AI-driven brain-computer interfaces for speech

by

Sean Metzger

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:

Chang, Edward — Chang, Edward

4B5B40824E04415...

Chair

DocuSigned by:

Reza Abbasi-Asl — Reza Abbasi-Asl

D3[illegible]749D...

DocuSigned by:

Rikky Muller — Rikky Muller

A5734EB12B7F4D2...

Committee Members

Acknowledgments

Firstly, I want to say a huge thank you to my advisor, Dr. Edward Chang. Thank you for taking a chance on me. Being able to work with Eddie was an amazing experience and I saw how incredible he was at asking the right question and thinking hard about why you are doing something and why it matters before you even start. His enthusiasm and excitement about the work in his lab is contagious and was a persistent source of motivation. Eddie is by far the hardest worker I know and he does it all because he's so excited about each project. He is inspiring and I can't wait to see what he does next.

Secondly, this dissertation would not be possible without the incredible dedication and bravery of our clinical trial participants. The BRAVO participants redefined what generosity means for me, as their sacrifice of having invasive brain surgery and giving up hundreds of hours of their time to work with us was incredible. Without their dedication - without a surgery, without thousands of repeats of sentences and words, and without their excitement when the system worked, this thesis would not exist, so I want to say a huge thank you. Getting to spend a day working with them was always the highlight of any week, and their smiles after we got something to work made the long hours worth it all. Hopefully the work in this dissertation can be further improved to be truly use-able at all times and give them their voice back.

I also want to thank my dissertation committee, Rikky Muller and Reza Abbasi-Asl, for their guidance and help not only while writing this dissertation but during the preparation

for my qualifying exam as well. Much of the material from that qualifying exam is part of Chapter 2 of this thesis.

I next want to thank my amazing colleagues across the lab who constantly inspired me with their brilliance Firstly, to everyone on the BRAVO team past and present; Margaret, Jessie, David, Kaylo, Alex, Max, Josh, Gopala, Joe, Pengfei, and Ran. Working with you was truly special and I feel lucky to have had such a talented and hardworking group of teammates. It was also an honor to work with Laura, Sarah, Many, Lingyun, Kristin, and Emily across many exciting projects that were not included in this dissertation. I also really want to thank Ilina, Terri, and Deb for their companionship, analysis of the neural data of the bachelorette participants, and particularly for their desire to make the lab a great place to work. A special thanks to everyone else in the lab who made it a great place to be, especially Itzik! I also want to shout out a special non-Chang Lab colleague - Colorado Reed, who showed true brilliance during our final project during deep unsupervised learning, and taught me a lot about structuring workflows, managing projects and doing high-impact work - thank you!

I also want to acknowledge the amazing staff of the Chang Lab that made life easier and helped us a lot - Todd, Viv, Robin, Ilona, all were amazing help throughout. I also want to say thanks to the BioE program administrators - Victoria and Rocio, and my graduate advisor Prof. Christoph Schneider. Thanks also to Zachary Knight for a great TA experience in NS221.

I also want to thank all my academic mentors and inspirations prior to my work here, at Stanford - Daniel Palanker, Tom Flores, Bill Dally, John Fox, and Mark Cappelli, as well as inspiring figures who I worked with as a young high school student - Mihaly Horanyi and Keith Julien at CU-Boulder, Drs. Petach and Silverman at Fairview, and Matt Greenberg and Bruce Edgerly at Backcountry Access.

Finally, this dissertation would not have been possible without an amazing support system outside of the lab. I want to thank Margaret Seaton, who has been the best partner anyone could ask for over the course of this challenging PhD, and who joined the lab and became the best colleague ever. I love you. I also want to thank my family - Andre, Mercedes, and Wyatt, for their incredible support and inspiration over the course of this journey. Spending time with them during COVID and their times in the bay has been truly special. I also want to thank my Grandma Patti, Grandma Mariana, Grandpa Ted, and all my aunts and uncles for their support and passion. Grandpa - I finally wrote up the thesis like you asked me to!

I finally want to thank my friends for their companionship over these years. A special shoutout to the 6 brew crew and gravel chad summer, as well as my special roommates. I want to give a special shoutout in particular to Ben, Daniel, Jorge, Cory, Gabe, Alaena, Chase, Thomas, Martin, Chris, Peter, Grant, Taylor, Connor, Nora, Skye, and Bronson, as well as everyone else who brightened my days beyond belief. I also want to specially shout out my academically minded inspirations - Chris Mathy, Sarah and Matt H, as well as my academic inspirations on our IM soccer team in Berkeley - Your GSI - that came up tragically short in not one, not two, but three IM finals.

