPrCG is causal to sequence execution deficits, the speech errors we observed were strikingly consistent with AOS. One

test of diagnosing AOS is assessing oral diadochokinesis, where one would expect a patient with AOS to be able to produce "papapa" but have increased speech errors or difficulty with "pataka" (Strand et al. 2014)–precisely mirroring the metric of sequence complexity (e.g. "bababa" vs "badaga"). In line with recent studies localizing AOS to the mPrCG and not Broca's area, we only find these apraxic errors from stimulation in the mPrCG.

The mPrCG's role in complex speech sequencing provides a neurobiological explanation for why patients with AOS have difficulty with utterances of increasing complexity. Understanding of the neurobiological causes of speech disorders can be used to better understand the etiologies of speech disorders in patients with damage to more than one cortical area. This information can potentially help clinicians better understand what types of treatment or speech therapy (if any) may be most effective. In cases where surgical resection is indicated in and around speech areas, patients will undergo clinical mapping for language and motor functions where stimulation is used to determine what areas, if removed, might result in speech or motor deficits. Typical clinical language mapping is not as fine-grained as what we present here and may only use speech tasks with monosyllabic stimuli (e.g. counting). It is possible that this type of testing would fail to identify sites that may result in AOS. Instead, testing with stimuli including simple and complex sequences can give clinicians and surgeons a more detailed understanding of where resections involving the mPrCG can be more aggressive and where resections might begin to cause chronic speech deficits including AOS.

Conclusion

We demonstrate the critical role of the mPrCG in speech-motor planning and establish a neurophysiological link between phonological sequencing, the mPrCG, and AOS. Clinically, this link has the potential to inform surgical resection of the mPrCG and diagnosing and understanding AOS. These results further our understanding of speech-motor planning and sequencing and stress the importance of accounting for the mPrCG in models of speech production.

3.5 Methods

Participants

Thirteen individuals voluntarily participated in this study (see Table 3.1 for demographic information). All participants were fluent English speakers with normal speaking abilities undergoing surgical treatment for epilepsy involving the subdural implantation of high-density electrocorticography (ECoG) grids for 1-3 weeks. Grids were placed according solely to clinical needs, usually covering the lateral cortical surface and in two subjects also covering portions of the medial cortex. All but one participant had ECoG grids implanted on their left hemisphere. For details on grid placement, see Figure 3.6. Depending on the amount and type of data collected, some participants were not included in all analyses (described in Table 3.1). Each participant gave written informed consent before participating in the study.

Experimental protocol was approved by the Institutional Review Board at the University of California, San Francisco.

Task design

The task was designed using a delayed go-cue paradigm. In this paradigm, a target utterance was presented on the screen for 2.5 seconds before being replaced with a white fixation cross for a variable delay (on average 1 s). After this delay, the fixation cross turns green and this is the go-cue for the participant to overtly produce the target utterance. For 1 participant (EC267), they found the visual presentation task too difficult so we instead auditorily presented the utterance and the delay period started immediately at the end of the utterance.

Target utterances were chosen to modulate sequence and articulatory complexity (similar to Bohland et al. 2006). Sequence complexity was defined as whether the syllable sequence was a repetition of the same syllable three times (simple, e.g. "bababa") or was composed of three unique syllables (complex, e.g. "badaga"). We additionally include trials where only one syllable is said in isolation. Articulatory complexity was defined as whether the syllable was composed of a single consonant-vowel pair (simple, e.g. "ba") or of a consonant cluster-vowel pair (complex, e.g. "bla"). Given three simple syllables and three complex syllables, and three sequence complexity conditions, there are a total of eighteen possible target utterances (see Table 3.2).

The utterances were split into three utterance sets and collected based on variable time allowances with each participant, with the goal of getting 20 trials per target utterance. The first and second utterance set covered the target utterances necessary to assess sustained activity, sequence and articulatory complexity, reaction time decoding, speech errors, and differences in the encoding of unique sequences for complex sequences of complex syllables. Ten participants completed these sets (one of these participants doing the auditory presentation version). Three of these ten also completed the third utterance set. The remaining three participants completed all or some of only the first utterance set.

Data collection and processing

Acquisition and neural signal processing

Neural signals from the ECoG grids were recorded using TDT preamplifiers and digital processors while participants performed the delayed go-cue task. A microphone and photodiode (attached to the screen of the laptop presenting the task to keep track of task phases) were recorded as analog signals. Raw voltage signals were acquired at 3051.76 Hz and analog signals were recorded simultaneously at 24414.06 Hz.

Neural signals were preprocessed using custom Python software and the NWB file format. Each participant's neural signals were downsampled to 400 Hz and then Notch filtered for 60 Hz (and its harmonics) to remove line noise. Raw signals were then visually inspected. Electrodes visually appearing to have no signal, excessive noise, or frequent artifacts were

noted as "bad channels" and excluded from analyses. We also applied an automated procedure to identify electrodes whose 99th percentile raw voltage values, voltage derivative, and root mean square were 5 stand deviations greater than the rest of their electrodes. These electrodes were also noted as bad channels.

For analyses using raw data, this raw signal was downsampled to 1000 Hz. To extract the time-varying high-gamma analytic amplitude (HGA), the Hilbert transform was applied to 8 Gaussian filters with center frequencies between 70 to 150 Hz (center frequencies of 73.0, 79.5, 87.8, 96.9, 107.0, 118.1, 130.4, and 144.0 Hz and standard deviations of 4.68, 4.92, 5.17, 5.43, 5.70, 5.99, 6.30, and 6.62 Hz). The HGA was calculated as the mean of these 8 analytic signals. Unless otherwise stated, HGA was downsampled to 100 Hz for all analyses.

HGA was visually inspected for each participant. If HGA signals appeared to be abnormally highly correlated or contain artifacts, the raw signals were re-referenced in 64 channel blocks (according to how the grids are connected to the pre-amplifier, and excluding bad channels) before Notch filtering and HGA was recomputed. In all analyses, HGA signals were z-scored relative to inter-trial periods of silence.

Time segment artifact rejection

For all participants, we identified time segments where the HGA signal likely had high magnitude artifacts. This was done by identifying time segments where the HGA magnitude or derivative was greater than 5 times the 99th percentile value. Once identified, these time segments (plus a 15 ms buffer on each side of the segment) were excluded from analyses by

replacing the HGA with NaN values and using metrics that ignored NaNs.

Annotation

All participants' microphone data were manually annotated in Audacity at the syllable level, for task and stimulation mapping blocks. Annotations were used to determine the acoustic onset of speech and whether extraneous speech occurred during inter-trial silence periods (these silence periods were then excluded from z-scoring and analyses). For one participant (EC254), they almost exclusively pronounced the target syllable "gloo" as "glow", regardless of the context in isolated, simple, or complex sequences. Of the 160 total trials where "gloo" was in the target sequence, they said "glow" in 121 of those trials. In only 4 trials did they say "gloo". Thus, we treated their utterances of "glow" as "gloo", and marked them as correct trials.

Data analysis

All data analysis, with few exceptions, was carried out in Python 3.8 using common scientific computing packages, including NumPy, SciPy, and Pandas. The exceptions, NMF clustering, task phase decoding, and computing of auditory and articulatory controls, were computed in Matlab R2019a.

Determining significant electrodes

To determine whether electrodes had neural activity significantly above baseline during each of the 4 task phases, we used a non-parametric one-sided Wilcoxon rank-sum test, comparing task-related neural activity to baseline activity. Task-related neural activity was taken from trials where the target sequence was a complex sequence of complex syllables and where the participant did not make mistakes. Baseline activity was taken from 1 second windows in between each trial where there was no extraneous noise, no produced speech, and nothing on the task laptop screen. We defined analysis windows for each task phase–0.5 to 1.5 seconds from target presentation for encoding, 0.0 to 0.75 seconds from the onset of delay for delay, the time from the go-cue to 0.1 seconds before the acoustic onset of speech for pre-speech, and finally the window covering the duration of the produced speech for speech. For task-related and baseline activity, the average over time was taken, leaving a distribution across single trials for task versus baseline activity. P-values were FDR corrected across electrodes and the 4 task phases within each participant.

Characterizing sustained neural activity patterns

In order to determine electrodes with sustained activity, we identified electrodes whose task-related neural activity was significantly above baseline for all of the 4 time periods. We used convex non-negative matrix factorization (NMF, as in Hamilton et al. 2018) to cluster trial-averaged neural activity patterns for sustained electrodes. The number of clusters was

determined by a combination of computing the elbow and selecting the number of clusters where the temporal patterns were all visually different (e.g. not having two clusters with the same relative characteristics but simply lower magnitude in one).

Task phase analyses

In order to determine whether sustained represented relative task information, we first sought to characterize whether neural activity was different across task phases. First, we examined the modulation of single electrode neural activity across task phases by averaging neural activity in the analysis windows we defined to determine significantly above baseline electrodes. This was calculated on trials where the target sequence was a complex sequence of complex syllables. We then performed a Friedman test to determine electrodes with significantly different neural activity across the four phases. We further computed post-hoc Wilcoxon signed-rank tests to determine which electrodes' neural activity was not only different across the four task phases but was different for every phase.

After determining that there was modulation at the single electrode level, we tested whether the sustained population could differentiate task phases from each other. In order to investigate population activity, we needed to combine neural activity from multiple participants. To accomplish this, we used a procedure of random resampling for all decoding. In brief, data from each participant was randomly split into 100 training and test sets (80% and 20%, respectively). Within each set, trials were upsampled to 1000 trials, shuffled, and then combined across participants. This yielded one measure of average accuracy for

each of the 100 test sets, which was used to calculate mean and standard deviations for the time course. This procedure ensured that train and test sets did not overlap and that a wide variety of combinations across trials and participants was used. Using this procedure, we time-averaged neural activity in non-overlapping 100 ms time bins across the four task phases and fit decoders to predict which of the four task phases each time bin pertained to. All decoding was performed using L1-regularized logistic regression (with liblinear solvers), predicted single trials, and balanced class labels during training.

After establishing that task phase could be differentiated across time, we sought to compare decoding accuracy across various neural populations. To avoid the bias introduced by different numbers of electrodes in each population, we performed a resampling procedure in which we randomly selected a subset of electrodes in each group to keep the number of electrodes (N) included the same for each comparison. For each iteration, we again fit a logistic regression decoder to predict the task phase. As before, we used the random resampling procedure to combine data across participants. For simplicity, we used time-averaged neural activity from the previously defined analysis windows, as computing the full time course was not necessary for this comparison. We repeated this procedure 500 times for each N and calculated the mean and standard deviation of the decoding accuracy. Significantly different decoding accuracy between size-matched populations was determined via two-sided Wilcoxon rank-sum tests.

Principal component analyses

In order to compare population dynamics between sustained and non-sustained populations of electrodes, we performed principal component analysis and projected trial-averaged population activity onto the first three principal components (PCs). To compare how trajectories behaved within task phases we first calculated the proportional distance traveled. For any given population, we computed the distance traveled by each individual task phase and divided this by the total distance traveled across all four phases, yielding a normalized measure that we compared across populations. From these projections, it appeared that some trajectories were moving along roughly two dimensional planes. We fit 2D planes for the encoding and speech portions of the trajectory (and the resulting fit, measured by R^2), which then allowed us to compute the angle between these planes.

Sequence and articulatory complexity

To determine encoding of sequence and articulatory complexity, we considered electrodes with activity above baseline and specifically looked at the encoding, delay, and pre-speech time periods. For each electrode and task phase, we compared the time averaged neural activity between simple and complex levels of complexity across trials, using two-sided Wilcoxon rank-sum tests. Time windows for each task-phase were defined the same as in determining above baseline activity. One difference is that for pre-speech, we only included trials with reaction times between 0.2 and 1.1 seconds. Because the pre-speech windows were variable

in terms of length, we wanted to first ensure there were at least 10 timepoints to use for time average. With a cutoff at 0.1 s before acoustic onset, this meant that trials needed to have reaction times greater than at least 0.2 seconds. The upper bound was chosen so that any window being considered during pre-speech was maximally at 1 second.

For sequence complexity, the comparison was between simple sequences of simple syllables and complex sequences of simple syllables (e.g. "bababa" versus "badaga"), where articulatory complexity was not modulated (both have simple syllables). For articulatory complexity, the comparison was between complex sequences of simple syllables and complex sequences of complex syllables (e.g. "badaga" versus "blaadraagloo"), where sequence complexity was not modulated (both are complex sequences). P-values were FDR corrected across electrodes and the two complexity comparisons within each participant.

Articulatory and auditory controls

For each participant where complexity encoding was assessed, we determined which of their electrodes had significant auditory responses or high-performance articulatory encoding. For auditory responses, two-sided Wilcoxon rank-sum tests were used to determine significant responses during passive listening compared to baseline. For articulatory encoding, we fit articulatory kinematic trajectory (AKT) models (Chartier et al. 2018). This yielded a correlation value for how well the AKT model fit at a particular electrode reconstructed high-gamma neural activity and model weights for each electrode. For one participant (EC237), there was not enough pitch variation in the articulatory dataset, and so the pitch feature (f0

pitch, to measure the larynx) was excluded. Details on what utterance sets were used for both auditory and articulatory measures are detailed in Table 3.3.

Reaction time decoding

For each electrode and each trial (where the target sequence was complex and there were no mistakes) we took 0.75 second long windows (from -0.85 seconds to -0.1 seconds before acoustic onset of speech) and computed 3 neural features for each window: the mean across time, the variance across time, and the estimated slope. We further only included trials where the reaction time was at least 0.2 seconds long. This was to ensure that there were at least 10 timepoints of activity between the go-cue and the acoustic onset. Single trial reaction times were predicted from these 3 neural features using a linear regression model and a leave-one-out training and testing scheme. We computed the Spearman rank correlation between the distributions of predicted and true reaction times and significance was determined by permutation testing where we shuffled one group and recomputed the correlation (1000 permutations). We additionally predicted reaction times based on single electrode, single trial activity for the encoding and delay task phases. The windows we used were identical to those used in finding significant above baseline activity.

Statistical analysis

Unless stated otherwise, distributions were compared using non-parametric two-sided Wilcoxon rank-sum tests. Multiple comparisons were corrected using FDR correction. The significance

threshold was set to 0.05 for all tests.

Density maps

All density maps, unless otherwise stated, were calculated as Gaussian-smoothed normalized histograms. In brief, two 2D histograms would be computed–one considering all electrodes with significant activity during a particular phase of interest, which was treated as the "baseline" histogram, and one including the same electrodes but with weighting. For sequence complexity and articulatory complexity densities, these weights were simply valued at 1 for electrodes with significant complexity encoding or 0 otherwise. The resulting weighted histogram would then be divided by the "baseline" histogram and smoothed with a Gaussian kernel (standard deviation of 2). For the overlap between reaction time and pre-speech sequence complexity electrodes, a density was computed for each. Since sequence complexity electrodes were pertaining specifically to those during pre-speech, the sequence complexity z-value was used as weighting with nonsignificant electrodes set to 0. In order to combine these densities, each density was first normalized to sum to 1 before summing the two densities and dividing by the maximum value.

Stimulation mapping

For 4 participants (EC260, EC267, EC276, and EC282), we investigated whether the mPrCG was causally involved in the execution of syllable sequences by using electrocortical stimulation. As part of their clinical treatment for epilepsy, participants normally undergo stimula-

tion mapping to determine cortical areas critical for motor function and language production and processing. We used this paradigm to additionally test whether stimulation caused a deficit for producing complex syllable sequences in the mPrCG as well as sites in Broca's area and the pSTG.

Stimulation was performed using a Natus clinical stimulator with standard settings. Delivered stimulation was recorded with a stimulation trigger (cable connecting the output stimulation from the Natus stimulator to the analog recordings) aligned to the recorded microphone signal (in one participant, the stimulation trigger was not recorded and so delivery of stimulation was instead estimated from the recorded neural activity). Sites at which stimulation resulted in after discharges were not tested further. Duration of stimulation was set to capture the duration of roughly two production attempts. Stimulation current varied between stimulation site, participant, and between some of the tasks. For sites involved in speech production, high enough current amplitudes can induce passive motor effects (e.g. lip twitch) which could contribute to speech errors. To divorce direct motor effects from any speech production deficits, we first identified this current level and then attempted to titrate to a lower current level at which we could observe production deficits in the absence of passive motor effects. If a site was found to have deficits producing 4-syllable real words (e.g. "catastrophe") at a particular current amplitude and/or higher current amplitudes caused motor effects, this site was further tested (referred to here as a potential sequencing site). If a lower amplitude could not be found to induce production deficits in the absence of passive motor effects or if only passive motor or sensory effects were observed, sites were

marked as “motor” or “sensory” sites.

Further testing of potential sequencing sites involved testing whether the lower current amplitude (which did not cause passive motor effects) also caused production deficits in complex syllable sequences (e.g. “badaga”), simple syllable sequences (e.g. “bababa”), vocalization (vocalizing and holding a vowel), sequences of finger movements, and sequences of orofacial movements (e.g. repeatedly opening and closing the jaw). Testing complex versus simple sequences identified whether production deficits were specific to complex sequencing demands. Testing whether stimulation disrupted vocalization controlled for a potential effect of speech arrest at these sites at these current amplitudes. Testing finger movements and orofacial movements determined whether production deficits were specific to speech-motor movements.

In one participant (EC260), we additionally tested whether mimed speech was affected by stimulation. With this participant, we also tested whether control tasks could be initiated just as the participant was noticing difficulty producing speech, while stimulation was still ongoing.

A full description of tasks and instructions to the participants is described in Table 3.4. Only stimulation sites at which we tested for sequencing errors are included in Figure 3.5. Many other sites were tested as a part of the clinical speech and motor mapping, inducing passive motor, sensory, and perceptual effects consistent with previous studies (Leonard et al. 2019; Lu et al. 2021) but as they were not additionally tested for sequencing we do not describe them here.

Characterization of stimulation induced production errors

Speech errors were manually annotated from the recorded microphone signal. Errors were categorized according to the following categories:

1. Syllable segmentation: characterized by increased silence between syllables within a sequence or word.

2. Distortion/distorted substitution: characterized by phonemic insertions, substitutions, or deletions that are phonetically distorted from typical pronunciation. Distorted substitutions are distinct from phonemic paraphasias where phonemic substitutions are made without phonetic distortion.

3. Syllable duration: characterized by increased duration of syllables relative to typical production.

4. Stuttering disfluencies: characterized by stuttering on phonemes and syllables, defined here to include false starts, halting, and articulatory groping.

5. Speech arrest: characterized by a pause in speech without recovery until after the end of the stimulation window.

6. Hesitation: characterized by a pause in speech with recovery during the stimulation window.

7. Phonological simplification: characterized by omissions of phonemes in syllables without phonological distortion.

In the event that multiple instances of the same error type were present in a single word or sequence attempt, only one was labeled. For hesitation and speech arrest, where the error may be occurring in between word or sequence attempts, the error was assigned to the previous word or sequence attempt.

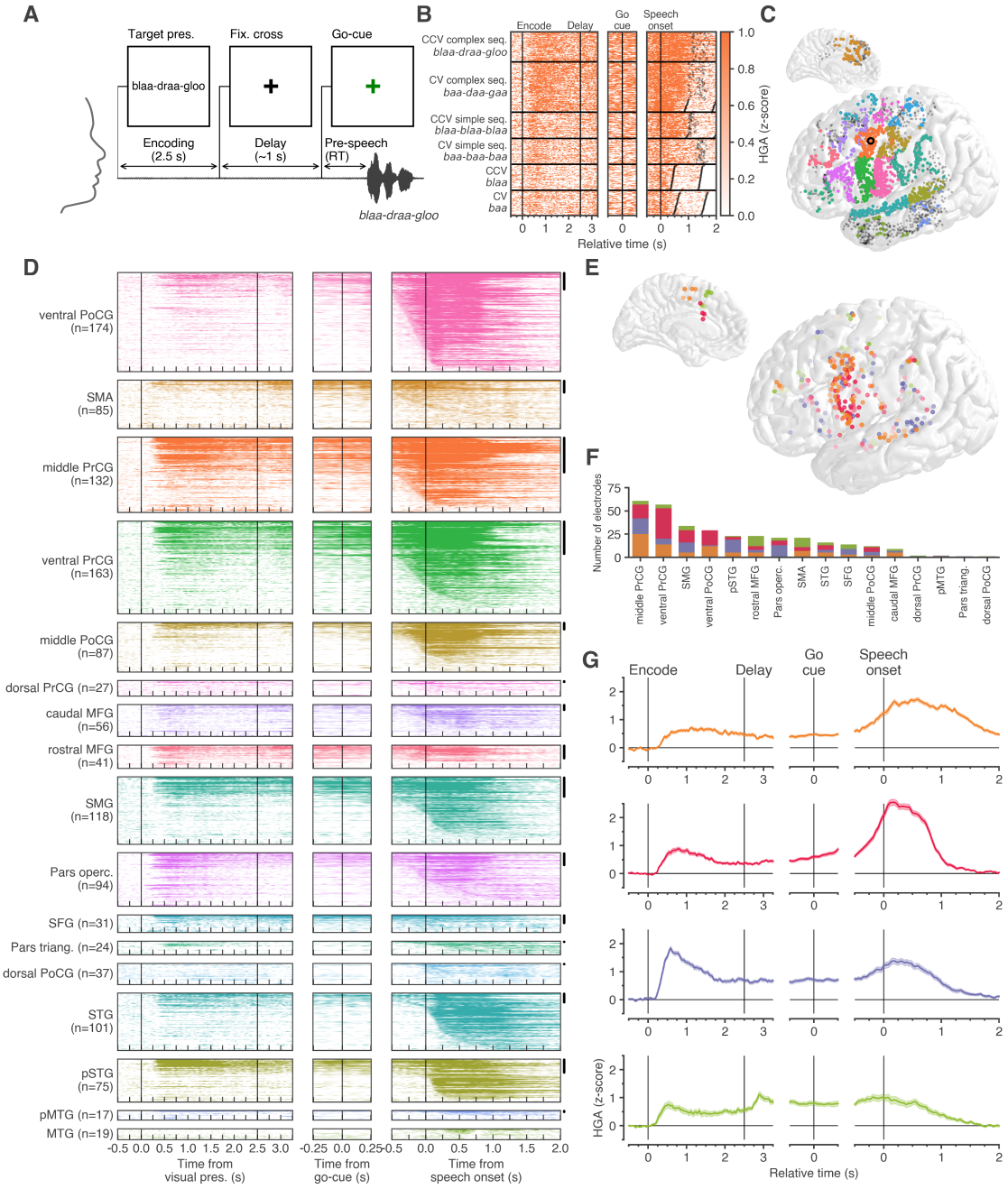


Figure 3.1. Sustained cortical activation from encoding to planning and production. (continued on next page).

(Previous page.) **Figure 3.1. Sustained cortical activation from encoding to planning and production. A.** The syllable sequence production task is a delayed go-cue paradigm. At the start of each trial, a target syllable sequence is visually displayed to the participant. After 2.5 seconds, the target sequence is replaced with a white fixation cross (the delay-cue) for an average of 1 second. When the cross turns green (the go-cue), the participant produces the target sequence. The task period during which the target sequence is on the screen is referred to as encoding and the time between the fixation cross and the go-cue is referred to as delay. Between go-cue and the acoustic onset (dependent on the participant's reaction time (RT)) is referred to as pre-speech. **B.** Single trial high-gamma activity (HGA) for an electrode in the middle precentral gyrus (highlighted in **C.**) for every condition, aligned to encoding, delay, the go-cue, and speech onset. If the offset of speech, or additionally the end of the trial, are within the plotted window, they are marked by black dots. Only trials where no speech production mistakes were made are shown. **C.** Electrodes with significant neural activity during at least one phase of the task are plotted as colored dots. Remaining non-significant electrodes are shown as smaller black dots. Electrodes are shown on an averaged brain reconstruction (MNI-152) and are colored according to anatomical region. For visualization purposes, only the medial and lateral left hemisphere is shown (n=10 left hemisphere participants) though cortical activity from electrodes in both hemispheres and their medial surfaces are included in other panels (n=1 right hemisphere participant) unless otherwise noted. **D.** Single electrode activations during syllable sequence production task. Heatmap showing the average high gamma activity during a complex sequence of complex syllables, where each row corresponds to an electrode from **C.** (except where the electrode is from the right hemisphere or medial surface). Colors correspond to anatomical regions in **C.** Each electrode that is defined as having sustained activity (significant neural activity during the encoding, delay, pre-speech, and speech phases) is marked with a black dot on the right side of the raster. **E.** Unsupervised clustering (via NMF) of sustained electrodes reveals four distinct temporal patterns of cortical activation. Spatial distribution of electrodes in each cluster is shown on the left lateral and medial surfaces on an average brain. For visual purposes, only the left lateral and medial surfaces are shown. The opacity of each electrode corresponds to its NMF weight for its assigned cluster (normalized by the maximum weight for that cluster) with a more opaque color meaning a stronger NMF weight. **F.** Each anatomical region and the number of electrodes that belong to each of the four clusters. The total height of each bar corresponds to the total number of sustained electrodes in each region, with the mPrCG containing the most. **G.** Average high gamma activity (mean $\pm$ standard error) relative to presentation of the target sequence (left), the go-cue (center), and production onset (right) across electrodes in each of the four clusters.

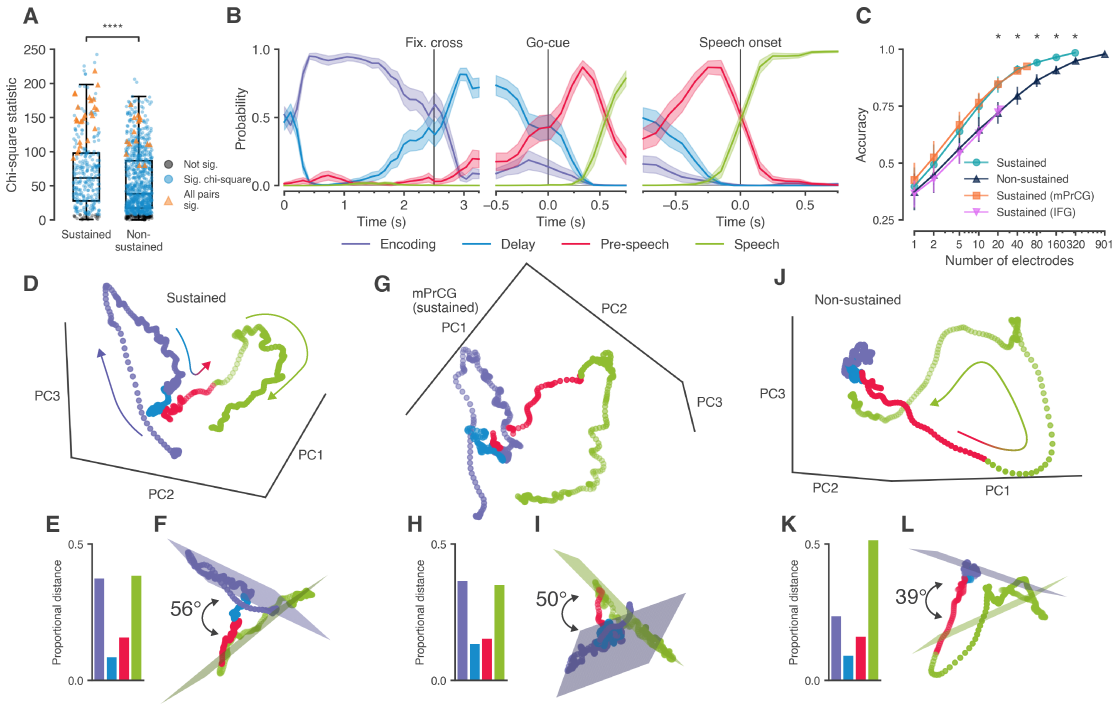


Figure 3.2. Sustained cortical activity differentiates task phases during speech planning and production. **A.** Single electrodes have significantly different time-averaged neural activity across task phases ($P < 0.05$, Friedman test) show differences in activity across task phases (tested by averaged activity in specific time windows during each phase). Chi-square values from the Friedman test are plotted for sustained and non-sustained electrodes, with significant electrodes shown in blue and nonsignificant electrodes shown in grey. Additional post-hoc testing identified some electrodes whose time-averaged neural activity was significantly different between all pairs of task phases ($P < 0.05$, two-sided Wilcoxon signed-rank test, shown as orange triangles). Overall, the sustained population has significantly greater chi-square values than the non-sustained population (**** $P < 1 \times 10^{-4}$, two-sided Wilcoxon rank-sum test). **A.** The sustained population correctly predicts the task phase across time. The time course of the decoded probability (mean $\pm$ standard deviation) of each task phase using the sustained population, aligned to target presentation (left), go-cue (center), and speech production onset (right). The probability of encoding, delay, pre-speech, and speech are shown as purple, blue, pink, and green lines, respectively. **A.** Comparison of task phase decoding accuracy using different groups of electrodes, matched for population size (number of electrodes) by a random resampling procedure. Error bars indicate standard deviation. Asterisks indicate significant differences between sustained (teal circles) and non-sustained groups (navy blue triangles) for matched population sizes. Additionally, decoding accuracy is shown for subsets of the sustained population, namely electrodes in the mPrCG (orange squares) and in the IFG (pink upside down triangles). **A.** Trajectories of trial-averaged activity from all sustained electrodes projected onto the first three principal components (PCs). The starting point of the trajectory is the onset of encoding and the endpoint is 1.5 seconds after the start of production, with annotated arrows showing the direction of time. Time points are colored according to their task phase, with purple for encoding, blue for delay, pink for pre-speech, and green for speech. **A.** The proportional distance traveled by the trajectory in **D.** during each phase. Proportional distance is calculated as the distance traveled during a particular phase divided by the total length of the trajectory. **A.** A rotated view of the same trajectory in **D.**, with 2D planes fitted for encoding and speech phases. The angle between the planes is annotated. (**G-I**) Similar to **D-F**, for electrodes in mPrCG with sustained activity. (**J-L**) Similar to **D-F**, for all non-sustained electrodes.