Finally, I want to thank my bikes for keeping me sane and making me want to push my limits, and all the people I had the blessing to share a day in the saddle with on a foggy early morning in one of the most beautiful places to ride! There's nothing better.

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.

Chapter 3 is directly adapted from:

> **Sean L. Metzger***, Kaylo T. Littlejohn*, Alex B. Silva*, David A. Moses*, Margaret P. Seaton* et. al. (2023). A high performance neuroprosthesis for speech decoding and avatar control. *Nature*, 1-10. doi: 10.1038/s41586-023-06443-4

* Denotes equal contributions.

Personal contributions are further detailed at the start of each chapter with a disclaimer concerning previous or future publication.

Abstract

AI-driven brain-computer interfaces for speech

Sean L. Metzger

Speech is a fundamental human behavior, and enables the fast, effortless expression of emotions, desires, and needs. Devastating conditions like paralysis and brain-stem stroke can rob individuals of the ability to speak, even though they retain intact cognitive abilities. Brain-Computer Interface (BCI) technology offers hope for such individuals by reading out these intact neural signals using a recording device, then deciphering what the person was trying to say using machine learning and artificial intelligence. Prior to beginning this thesis, many questions remained in the development of speech-BCIs. Could speech be decoded from the brain of a person who hadn't spoken for many years and was suffering from paralysis? What algorithms could do this? What recording technologies could be used? Would the brain activity look similar to healthy speakers, or would it have evolved? Could we enable someone to speak quickly and naturalistically with these devices?

With these questions in mind, we launched the BCI restoration of arm and voice (BRAVO) clinical trial. This trial explores the use of Electrocorticography (ECoG), a high-resolution neural recording modality, to read out and decode neural activity during intended speech. This thesis presents results on work I have done as part of this clinical trial, and demonstrates successful speech decoding with two incredible participants who have each been unable to

speak for over 15 years.

Chapter 1 introduces a proof-of-concept speech BCI in someone who cannot speak, showcasing real-time decoding of 50 words and restoring communication at 15 words per minute. Chapter 2 expands the scope of speech BCIs, enabling speech-based spelling with NATO codewords, achieving 29 characters per minute and allowing communication with a practically unlimited vocabulary. We also show silent-speech attempts can be decoded, and that low-frequency neural activity (between .6 and 16.67 Hz) is critical to decoding. Chapter 3 introduces a state-of-the-art BCI which can decode text from neural activity during silent speech attempts at 80 words per minute. Sound units and facial gestures were also predictable, enabling auditory and visual representations of attempted speech. These advancements illustrate the potential of speech BCIs to restore communication for those who have lost their voice, and demonstrated that speech representations are maintained for years even after paralysis.

Taken together, this research presents progress in the development of speech BCIs and signifies a significant step towards improving the lives of individuals with communication impairments through versatile and comprehensive speech restoration systems.

Contents

List of Figures

List of Tables

Introduction

Speech is a unique and defining characteristic of the human species and serves as our primary mode of communication. It enables us to express our desires, emotions, and needs. It lets us interact with others and form bonds and relationships, and is the default mode of communication we use in our day to day lives. It is a beautiful and intricate motor process that requires the brain to control and coordination of hundreds of muscles in our vocal tract in order to shape breath into words and sentences.

Unfortunately, many individuals lose their ability to speak due to devastating conditions such as brain-stem stroke, paralysis, and amyotrophic lateral sclerosis (ALS). Despite this loss, many of these individuals retain intact cognitive abilities and brain signals necessary for speech production. It is only because of disrupted pathways between the brain and muscles that the neural commands for speech generation do not reach the vocal tract. While some individuals with these conditions can resort to communication methods like eye-tracking based spelling, these solutions can be slow and cumbersome, significantly reducing quality of life.(Felgoise et al. 2016)

Brain-Computer Interface (BCI) technology has emerged as a promising approach to restore communication for people with these conditions by circumventing damaged neural pathways. Fundamentally, a BCI functions by directly decoding neural activity into intended messages (Brumberg et al. 2018). The majority of intracortical BCIs that have been demonstrated focus on decoding neural activity related to hand and arm movements into a binary click (Vansteensel et al. 2016), cursor control to select letters one by one, Pandarinath et al. 2017 or handwritten letters (Willett et al. 2021) to drive spelling-based devices. While promising, these devices lag far behind natural rates of communication from speech, achieving rates of at best 18 words per minute, whereas natural speech enables communication at 150 words per minute (8x higher) (Edward F. Chang and Anumanchipalli 2019). Non-invasive approaches also show great promise - for example visually-based P300 approaches Nijboer, Plass-Oude Bos, et al. 2014 were also promising. A full comparison of device communication rates with paralyzed participants appears in Table 0.1. In addition to their slower rates of communication, these devices are notably less naturalistic to use than a device where you would simply speak the words you are intend to say.