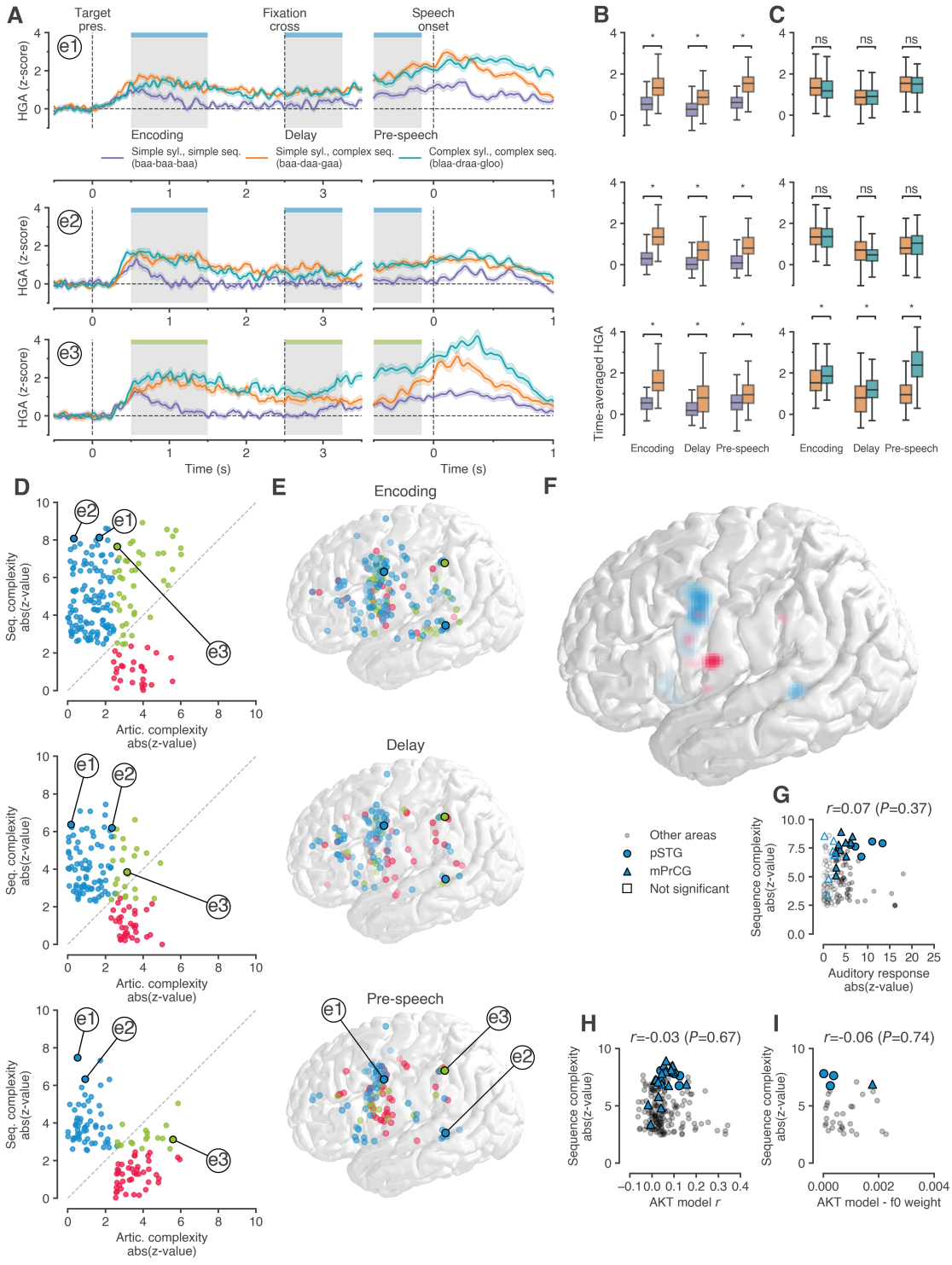


Figure 3.3. Sequence complexity and articulatory complexity modulate neural activity for planning spoken syllable sequences. (continued on next page).

(Previous page.) **Figure 3.3. Sequence complexity and articulatory complexity modulate neural activity for planning spoken syllable sequences. A.** The average high-gamma neural activity (HGA) (mean ± standard error) from example electrodes in the mPrCG, pSTG, and SMG (top to bottom) for three types of sequences: simple sequences with simple syllables (purple), complex sequences with simple syllables (orange), and complex sequences with complex syllables (teal). HGA is aligned to target presentation (left) and production onset (right). For complex sequences, HGA remains sustained throughout the delay period at a greater magnitude than simple sequences. Shaded regions indicate the encoding, delay, and pre-speech periods used to determine the effect of sequence and articulatory complexity. Colored bars at the top of each shaded region indicate the type of complexity effect observed (blue for sequence complexity only and green for both sequence and articulatory complexity). **B.** Time-averaged HGA between trials of simple and complex sequences for the encoding, delay, and pre-speech periods. Wilcoxon rank-sum testing between simple and complex sequences determines whether the electrode has an effect of sequence complexity for that time period. Panels from top to bottom are for the three example electrodes in **A.** Simple sequences are shown in purple and complex sequences are shown in orange. **C.** Same as **B.**, for determining articulatory complexity. Sequences of simple syllables are shown in orange and sequences of complex syllables are shown in teal. **D.** Scatter plot comparing the size of complexity effects (absolute value of the z-statistic) for sustained electrodes during the encoding, delay, and pre-speech periods. The x-axis indicates articulatory complexity (a difference between sequences of simple and complex syllables) and the y-axis indicates sequence complexity (a difference between simple and complex sequences). Electrodes with a significant effect of sequence complexity but not articulatory complexity are shown in blue. Electrodes with a significant effect of articulatory complexity but not sequence complexity are shown in pink. Electrodes with significant effects of both sequence and articulatory complexity are shown in green. The example electrodes from **A.** are indicated with labels. **E.** Electrode locations from left hemisphere participants (n=9). Electrodes from **D.** with either significant sequence complexity or articulatory complexity effects are shown in blue and pink. Electrodes with both effects are shown in green. The example electrodes from **A.** are indicated with labels. **F.** Density map of electrodes that maintain sequence (blue) or articulatory (pink) complexity effects through the encoding, delay, and pre-speech phases. This localizes the mPrCG and the pSTG as key areas encoding sequence complexity. Articulatory complexity is more consistently localized to the vSMC with smaller clusters near the mPrCG and the SMG. **G.** Sequence complexity effect size compared to auditory response size. The sequence complexity z-value compared here is the greatest effect per electrode across the 3 task periods shown. Auditory responses are calculated from neural activity recorded while participants listened to sentences. The effect of sequence complexity is not significantly correlated with auditory responses. Electrodes contributing to the density map in **F.** from the pSTG and mPrCG are represented by blue circles and triangles, respectively. All other sequence complexity electrodes are represented by gray circles. Electrodes without significant auditory responses have open markers while electrodes with significant auditory responses have color-filled markers. **H.** Similar to **G.**, but comparing sequence complexity effect size to the performance of an articulatory kinematic trajectory (AKT) model. **I.** Similar to **G.**, but comparing sequence complexity effect size to the f0 weight (a proxy for laryngeal movement) from the AKT model.

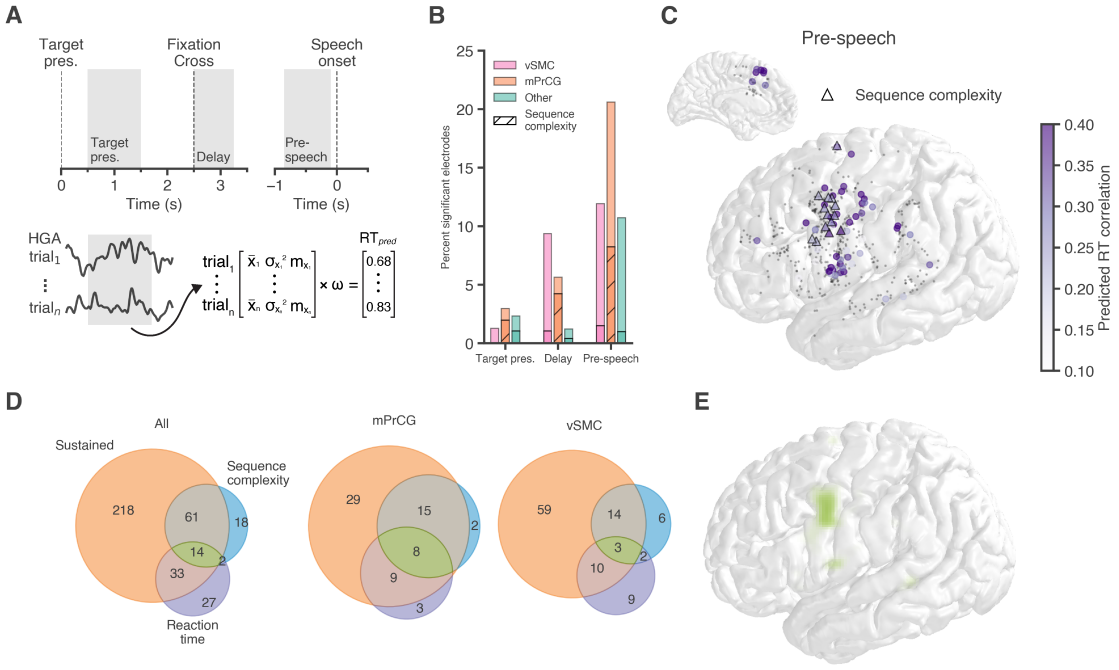


Figure 3.4. mPrCG pre-speech activity predicts behavioral reaction time. A. Definition of task phases and linear regression model for predicting reaction time. For each task phase, electrode, and trial, the mean, variance, and estimated slope of the HGA in that window is computed and used as features for predicting reaction time. The Spearman rank correlation coefficient is computed between the predicted and true reaction time values. Permutation testing (P=1000) determined which electrodes predicted reaction time distributions that were significantly correlated ($P < 0.05$) with the true reaction times. **B.** Percent of electrodes in each region with significant neural activity during that task phase that significantly predicted reaction time. These percentages exclude electrodes that encode articulatory movement. Hashed lines indicate what percentage were also sequence complexity electrodes. The mPrCG has the highest proportion of electrodes during pre-speech. **C.** Spatial distribution of electrodes that significantly predict reaction time based on pre-speech neural activity projected on a common MNI brain. Electrodes that are also sequence complexity are marked with triangles. Color scale reflects the Spearman rank correlation coefficient for each electrode. Electrodes with significant neural activity above baseline during pre-speech but that did not significantly predict reaction time or did but also encode articulatory movements are shown as small black dots. **D.** Overlap of electrodes that correlate with reaction time, have sustained neural activity, and encode sequence complexity, shown as venn diagrams. From left to right, the venn diagrams show the overlap when considering all electrodes, only the mPrCG, or only the vSMC. In contrast to all electrodes and the vSMC, the mPrCG has a much greater proportion of overlap between sustained activity, reaction time, and sequence complexity electrodes. **E.** Spatial density map of the overlap between reaction time electrodes and sequence complexity. The density map is computed as a Gaussian smoothed histogram, normalized by the number of electrodes with above baseline pre-speech activity.

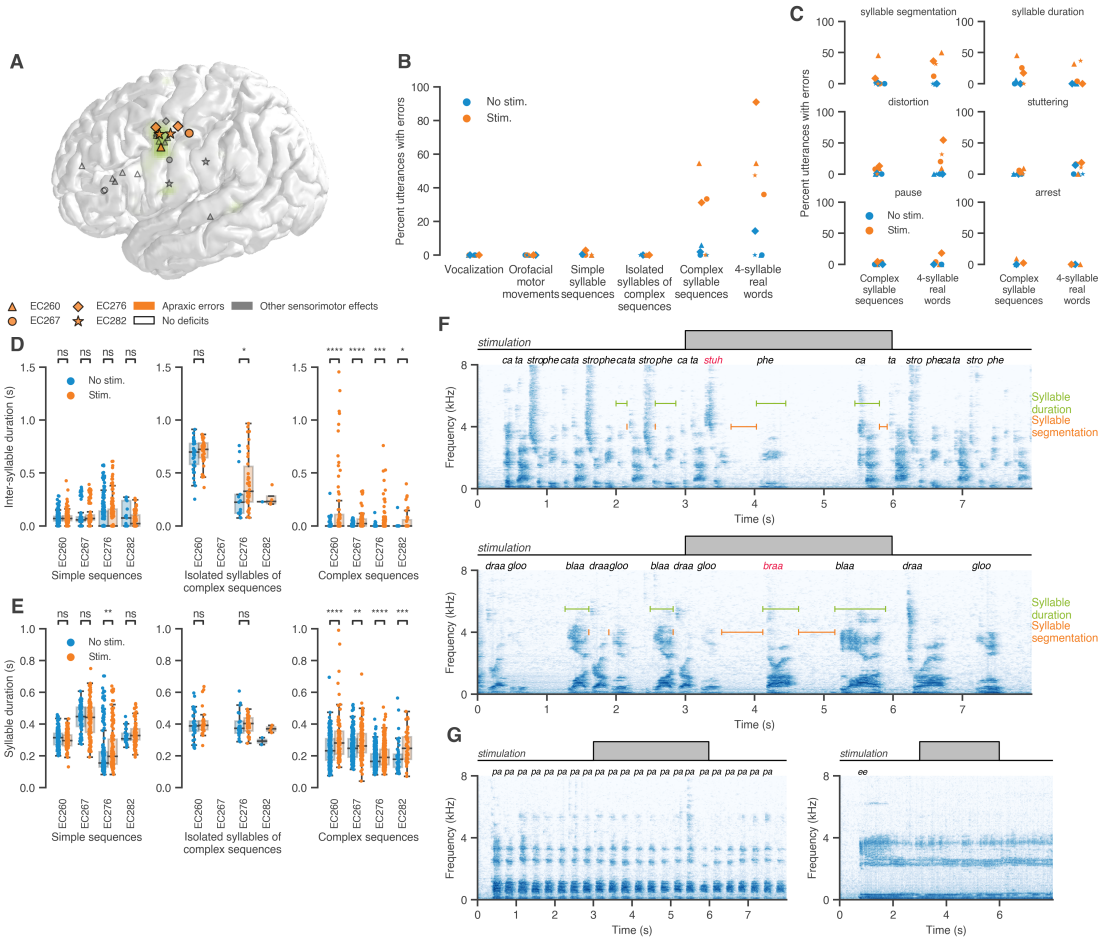


Figure 3.5. Direct cortical stimulation of the mPrCG results in apraxic speech errors. (continued on next page).

(Previous page.) **Figure 3.5. Direct cortical stimulation of the mPrCG results in apraxic speech errors. A.** Location of electrode pairs on a common MNI brain used during direct cortical stimulation in 4 patients. Marker shapes correspond to each of the 4 patients while marker color refers to the effect. A density computed on all subjects depicts the overlap between reaction time and sequence complexity electrodes (green, same as Figure 3.4E). Sites at which there were apraxic errors separable from direct motor effects are in orange. Each stimulation site with apraxic errors involved electrodes with sequence complexity effects during pre-speech (Figure 3.11, Figure 3.12, Figure 3.13, Figure 3.14). Sites at which there were no apraxic errors but instead direct motor effects or sensory/perceptual effects are in gray. Sites at which there were no effects, perceptual or to production, have no color and are outlined in black. **B.** The percent of utterances with a speech error with and without stimulation. Speech errors resulting from stimulation occur during complex syllable sequences of both pseudo- and real words. Marker shapes correspond to each of the 4 patients (as in **A.**). **C.** The percent of complex sequence utterances with speech errors, broken down by 6 error types. Error profile proportions differ across patients, but all are characteristic of apraxia of speech. Marker shapes correspond to each of the 4 patients (as in **A.**). **D.** Quantification of the inter-syllable duration (to measure increased syllable segmentation) for simple and complex sequences with and without stimulation. Inter-syllable duration is significantly greater for complex sequences, but not simple sequences ($P < 0.05$, one-sided Wilcoxon rank-sum test). or complex sequences where each syllable is executed in isolation, during stimulation. For isolated syllables of complex sequences, EC276 showed also significantly greater inter-syllable duration. EC267 did not perform this task during stimulation testing and EC282 had too few samples to statistically evaluate. **E.** Quantification of the duration of syllables in simple and complex sequences while stimulation was and was not occurring. For 3 out of 4 patients, only syllables during complex sequences were significantly longer in duration when stimulation was occurring ($P < 0.05$, one-sided Wilcoxon rank-sum test). For EC276, syllables during both simple and complex sequences were significantly longer during stimulation. For isolated syllables of complex sequences, EC267 did not perform this task during stimulation testing and EC282 had too few samples to statistically evaluate. **F.** Representative examples of stimulation causing apraxic errors during complex syllable sequence production for the 4-syllable word "catastrophe" and the 3-syllable sequence "blaa-draa-gloo". In each panel, the stimulation pulse applied to the orange triangle stimulation site in **A.** is plotted above a spectrogram of the participant's speech. Syllable annotations are marked at their onset. Both examples show a slowed syllable rate (green bars marking syllable duration increasing during stimulation), increased syllable segmentation (orange bars marking inter-syllable duration increasing during stimulation), and phonological distortions (marked by red annotated syllables). **G.** Representation examples of stimulation evoking no errors during simple syllable sequence production (fast "pa" repetitions) and sustained vocalization of a vowel.

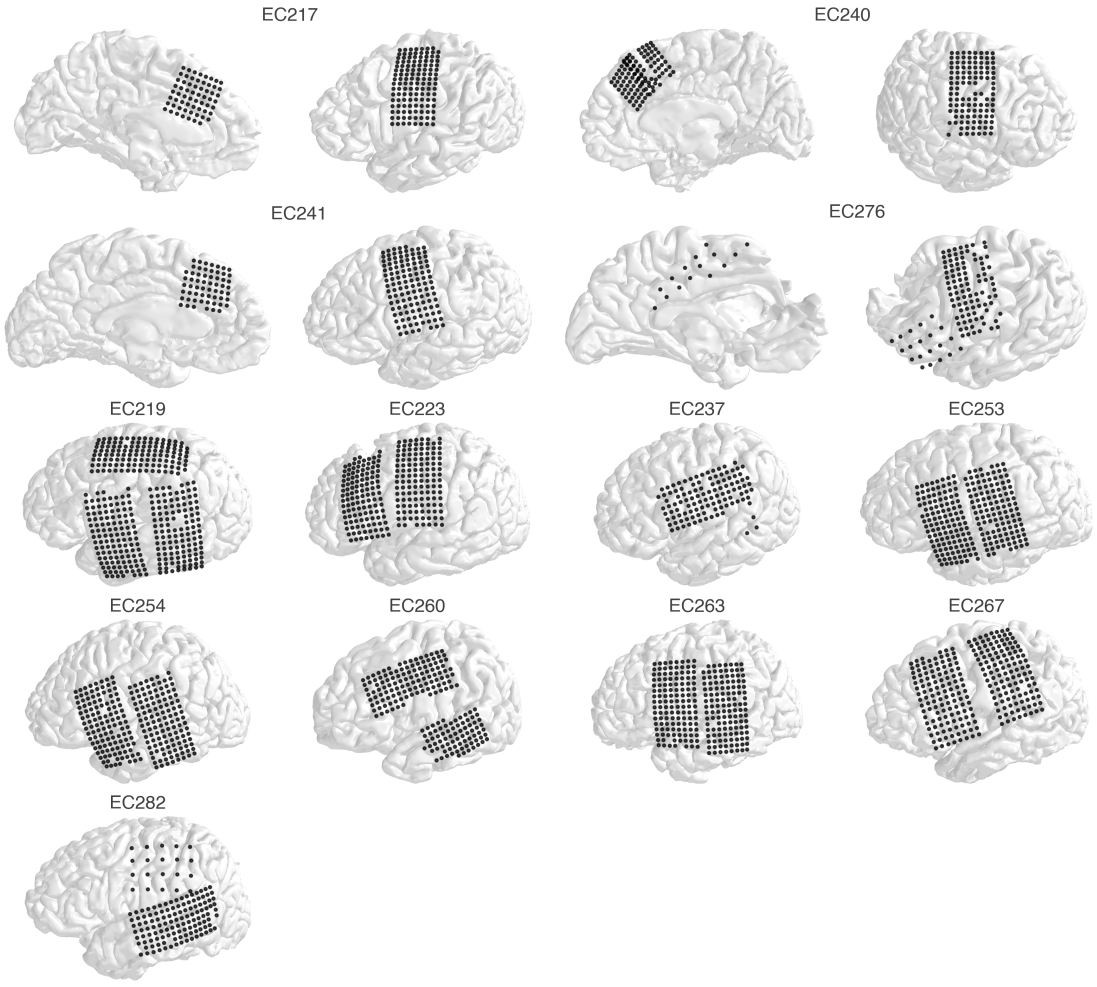


Figure 3.6. Participant electrode coverage. Electrode coverage for each participant (n=13, 12 left hemisphere, 4 included medial coverage) plotted on reconstructions of their brain. Electrodes that were continually noisy, dead, or artifactual for all recorded blocks are not shown as they were excluded from all considered analyses.

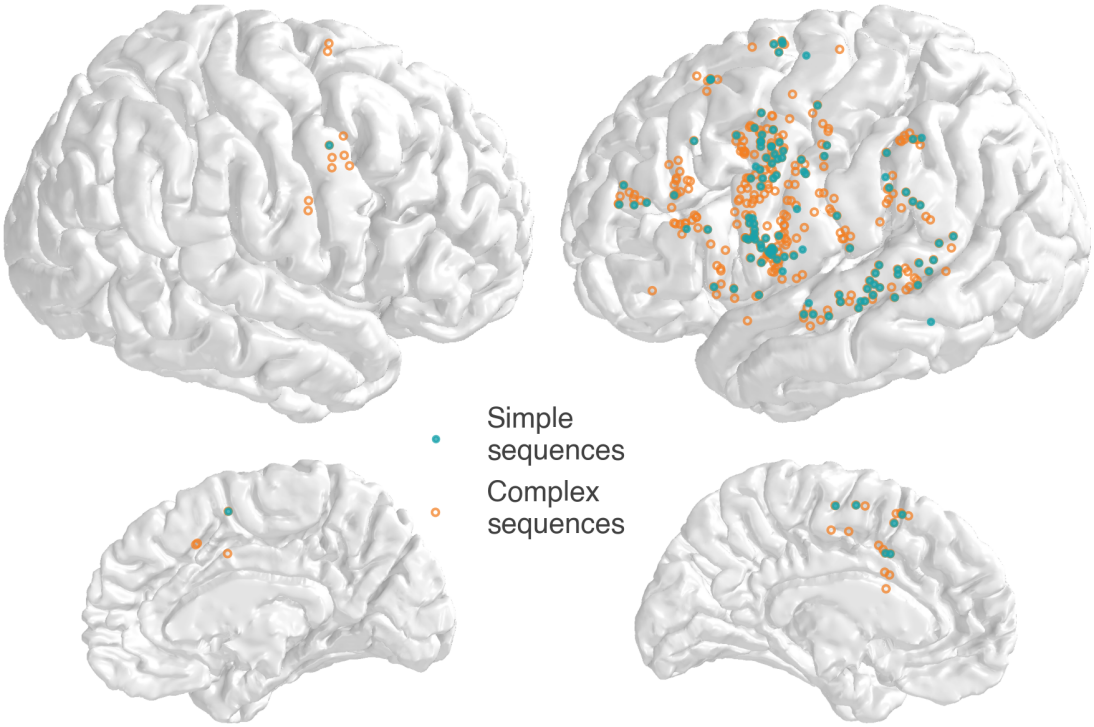


Figure 3.7. Electrodes with sustained activity during simple sequences. Electrodes with sustained activity during simple sequences of simple syllables (i.e. "bababa", filled green dots) compared to during complex sequences of complex syllables (i.e. "blaadraagloo", unfilled orange dots), plotted on an average brain (MNI-152). In the main text, we consider sustained activity defined by complex sequences of complex syllables. Considering only simple sequences there is still sustained activity, but it is a subset of the larger network and most prominent in the precentral gyrus, suggesting that sustained planning activity needed for any articulated utterance is common to the precentral gyrus while more areas are recruited for more complex sequences.

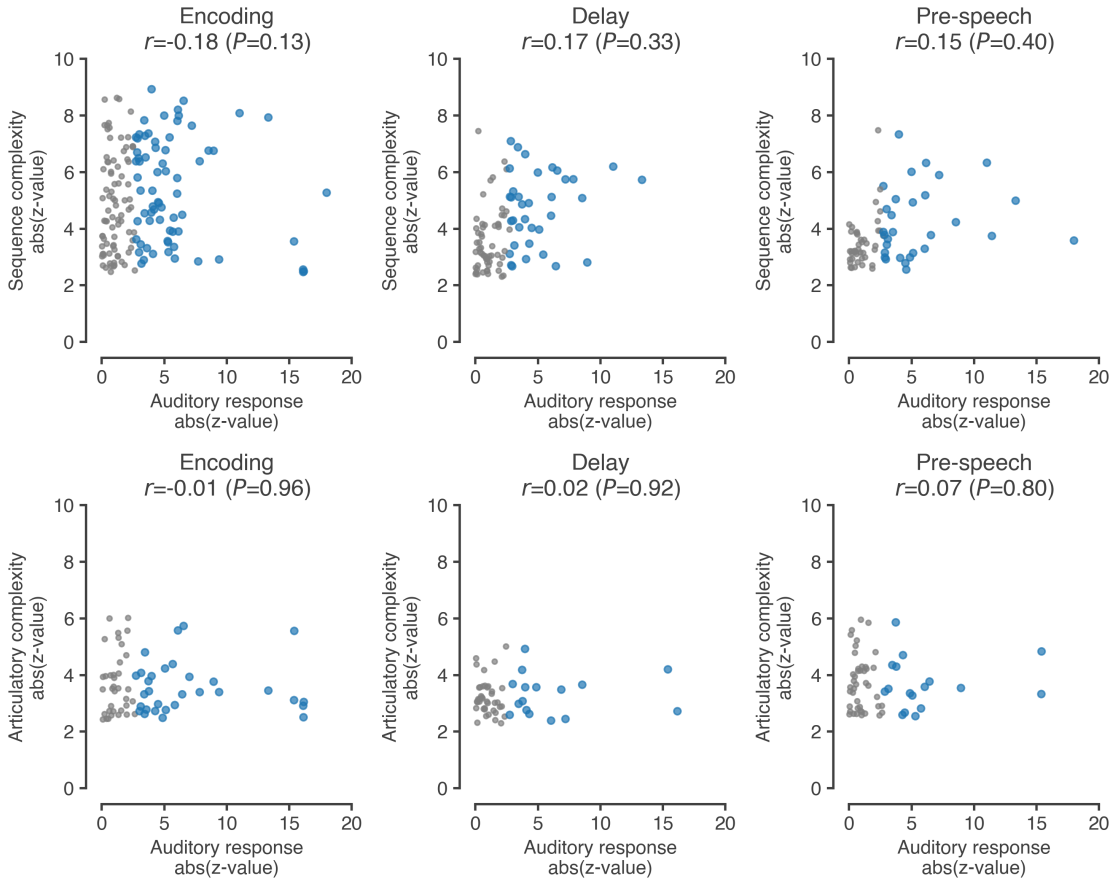


Figure 3.8. Encoding of sequence and articulatory complexity compared to auditory responses. Absolute valued z-values (from Wilcoxon Rank-Sum testing) of sequence complexity (top row) and articulatory complexity (bottom row) for the encoding (left), delay (middle), and pre-speech (right) periods versus auditory responses. Only electrodes with significant sequence or articulatory complexity effects are shown. Electrodes without significant auditory responses are shown in gray. The Pearson correlation coefficient (p-value determined by permutation testing) was computed for electrodes with both significant complexity encoding and auditory responses, with no significant correlations.