Thus, a more ideal BCI would translate intact brain signals directly into intended speech. However, the development of such technology presents several challenges, including the need for a neural recording technology that enables precise readout of neural signals, selecting the right brain areas to record from, and the development of high-performance decoding strategies.

One potential recording modality for a speech BCI is Electrocorticography (ECoG), a method that provides high-quality neural signals with excellent temporal resolution across extensive areas of the motor cortex (Edward F Chang 2015). Because it can record directly from the brain's surface, ECoG holds promise in decoding complex neural activity associated with speech control that non-invasive recording modalities may not be able to capture. ECoG electrodes are also relatively large in size (1mm in diameter) and sit on the surface of the brain, making them less invasive than recording technologies that penetrate into brain tissue and less prone to instability compared to other devices like the Utah array which record with higher spatial precision but are susceptible to disruptions in daily use due to its much smaller electrodes shifting (Pandarinath et al. 2017; Willett et al. 2021). The combination of high temporal and spatial resolution, along with its potential for stable chronic recordings make ECoG a promising recording modality for BCIs.

Another crucial question is where to record neural activity to facilitate speech restoration. The ventral sensorimotor cortex (vSMC) is a critical area involved in articulation control (Chartier et al. 2018; Carey et al. 2017; Bouchard et al. 2013), known to contain cortical representations that can be leveraged to decode intended speech from brain signals in healthy speakers with ECoG implanted for the treatment of intractable epilepsy (Makin et al. 2020; Anumanchipalli et al. 2019). Prior to this thesis, it remained uncertain whether these representations persisted in individuals with long-term paralysis.

With these questions in mind, our lab initiated the BRAVO (BCI Restoration of Arm and Voice) trial in 2019, focusing on studying a chronic ECoG-based speech BCI for participants unable to speak. This thesis presents a series of results from this trial, which have laid the groundwork towards a clinically viable speech neuroprosthesis and shown the effectiveness of ECoG as a recording modality for speech brain-computer interfaces.

Chapter 1 presents a proof-of-concept study from the BRAVO trial, showcasing the successful implementation of a speech BCI in an individual with severe paralysis and anarthria caused by a brainstem stroke. The clinical-trial participant highlighted in this work, BRAVO-1, had a brainstem stroke following a car accident over 15 years ago. This accident left him unable to speak or type to communicate and prior to starting the trial, he used small residual neck movements to type out messages using a stylus attached to a baseball cap on his head at a painstaking rate of 5 words per minute. In early 2019, Bravo-1 underwent surgery to be implanted with a 128-electrode ECoG grid. The grid was implanted to maximize coverage of the vSMC and was attached to a pedestal that was surgically embedded in his skull that enabled chronic recording with the array.

We then recorded neural activity while Bravo-1 attempted to articulate 50 English words aloud. These 50 words were selected to be highly usable and important in clinical care setting and could be combined to form over 1,000 sentences. Next, we trained two artificial neural network models using Bravo-1's neural data. We first created a speech-detection model that could distinguish when Bravo-1 was attempting to speak vs not. Then we trained a classifier to predict which of the 50 words Bravo-1 was saying during a speech attempt. Using these models, we created a real-time system that could detect when he was attempting to speak with the speech-detection model, then predict the probability of each of the 50 words given the neural activity during the speech attempt with the classifier. We accumulated sequences of these word predictions as BRAVO-1 attempted to string together words in a sentence. Natural language processing algorithms infused the sequences of predictions with the statistics of english to produce the most lingustically probable sentence, given the model predictions. We were able to use this system in real-time to decode sentences composed of the 50 words at 15 words per minute and with 25% word error rate. These results highlighted the persistence of neural signals related to attempted articulation even after long-term paralysis and demonstrated that it was possible to restore communication via speech decoding in someone with paralysis and anarthria.

Chapter 2 builds on the technology developed in Chapter 1 by developing a spelling BCI driven by speech attempts. Although chapter 1 served as an important proof of concept, its practical use was quite limited, as Bravo-1 could only speak using words that fell within the 50-word vocabulary. We hence devised a spelling system that decoded NATO codewords (e.g. Alpha for A, Bravo for B and so forth) from neural activity as BRAVO-1 attempted to say the NATO codewords. The idea was that sequences of these codewords could be combined to spell out any word in the English language (e.g. Charlie-Alpha-Tango for cat), greatly expanding the vocabulary BRAVO-1 could use to express himself.