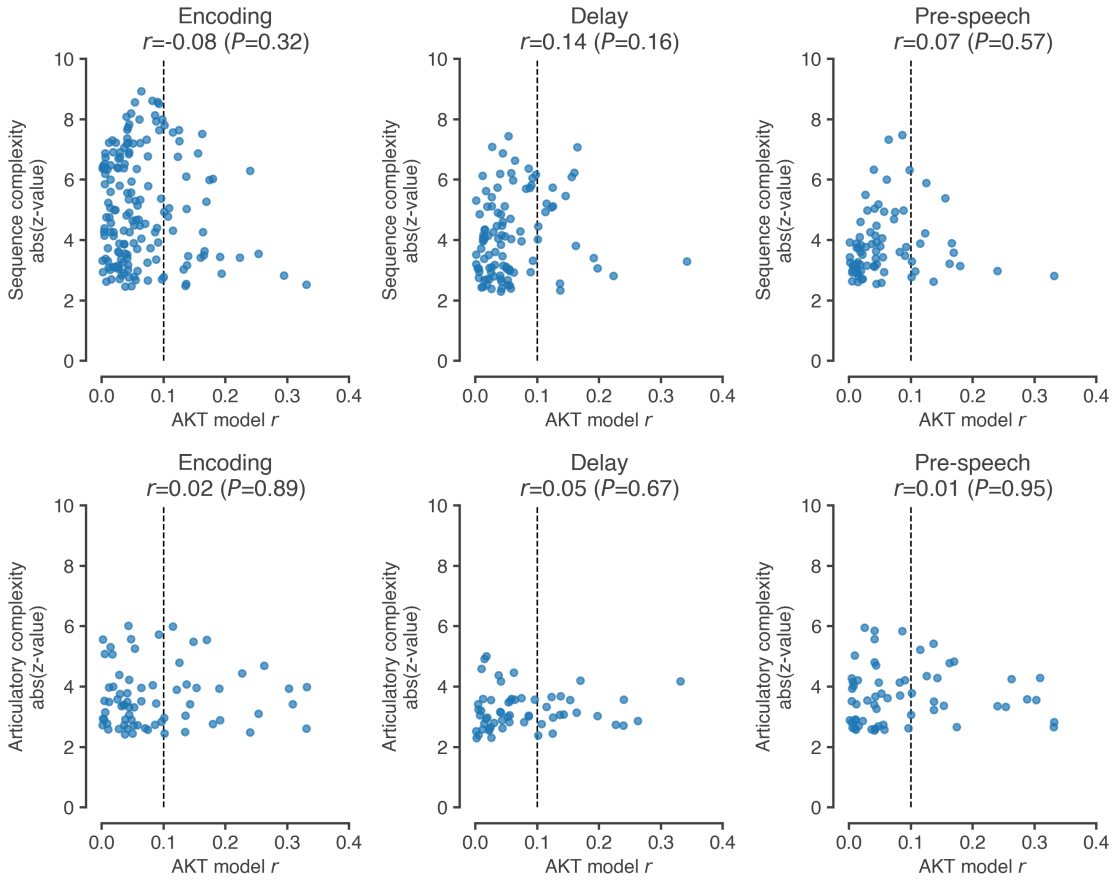


Figure 3.9. Encoding of sequence and articulatory complexity compared to articulatory encoding. Absolute valued z-values (from Wilcoxon Rank-Sum testing) of sequence complexity (top row) and articulatory complexity (bottom row) for the encoding (left), delay (middle), and pre-speech (right) periods versus articulatory kinematic trajectory (AKT) model performance (correlation of reconstructed neural activity). The Pearson correlation and p-value (determined by permutation testing) is shown for each scatter plot. AKT model performance was not significantly correlated to sequence or articulatory complexity for any time period. AKT model performance of r=0.1 was chosen as a cut-off for considering an electrode to encode AKTs (vertical dashed line).

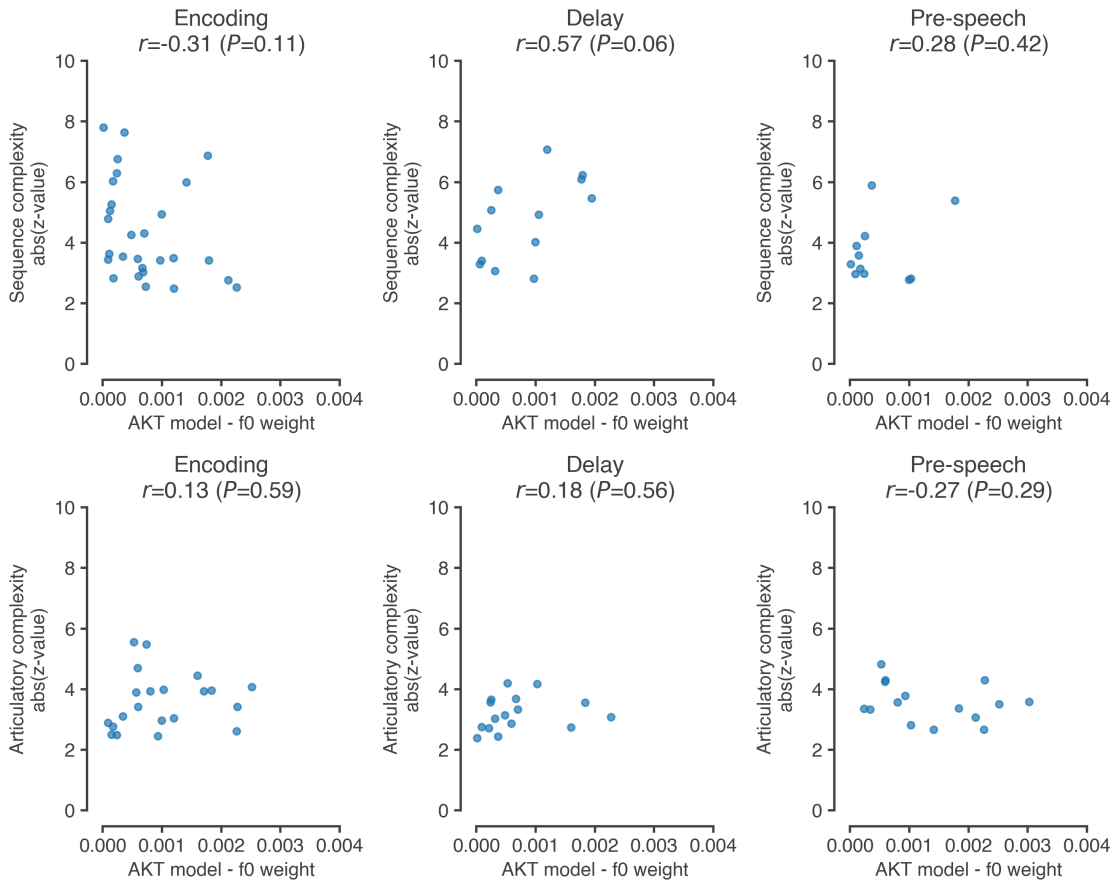


Figure 3.10. Encoding of sequence and articulatory complexity compared to laryngeal (f0) encoding in articulatory models. Absolute valued z-values (from Wilcoxon Rank-Sum testing) of sequence complexity (top row) and articulatory complexity (bottom row) for the encoding (left), delay (middle), and pre-speech (right) periods versus the AKT model laryngeal encoding weight (f0). Only electrodes with an AKT model correlation greater than 0.1 are plotted. The Pearson correlation and p-value (determined by permutation testing) is shown for each scatter plot. AKT laryngeal encoding was not significantly positively correlated to sequence or articulatory complexity for any time period.

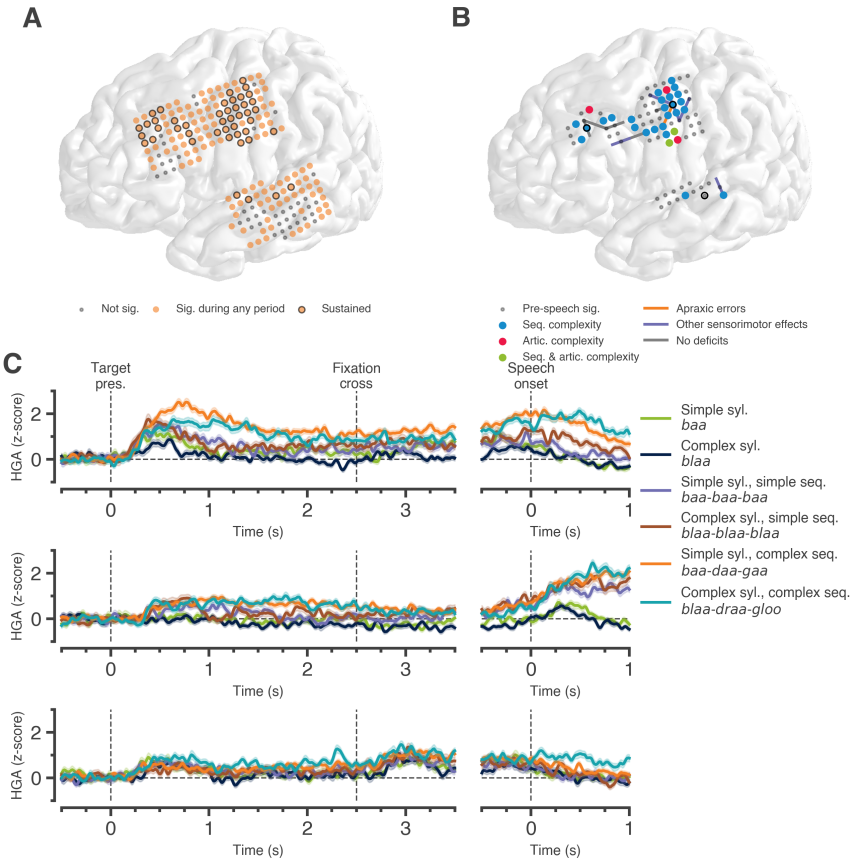


Figure 3.11. Sustained activity, complexity encoding, and stimulation sites for EC260. A. Electrode coverage for this participant on a 3D reconstruction of their brain. Electrodes that had neural activity significantly above baseline during at least one task phase are plotted in orange, electrodes with significantly above baseline activity during the encoding, delay, pre-speech, and speech task periods are plotted as orange with black outlines, and all other electrodes are plotted in gray. Electrodes that were noisy, dead, or artifactual during all recorded data are not plotted. **B.** Electrodes with sequence complexity (blue) or articulatory complexity (red) (or both in green) during the pre-speech period. Other electrodes with significant activity during pre-speech, but no complexity encoding, are shown in gray. The electrodes with black outlines correspond to ERPs shown in **C.** Lines connecting two recording sites correspond to stimulation sites. Lines are colored by the effect (if any) at that site; orange for apraxic error, purple for direct motor or other sensory effects, and gray for no deficits. For EC260, the sensory effect at the site in the IFG was altered visual and auditory perceptions associated with a former teacher they knew as a child. The sensory effect at the site in the STG was slightly altered pitch, seemingly only for speech sounds and not for other sounds like the sound of hands rubbing together. **C.** Selected high-gamma activity (HGA) ERPs for the 6 task conditions this participant completed. The top row corresponds to the mPrCG electrode, the middle to the STG electrode, and the bottom to the IFG electrode in **B.** The site in the STG has sustained activity, but is not selective for complex sequences, rather this electrode seems selective for any multi-syllabic sequence.

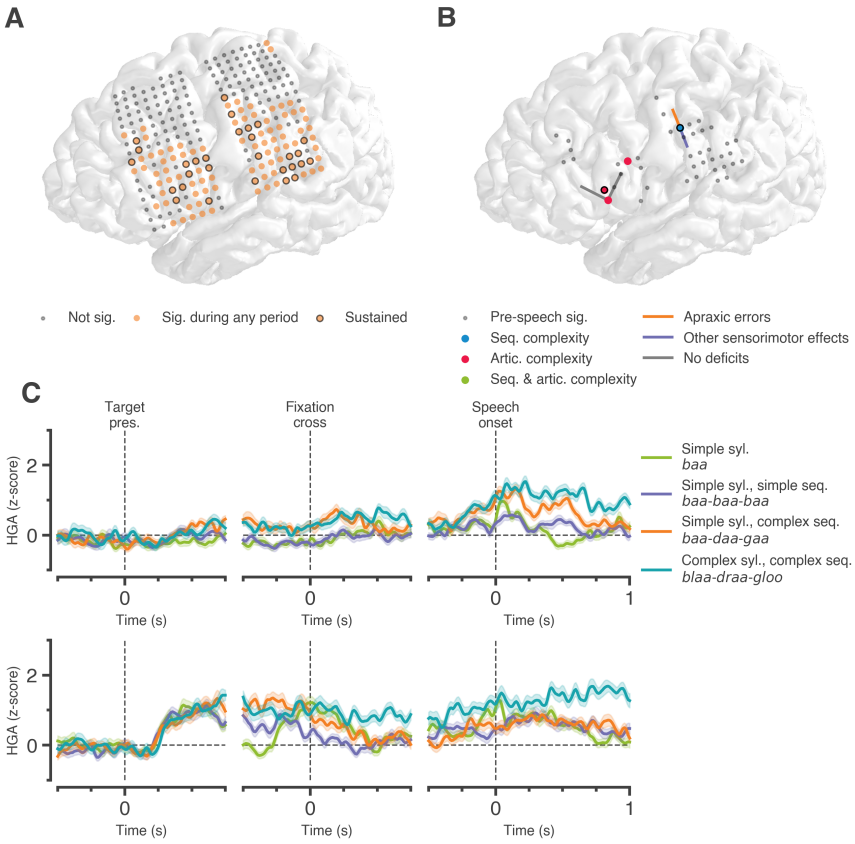


Figure 3.12. Sustained activity, complexity encoding, and stimulation sites for EC267. EC267 was only able to complete the task by completing a listen and repeat version (instead of reading the target sequences on the screen), and so they were excluded from the analyses of Figure 3.1, Figure 3.2, and Figure 3.3 for consistency. This figure depicts the key results from this participant, if they had been included. **A.** Electrode coverage for this participant on a 3D reconstruction of their brain. Electrodes that had neural activity significantly above baseline during at least one task phase are plotted in orange, electrodes with significantly above baseline activity during the delay, pre-speech, and speech task periods are plotted as orange with black outlines, and all other electrodes are plotted in gray. Due to this participant's encoding period being listening instead of reading, significant activity during this time period was not required for the electrode to be considered "sustained" for this figure. Electrodes that were noisy, dead, or artifactual during all recorded data are not plotted. **B.** Electrodes with sequence complexity (blue) or articulatory complexity (red) (or both in green) during the pre-speech period. Other electrodes with significant activity during pre-speech, but no complexity encoding, are shown in gray. The electrodes with black outlines correspond to ERPs shown in **C.** Lines connecting two recording sites correspond to stimulation sites. Lines are colored by the effect (if any) at that site; orange for apraxic error, purple for direct motor or other sensory effects, and gray for no deficits. For EC267, the sensorimotor deficit at the stimulation site over the postcentral gyrus was a direct motor effect. At the stimulation site in the mPrCG that caused apraxic errors, speech arrest due to motor effects was observed at higher stimulation current amplitudes. **C.** Selected high-gamma activity (HGA) ERPs for the 4 task conditions this participant completed. The top row corresponds to the mPrCG electrode in **B.** while the bottom row corresponds to the IFG electrode in **B.**

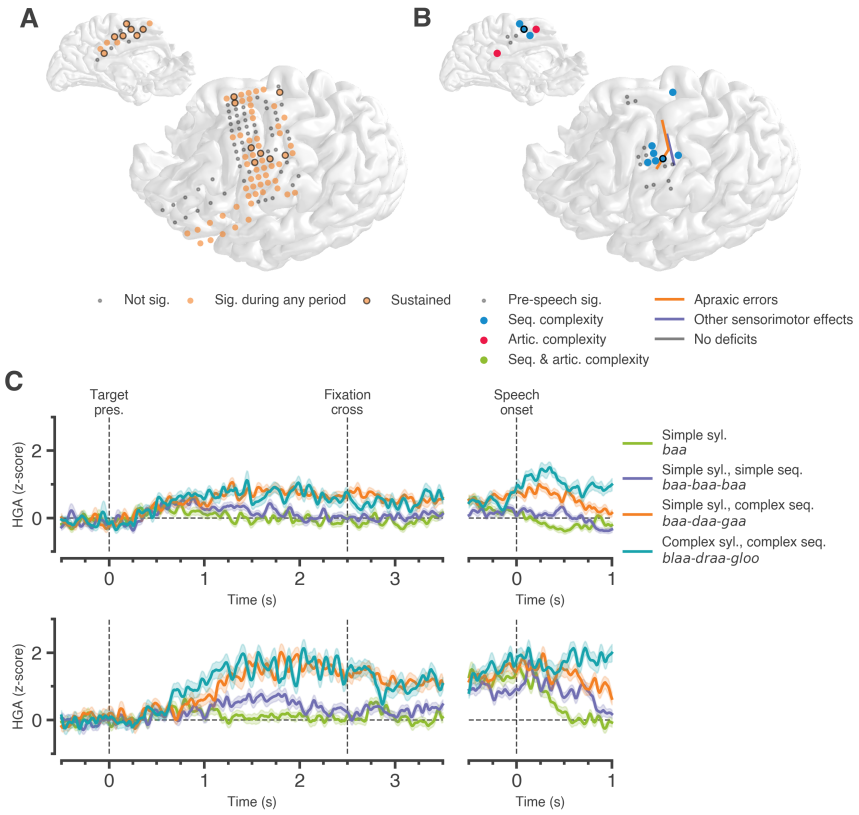


Figure 3.13. Sustained activity, complexity encoding, and stimulation sites for EC276. EC276 had reaction times that were much slower than the other patients, and so they were excluded from analyses involving the pre-speech period in Figure 3.2 and Figure 3.3 for consistency. This figure depicts the key results from this participant, if they had been included. **A.** Electrode coverage for this participant on a 3D reconstruction of their brain. Electrodes that had neural activity significantly above baseline during at least one task phase are plotted in orange, electrodes with significantly above baseline activity during the encoding, delay, pre-speech, and speech task periods are plotted as orange with black outlines, and all other electrodes are plotted in gray. Electrodes that were noisy, dead, or artifactual during all recorded data are not plotted. **B.** Electrodes with sequence complexity (blue) or articulatory complexity (red) (or both in green) during the pre-speech period. Other electrodes with significant activity during pre-speech, but no complexity encoding, are shown in gray. The electrodes with black outlines correspond to ERPs shown in **C.** Lines connecting two recording sites correspond to stimulation sites. Lines are colored by the effect (if any) at that site; orange for apraxic error, purple for direct motor or other sensory effects, and gray for no deficits. For EC276, the sensorimotor deficit at the stimulation site in the mPrCG was spontaneous vocalization. At the stimulation site across the sequence complexity site outlined in black, spontaneous vocalization was observed at higher stimulation amplitudes. **C.** Selected high-gamma activity (HGA) ERPs for the 4 task conditions this participant completed. The top row corresponds to the mPrCG electrode in **B.** while the bottom row corresponds to the SMA electrode in **B.**

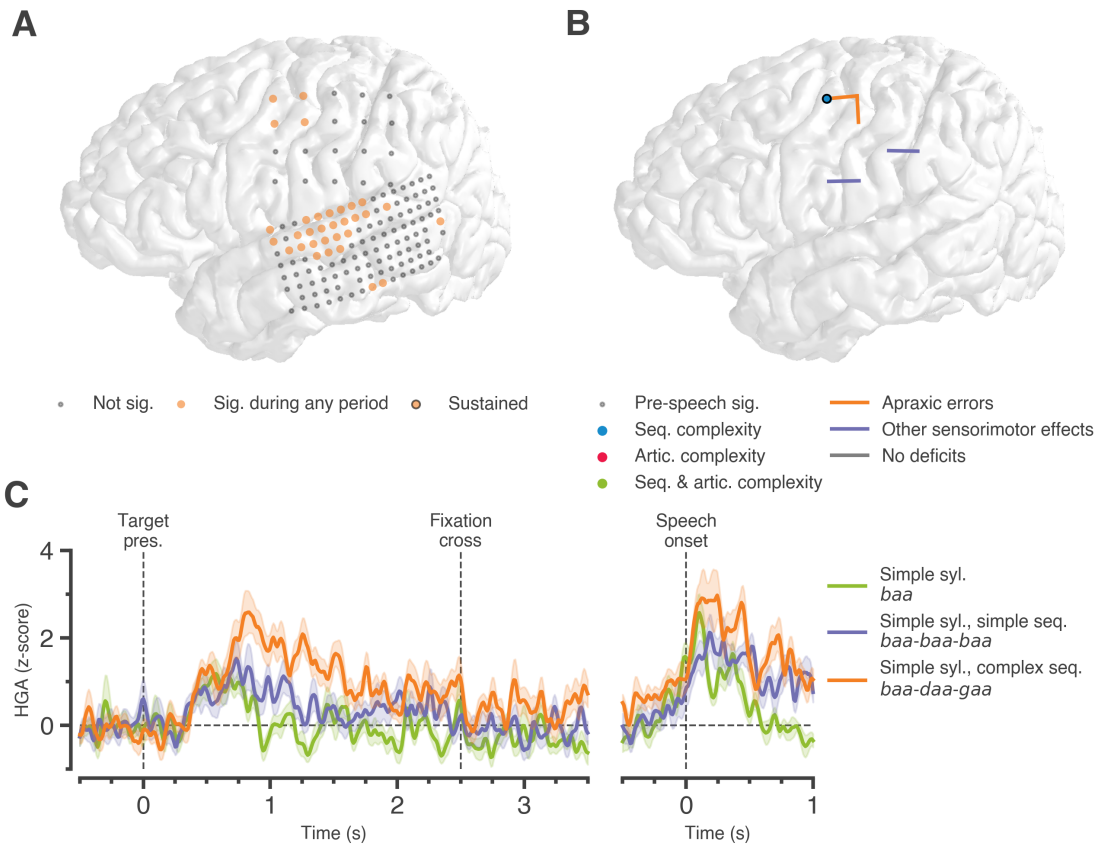


Figure 3.14. Sustained activity, complexity encoding, and stimulation sites for EC282. EC282 made errors on all complex sequences of complex syllables and only completed 1 block (5 repetitions per utterance) and so they were excluded from analyses in Figure 3.1, Figure 3.2, Figure 3.3, and Figure 3.4 for consistency. This figure depicts the key results from this participant, if they had been included. **A.** Electrode coverage for this participant on a 3D reconstruction of their brain. Electrodes that had neural activity significantly above baseline during at least one task phase are plotted in orange, electrodes with significantly above baseline activity during the encoding, delay, pre-speech, and speech task periods are plotted as orange with black outlines, and all other electrodes are plotted in gray. Electrodes that were noisy, dead, or artifactual during all recorded data are not plotted. **B.** Electrodes with sequence complexity (blue) or articulatory complexity (red) (or both in green) during the pre-speech period. Other electrodes with significant activity during pre-speech, but no complexity encoding, are shown in gray. The electrode with a black outline corresponds to the ERP shown in **C.** Lines connecting two recording sites correspond to stimulation sites. Lines are colored by the effect (if any) at that site; orange for apraxic error, purple for direct motor or other sensory effects, and gray for no deficits. For EC282, the sensorimotor deficit at the stimulation site over the central sulcus in the vSMC was a painful tingling sensation on the patient's lip and tongue that increased with stimulation current amplitude. The sensorimotor deficit at the stimulation site over the postcentral gyrus and supramarginal gyrus was a tingling sensation on the patient's head. At the stimulation sites in the mPrCG that caused apraxic errors, direct motor effects were observed at higher stimulation amplitudes. **C.** Selected high-gamma activity (HGA) ERP for the 3 task conditions this participant completed, corresponding to the mPrCG electrode in **B.**

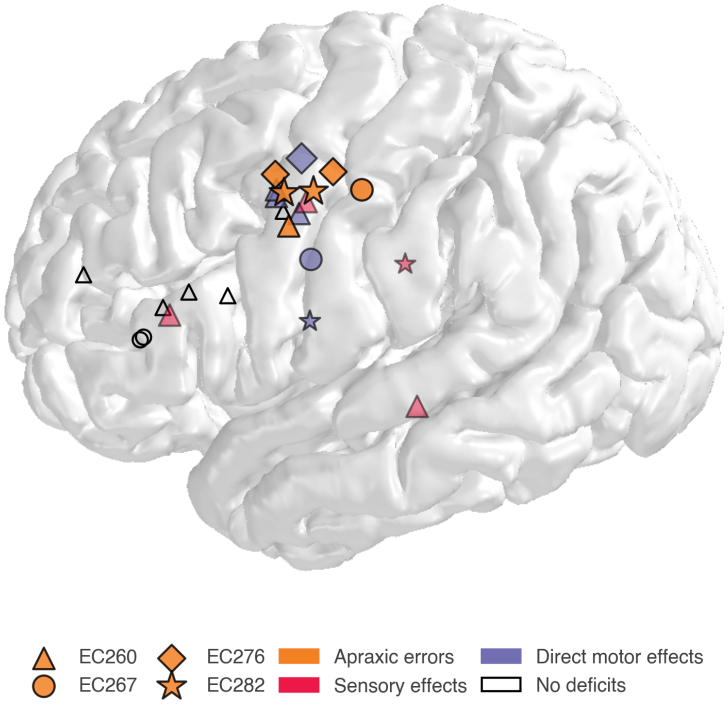


Figure 3.15. Stimulation sites for each participant and description of sensory effects. Stimulation sites are plotted on a common brain (MNI-152). Sites where stimulation caused apraxic speech errors are shown in orange. Sites where stimulation caused direct motor effects (such as movement or spontaneous vocalizations) are shown in red. Sites where stimulation caused sensory effects are shown in purple. The sensory effect at the site in the IFG for EC260 (triangles) was altered visual and auditory perceptions associated with a former teacher they knew as a child. The sensory effect at the site in the STG for EC260 was slightly altered pitch, seemingly only for speech sounds and not for other sounds like the sound of hands rubbing together. The sensory effect in the mPrCG for EC260 was right hand numbness. Finally, the sensory effect in the supramarginal gyrus for EC282 was a tingling sensation on the patient's head.

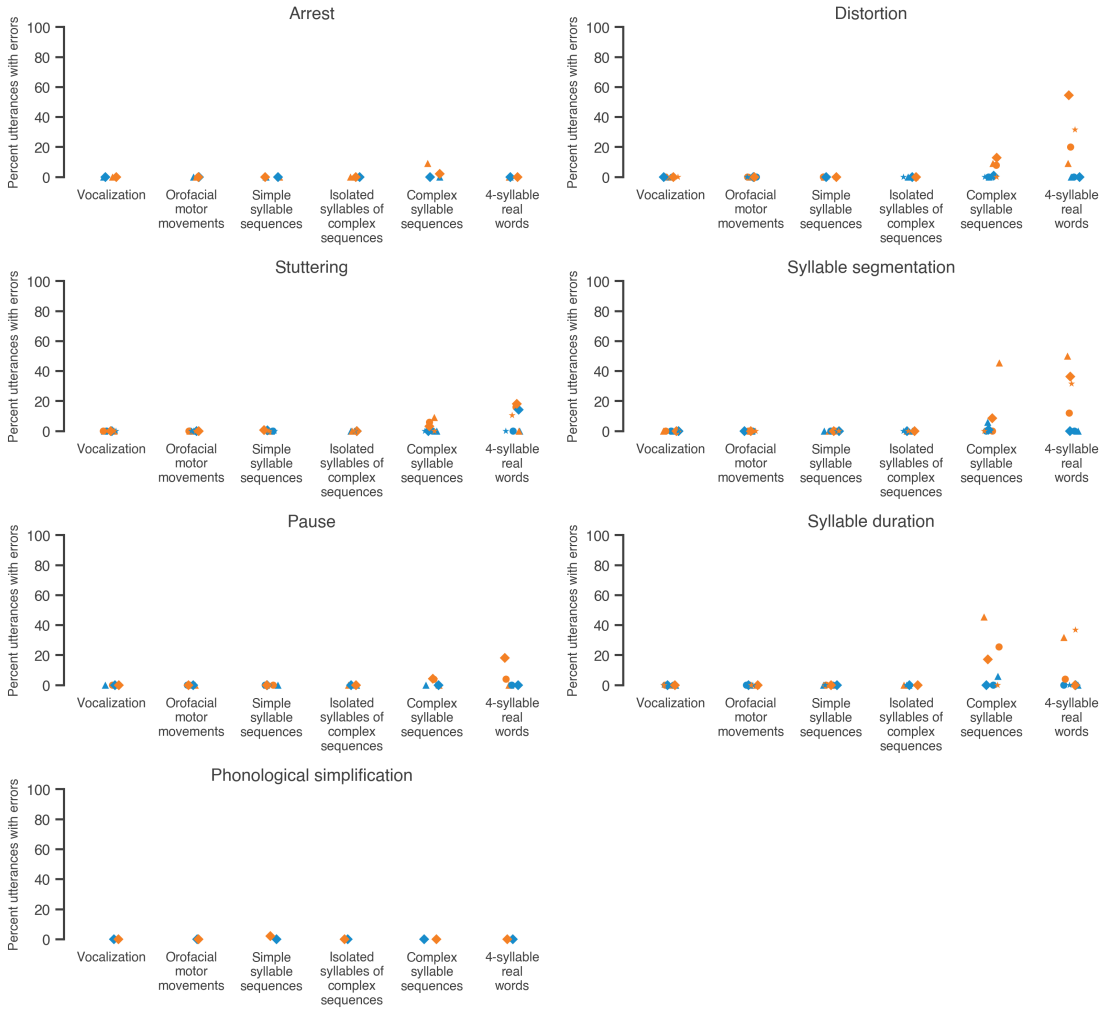


Figure 3.16. Error type and frequency for all tasks and all error types. Each panel depicts the percent of utterances, without stimulation (blue) and during stimulation (orange), that had a certain type of speech error. Each patient is depicted with a different shape; triangles for EC260, circles for EC267, diamonds for EC276, and stars for EC282. For phonological simplification, this occurred in one patient (EC276) where they said "jec" instead of "ject" during one stimulation pulse.

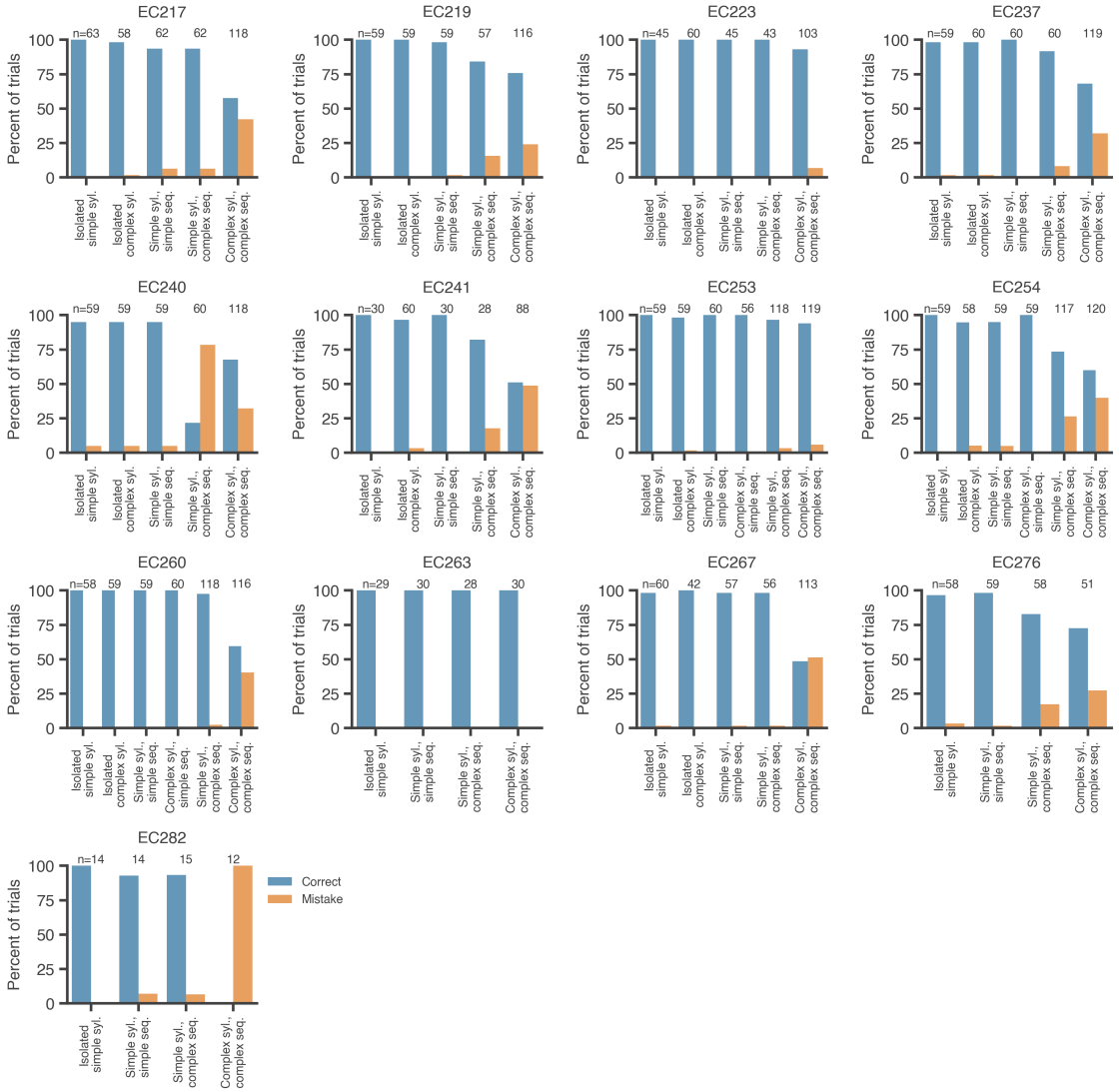


Figure 3.17. Percent correct and mistake trials for each participant. The total number of trials for each condition is labeled above the bars. Trials where there was extraneous speech at the start of the production (i.e. "oh sorry bababa" or "¡cough¿ badaga") were excluded from these counts. For each subject, the following number of trials were excluded from these counts: 5 for EC217, 8 for EC219, 4 for EC223, 2 for EC237, 4 for EC240, 4 for EC241, 9 for EC253, 8 for EC254, 10 for EC260, 3 for EC263, 32 for EC267, 13 for EC276, and 5 for EC282.

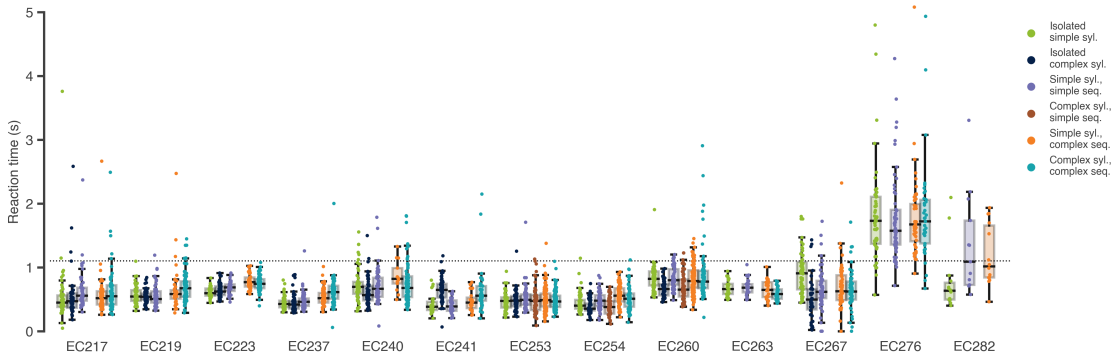


Figure 3.18. Reaction time distributions for each participant. Reaction times for each correct trial in each condition for each participant. Individual trials are shown as dots while the median and interquartile range is shown by box plots. The horizontal dashed line denotes the upper limit of a 1.1 second reaction time that was used to reject trials in some analyses. 1.1 seconds was chosen because this was applied to the pre-speech period where the window could be from the go-cue to 0.1 seconds before acoustic onset. This allowed for a maximum window of 1 second to be used in some analyses (such as sequence complexity encoding).

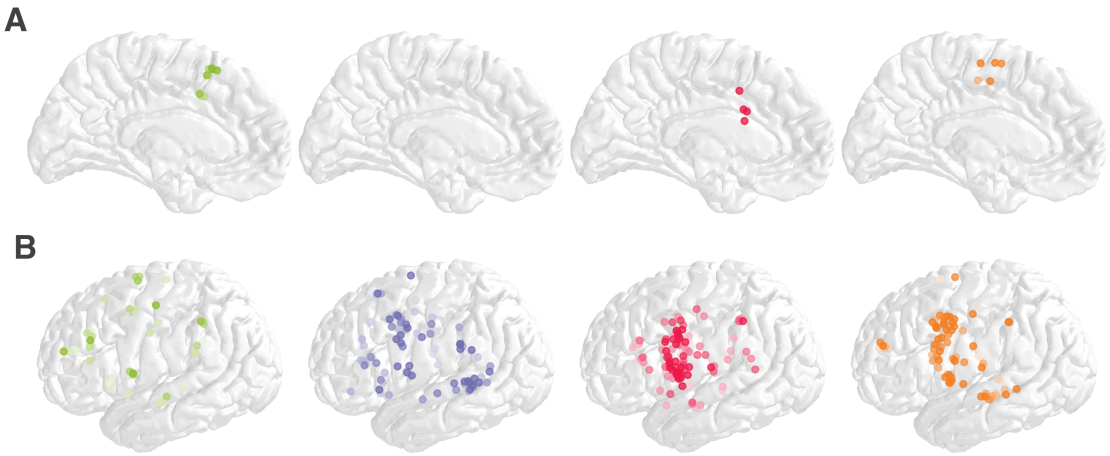


Figure 3.19. Spatial distributions of each sustained cluster. A. Spatial distribution of each temporal pattern of sustained activity found through unsupervised clustering (NMF) on the left medial surface of a common brain (MNI-152). **B.** Same as **A.** for the left lateral surface.

Table 3.1. Participant demographics, language experience, and analysis details. Languages are listed in order of acquisition, though all participants were fluent in English. Footnotes indicate what analyses each participant was included in.

Participant	Age	Gender	Fluent languages
EC217[1]	25	Female	English
EC219[1]	22	Female	English (Spanish-basic conversation)
EC223[1]	40	Female	English
EC237[1]	33	Female	Russian-English
EC240[1]	19	Male	English
EC241[1]	38	Male	English
EC253[1]	24	Male	English
EC254[1]	28	Male	English
EC260[2]	23	Male	Spanish-English
EC263[2]	24	Male	English
EC267[3]	44	Male	English-Spanish
EC276[4]	39	Male	Spanish-English
EC282[5]	22	Male	English-Spanish

[1] All analyses except stimulation.

[2] All analyses.

[3] Only reaction time prediction and stimulation. Participant's data was excluded from other analyses since they completed a version with an auditory stimulus instead of visual. Despite being excluded from the main figures, Figure 3.12 depicts the results of these analyses.

[4] All analyses, except for task phase decoding and pre-speech complexity encoding. Participant's reaction times were much greater and so they were excluded from analyses that averaged over the pre-speech period. Despite being excluded from some of the main figure panels, Figure 3.13 depicts the results of these analyses.

[5] Only stimulation analyses. Participant made errors on all complex sequences of complex syllables and so was excluded from most analyses using these trials. Despite being excluded from some of the main figure panels, Figure 3.14 depicts the results of these analyses.

Table 3.2. Utterance sets with varied sequence and articulatory complexity. Sequences in italics were only additionally collected only for participants EC253, EC254, and EC260 due to time constraints.

Syllable type (articulatory complexity)	Sequence complexity		
	Isolated	Simple	Complex
CV (simple)	baa	baa-baa-baa	baa-daa-gaa
	daa	daa-daa-daa	daa-baa-gaa
	gaa	gaa-gaa-gaa	gaa-daa-baa
			baa-gaa-daa
			daa-gaa-baa
			gaa-baa-daa
CCV (complex)	blaa	*blaa-blaa-blaa*	blaa-draa-gloo
	draa	*draa-draa-draa*	blaa-gloo-draa
	gloo	*gloo-gloo-gloo*	draa-blaa-gloo
			draa-gloo-blaa
			gloo-blaa-draa
			gloo-draa-blaa

Table 3.3. Auditory and articulatory datasets used to find significant auditory responses and fit articulatory kinematic trajectory (AKT) models.

Participant	Auditory dataset	Articulatory dataset
EC217	Isolated words	Days of the week
EC219	TIMIT sentences[1]	Days of the week
EC223	CV pairs	Days of the week
EC237	TIMIT sentences	Syllable sequencing (f0 laryngeal variable not fitted[2])
EC240	Instructions to a task	Days of the week
EC241	CV pairs	Days of the week
EC253	TIMIT sentences	MOCHA sentences[1]
EC254	TIMIT sentences	MOCHA sentences
EC260	TIMIT sentences	MOCHA sentences
EC263	TIMIT sentences	Subset of TIMIT sentences
EC267	TIMIT sentences	MOCHA sentences
EC276	Subset of TIMIT sentences	Subset of TIMIT sentences

[1] TIMIT and MOCHA are both corpora of sentences.

[2] For EC237, the only available production dataset was the syllable sequencing task. Compared to natural speech there is much less pitch variation needed in this task, and so the laryngeal variable (f0) was unable to be fit well by the model and was then excluded.

Table 3.4. Description of all tasks used during stimulation. The description of each task includes the instructions given to the patient. The example utterances are a subset of the utterances that may have been used during testing.

Task	Description	Example utterances or movements
4-syllable real word repetition	Patient instructed to repeat the word said aloud by an experimenter. This task was used as a go/no-go test whether to further investigate a site for complex sequencing errors. Sometimes, the patient would be asked to repeatedly produce the word, and the experimenter would deliver stimulation at different times in order to determine the appropriate stimulation current amplitude.	"catastrophe", "interjection", "honeysuckle"
Complex syllable sequences	Similar instructions as the 4-syllable real word task. Patients were instructed to say the sequences as if they were a single word. Test sequences chosen for each patient were based on whether the patient could say the sequence at baseline (i.e. no stimulation) with no errors.	"badaga", "pataka", "blaadraagloo", "chuhshuhjuh"
Isolated syllables of complex sequences	Similar instructions as the 4-syllable real word task. Patients were instructed to say the sequences with each syllable in isolation. That is, they were holding a complex syllable sequence in working memory, but were producing them as isolated syllables.	"ba. . . da. . . ga"
Simple syllable sequences	Similar instructions as the 4-syllable real word task.	"bababa", "papapa", "tatata"
Orofacial motor movements	Patients were instructed to repeatedly perform an orofacial movement. Three orofacial movements were tested for each patient.	Lip pucker, sticking tongue in and out, opening and closing jaw
Vocalization	Patients were instructed to hold a vowel for an extended period of time.	"ee", "oo"
Natural speech	Only tested with EC260. Patient was asked to describe their first job for a few minutes, while stimulation was applied during some sentences.	"I stayed after school"
Hand-motor sequences	Patients instructed to continually perform a complex hand movement sequence.	Using dominant hand, tap index finger to thumb, then middle finger to thumb, then ring, then pinky, then reverse and repeat.
Task switching	Only tested with EC260. Patient was asked to say the days of week repeatedly. Patient was instructed to monitor themselves for speech errors/difficulty. If they felt they were making errors, they were instructed to switch to a new task as soon as they could. Tasks they switched to included vocalization, orofacial movements, and mimed speech.	Example switching to vocalization: "monday tuesday wed-aaaah"

References

Andrews, John P., Nathan Cahn, Benjamin A. Speidel, et al. (Aug. 1, 2022). "Dissociation of Broca's area from Broca's aphasia in patients undergoing neurosurgical resections". *Journal of Neurosurgery*, pp. 1–11. ISSN: 0022-3085, 1933-0693. DOI: 10.3171/2022.6.JNS2297.

Averbeck, B. B., M. V. Chafee, D. A. Crowe, and A. P. Georgopoulos (Oct. 1, 2002). "Parallel processing of serial movements in prefrontal cortex". *Proceedings of the National Academy of Sciences* 99.20, pp. 13172–13177. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.162485599.

Binder, Jeffrey R. (Dec. 15, 2015). "The Wernicke area: Modern evidence and a reinterpretation". *Neurology* 85.24, pp. 2170–2175. ISSN: 0028-3878, 1526-632X. DOI: 10.1212/WNL.0000000000002219.

Binder, Jeffrey R. (Aug. 2017). "Current Controversies on Wernicke's Area and its Role in Language". *Current Neurology and Neuroscience Reports* 17.8, p. 58. ISSN: 1528-4042, 1534-6293. DOI: 10.1007/s11910-017-0764-8.

Bohland, Jason W. and Frank H. Guenther (Aug. 2006). "An fMRI investigation of syllable sequence production". *NeuroImage* 32.2, pp. 821–841. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2006.04.173.

Bouchard, Kristofer E., Nima Mesgarani, Keith Johnson, and Edward F. Chang (Mar. 2013). "Functional organization of human sensorimotor cortex for speech articulation". *Nature*

495.7441, pp. 327–332. ISSN: 1476-4687 (Electronic)\r0028-0836 (Linking). DOI: 10.1038/nature11911.

Castellucci, Gregg A., Christopher K. Kovach, Matthew A. Howard, et al. (Feb. 3, 2022). "A speech planning network for interactive language use". *Nature* 602.7895, pp. 117–122. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-021-04270-z.

Chang, Edward F., Garret Kurteff, John P. Andrews, et al. (Sept. 1, 2020). "Pure Apraxia of Speech After Resection Based in the Posterior Middle Frontal Gyrus". *Neurosurgery* 87.3, E383–E389. ISSN: 0148-396X, 1524-4040. DOI: 10.1093/neuros/nyaa002.

Chartier, Josh, Gopala K. Anumanchipalli, Keith Johnson, and Edward F. Chang (2018). "Encoding of Articulatory Kinematic Trajectories in Human Speech Sensorimotor Cortex". *Neuron* 98.5, 1042–1054.e4. DOI: 10.1016/j.neuron.2018.04.031.

Cheung, Connie, Liberty S Hamilton, Keith Johnson, and Edward F Chang (Mar. 4, 2016). "The auditory representation of speech sounds in human motor cortex". *eLife* 5, e12577. ISSN: 2050-084X. DOI: 10.7554/eLife.12577.

Cogan, Gregory B, Asha Iyer, Lucia Melloni, et al. (Feb. 2017). "Manipulating stored phonological input during verbal working memory". *Nature Neuroscience* 20.2, pp. 279–286. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/nn.4459.

Cogan, Gregory B., Thomas Thesen, Chad Carlson, et al. (Mar. 2014). "Sensory–motor transformations for speech occur bilaterally". *Nature* 507.7490, pp. 94–98. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature12935.

Dehaene, Stanislas, Lionel Naccache, Laurent Cohen, et al. (July 2001). "Cerebral mechanisms of word masking and unconscious repetition priming". *Nature Neuroscience* 4.7. Number: 7 Publisher: Nature Publishing Group, pp. 752–758. ISSN: 1546-1726. DOI: 10.1038/89551.

Dichter, Benjamin K., Jonathan D. Breshears, Matthew K. Leonard, and Edward F. Chang (2018). "The control of vocal pitch in human laryngeal motor cortex". *Cell* 174.1, pp. 1–11. DOI: 10.1016/j.cell.2018.05.016.

Eichert, Nicole, Daniel Papp, Rogier B Mars, and Kate E Watkins (Nov. 3, 2020). "Mapping Human Laryngeal Motor Cortex during Vocalization". *Cerebral Cortex* 30.12, pp. 6254–6269. ISSN: 1047-3211. DOI: 10.1093/cercor/bhaa182.

Flinker, Adeen, Anna Korzeniewska, Avgusta Y. Shestyuk, et al. (Mar. 3, 2015). "Redefining the role of Broca's area in speech". *Proceedings of the National Academy of Sciences* 112.9, pp. 2871–2875. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1414491112.

Gnadt, J.W. and Richard Alan Andersen (1988). *Memory related motor planning activity in posterior parietal cortex of macaque.*

Graff-Radford, Jonathan, David T. Jones, Edythe A. Strand, et al. (Feb. 2014). "The neuroanatomy of pure apraxia of speech in stroke". *Brain and Language* 129, pp. 43–46. ISSN: 0093934X. DOI: 10.1016/j.bandl.2014.01.004.

Guenther, Frank H. and Gregory Hickok (2016). "Neural Models of Motor Speech Control". *Neurobiology of Language*. Elsevier, pp. 725–740. ISBN: 978-0-12-407794-2.

Guo, Zengcai V., Hidehiko K. Inagaki, Kayvon Daie, et al. (May 2017). "Maintenance of persistent activity in a frontal thalamocortical loop". *Nature* 545.7653, pp. 181–186. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature22324.

Hamilton, Liberty S., Erik Edwards, and Edward F. Chang (June 2018). "A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus". *Current Biology* 28.12, 1860–1871.e4. ISSN: 09609822. DOI: 10.1016/j.cub.2018.04.033.

Hickok, Gregory (Feb. 2012). "Computational neuroanatomy of speech production". *Nature Reviews Neuroscience* 13.2, pp. 135–145. ISSN: 1471-003X, 1471-0048. DOI: 10.1038/nrn3158.

Hickok, Gregory, Jonathan Venezia, and Alex Teghipco (Nov. 30, 2022). "Beyond Broca: neural architecture and evolution of a dual motor speech coordination system". *Brain*, awac454. ISSN: 0006-8950, 1460-2156. DOI: 10.1093/brain/awac454.

Inagaki, Hidehiko K., Susu Chen, Kayvon Daie, et al. (July 8, 2022). "Neural Algorithms and Circuits for Motor Planning". *Annual Review of Neuroscience* 45.1, pp. 249–271. ISSN: 0147-006X, 1545-4126. DOI: 10.1146/annurev-neuro-092021-121730.

Itabashi, Ryo, Yoshiyuki Nishio, Yuka Kataoka, et al. (Jan. 2016). "Damage to the Left Precentral Gyrus Is Associated With Apraxia of Speech in Acute Stroke". *Stroke* 47.1, pp. 31–36. ISSN: 0039-2499, 1524-4628. DOI: 10.1161/STROKEAHA.115.010402.

Jordan, Michael I. (1997). "Serial Order: A Parallel Distributed Processing Approach". *Advances in Psychology*. Vol. 121. Elsevier, pp. 471–495. ISBN: 978-0-444-81931-4. DOI: 10.1016/S0166-4115(97)80111-2.

Kaestner, Erik, Xiaojing Wu, Daniel Friedman, et al. (Feb. 10, 2022). "The Precentral Gyrus Contributions to the Early Time-Course of Grapheme-to-Phoneme Conversion". *Neurobiology of Language* 3.1, pp. 18–45. ISSN: 2641-4368. DOI: 10.1162/nol_a_00047.

Kornysheva, Katja, Daniel Bush, Sofie S. Meyer, et al. (Mar. 2019). "Neural Competitive Queuing of Ordinal Structure Underlies Skilled Sequential Action". *Neuron* 101.6, 1166–1180.e3. ISSN: 08966273. DOI: 10.1016/j.neuron.2019.01.018.

Kornysheva, Katja, Anika Sierk, and Jörn Diedrichsen (Mar. 1, 2013). "Interaction of temporal and ordinal representations in movement sequences". *Journal of Neurophysiology* 109.5, pp. 1416–1424. ISSN: 0022-3077, 1522-1598. DOI: 10.1152/jn.00509.2012.

Lashley, Karl S. (1951). "The Problem of Serial Order in Behavior". *Cerebral mechanisms in behavior; the Hixon Symposium*, pp. 112–146.

Leonard, Matthew K., Ruofan Cai, Miranda C. Babiak, et al. (June 2019). "The peri-Sylvian cortical network underlying single word repetition revealed by electrocortical stimulation and direct neural recordings". *Brain and Language* 193, pp. 58–72. ISSN: 0093934X. DOI: 10.1016/j.bandl.2016.06.001.

Levelt, Willem J. M. (1993). *Speaking: From Intention to Articulation*. The MIT Press. ISBN: 978-0-262-27822-5. DOI: 10.7551/mitpress/6393.001.0001.

Levy, Deborah F., Alexander B. Silva, Terri L. Scott, et al. (Mar. 27, 2023). "Apraxia of speech with phonological alexia and agraphia following resection of the left middle precentral gyrus: illustrative case". *Journal of Neurosurgery: Case Lessons* 5.13, CASE22504. ISSN: 2694-1902. DOI: 10.3171/CASE22504.

Long, Michael A. and Michale S. Fee (Nov. 2008). "Using temperature to analyse temporal dynamics in the songbird motor pathway". *Nature* 456.7219, pp. 189–194. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature07448.

Lu, Junfeng, Zehao Zhao, Jie Zhang, et al. (Sept. 4, 2021). "Functional maps of direct electrical stimulation-induced speech arrest and anomia: a multicentre retrospective study". *Brain* 144.8, pp. 2541–2553. ISSN: 0006-8950, 1460-2156. DOI: 10.1093/brain/awab125.

MacKay, Donald G. (July 1970). "Spoonerisms: The structure of errors in the serial order of speech". *Neuropsychologia* 8.3, pp. 323–350. ISSN: 00283932. DOI: 10.1016/0028-3932(70)90078-3.

MacNeilage, Peter F. (Aug. 1998). "The frame/content theory of evolution of speech production". *Behavioral and Brain Sciences* 21.4, pp. 499–511. ISSN: 0140-525X, 1469-1825. DOI: 10.1017/S0140525X98001265.

Mohr, J. P., M. S. Pessin, S. Finkelstein, et al. (Apr. 1, 1978). "Broca aphasia: Pathologic and clinical". *Neurology* 28.4, pp. 311–311. ISSN: 0028-3878, 1526-632X. DOI: 10.1212/WNL.28.4.311.

Mugler, Emily M., Matthew C. Tate, Karen Livescu, et al. (2018). "Differential Representation of Articulatory Gestures and Phonemes in Precentral and Inferior Frontal Gyri". *The Journal of Neuroscience* 4653, pp. 1206–18. DOI: 10.1523/JNEUROSCI.1206-18.2018.

Peeva, Maya G., Frank H. Guenther, Jason A. Tourville, et al. (Apr. 2010). "Distinct representations of phonemes, syllables, and supra-syllabic sequences in the speech production network". *NeuroImage* 50.2, pp. 626–638. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2009.12.065.

Pillay, Sara B., Benjamin C. Stengel, Colin Humphries, et al. (Nov. 2014). "Cerebral localization of impaired phonological retrieval during rhyme judgment: Phonological Retrieval". *Annals of Neurology* 76.5, pp. 738–746. ISSN: 03645134. DOI: 10.1002/ana.24266.

Quigg, Mark, David S. Geldmacher, and W. Jeff Elias (May 2006). "Conduction aphasia as a function of the dominant posterior perisylvian cortex: Report of two cases". *Journal of Neurosurgery* 104.5, pp. 845–848. ISSN: 0022-3085. DOI: 10.3171/jns.2006.104.5.845.

Ray, Supratim and John H.R. Maunsell (Sept. 2010). "Differences in Gamma Frequencies across Visual Cortex Restrict Their Possible Use in Computation". *Neuron* 67.5, pp. 885–896. ISSN: 08966273. DOI: 10.1016/j.neuron.2010.08.004.

Rong, Feng, A. Lisette Isenberg, Erica Sun, and Gregory Hickok (Oct. 9, 2018). "The neuroanatomy of speech sequencing at the syllable level". *PLOS ONE* 13.10. Ed. by Claude Alain, e0196381. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0196381.

Silva, Alexander B., Jessie R. Liu, Lingyun Zhao, et al. (Nov. 9, 2022). "A Neurosurgical Functional Dissection of the Middle Precentral Gyrus during Speech Production". *The*

Journal of Neuroscience 42.45, pp. 8416–8426. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.1614-22.2022.

Steinschneider, M., Y. I. Fishman, and J. C. Arezzo (Mar. 1, 2008). "Spectrotemporal Analysis of Evoked and Induced Electroencephalographic Responses in Primary Auditory Cortex (A1) of the Awake Monkey". *Cerebral Cortex* 18.3, pp. 610–625. ISSN: 1047-3211, 1460-2199. DOI: 10.1093/cercor/bhm094.

Strand, Edythe A., Joseph R. Duffy, Heather M. Clark, and Keith Josephs (Sept. 2014). "The apraxia of speech rating scale: A tool for diagnosis and description of apraxia of speech". *Journal of Communication Disorders* 51, pp. 43–50. ISSN: 00219924. DOI: 10.1016/j.jcomdis.2014.06.008.

Svoboda, Karel and Nuo Li (Apr. 2018). "Neural mechanisms of movement planning: motor cortex and beyond". *Current Opinion in Neurobiology* 49, pp. 33–41. ISSN: 09594388. DOI: 10.1016/j.conb.2017.10.023.

Venezia, Jonathan H., Virginia M. Richards, and Gregory Hickok (Sept. 1, 2021). "Speech-Driven Spectrotemporal Receptive Fields Beyond the Auditory Cortex". *Hearing Research* 408, p. 108307. ISSN: 0378-5955. DOI: 10.1016/j.heares.2021.108307.

Zimnik, Andrew J. and Mark M. Churchland (Feb. 22, 2021). "Independent generation of sequence elements by motor cortex". *Nature Neuroscience*. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/s41593-021-00798-5.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Jessie R. Liu
E6A28C566B5B497...
Author Signature

5/17/2023
Date