In this work, we also showed it was possible to decode the NATO codewords during silent-speech attempts, which involve attempted mouth movement without vocalization, rather than 'overt' speech attempts which invoked our participants residual vocalization as in our previous works. Silent-speech attempts are advantageous since they are less effortful to complete and more discreet than overt speech attempts. However, we found they were slightly more difficult to decode than 'overt' speech attempts. To overcome this challenge, we expanded the neural features used during decoding to include both low-frequency signals and high-gamma activity. We found that important and unique spatial and temporal information was present in low-frequency features, and that these features enabled high-accuracy decoding of the NATO codewords.

We then used this decoding to drive a real-time speech-driven spelling system that successfully decoded sentences with 94% character level accuracy using a large vocabulary of over 1,000 words. Offline simulations indicated that limiting words decoded with this system to an even larger vocabulary of 9,000 words did not significantly degrade performance. Importantly, our system operated at 29 characters per minute, 70% faster than our participants communication prior to the study (17 characters per minute). While this speed corresponded to a communication rate of 7 words per minute that was slower than our previous direct word decoding approach at 7 words , an advantage of this approach relative to spelling BCIs driven by hand activity is that it could naturally be paired with a direct word-decoding approach, making for a fast and generalizable BCI experience.

In summary, chapter 2 demonstrates that speech BCIs can be used with large vocabularies of over 1,000 words and that silent-speech attempts are a viable recording modality. Its high accuracy and expressive decoding established that speech BCIs could be a much more clinically viable pathway for restoring communication.

In 2022, we enrolled another participant into our clinical trial. This participant, similar to our first, had suffered a tragic brainstem stroke and had not been able to speak for 18 years. This heartbreakingly left her completely unable to vocalize and move. She enrolled in the BRAVO trial and underwent surgical implantation of an ECoG array. In the time between our first implantand the second, recording technology had improved such that this array had nearly 2x as many electrodes as BRAVO-1's array and the electrodes were much closer together (3mm spacing vs 4mm with BRAVO-1). This enabled more detailed recordings than we had with BRAVO-1, and we asked if we could see better results than what we had seen with BRAVO-1

Chapter 3 demonstrated that recordings from this grid could be used to develop a state-of-the-art speech BCI. In this chapter, we trained artificial neural networks to decode sequences of phonemes from neural activity as our participant attempted to silently speak sentences scraped from tweets and movie dialogues. The decoded sequences of phonemes could then be assembled to form words. A significant challenge in training models with this approach is that we could never know at which timepoint our participant was trying to say each phoneme, since they cannot speak. We addressed this problem by using the connectionist temporal classification loss, which enables the training of decoding algorithms without alignment between the neural activity and the labels being used for training. This approach led us to achieve speeds of nearly 80 words per minute during real-time decoding as our participant produced sentences composed of words from a vocabulary of over 1,000 words with high accuracy.

While this represents the state-of-the art in communication rates achieved with a brain-computer interface, speech-based communication relies on many elements beyond just the text of what we are saying. Sound can convey tone and emphasis in what we are trying to express. Hence, we modified our system to predict sound units derived from self-supervised learning instead of phones. These sound-units could then be combined to synthesize an intelligible auditory representation of the speech that the participant was attempting to say. Similarly, facial gestures play a large role in communication with others, as they can express emotion and reactions to speech in addition to carrying non-verbal meaning. We therefore further modified our system to produce discretized movement units that could then be used to reconstruct the continuous articulatory movements the participant was attempting to produce. These movements could then be used as inputs to a 3d avatar, which produced life-like movements that carried information about what the participant was trying to say and were strongly correlated with neurotypical speaker's orofacial movements during speech.

Overall, chapter 3 represents the culmination of years of research in the BRAVO trial, as we leaned on the research in chatpers 1 and 2 demonstrating that speech and silently attempted speech could be decoded, and used many of the signal processing techniques learned along the way in this chapter. More importantly, chapter 3 demonstrates the true promise of speech BCIs - that they can be used to restore communication using the natural way we communicate at far higher speeds than handwriting driven BCIs.

Together, these chapters present an important series of developments in the development of Brain-Computer Interface technology for speech restoration in individuals with communication impairments and expands the possibilities for improving quality of life for those affected by losing the ability to speak. The integration of direct speech and spelling BCI approaches, alongside novel phoneme decoding techniques, holds promise for realizing a comprehensive and versatile speech restoration system, bringing us one step closer to bridging the communication divide for the many people around the world who suffer from the devastating loss of the ability to speak.

Table 0.1. Comparison of BCI communication speeds for participants with paralysis.

Study	Recording modality	Words per minute	Control Strategy
Willett et al. 2021	Utah array	18	Handwriting
Pandarinath et al. 2017	Utah Array	6.3	Cursor control
Nijboer, Sellers, et al. 2008	EEG	0.31-0.82	Visual spelling
Vansteensel et al. 2016	ECoG	0.23	Movement attempts

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.

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.

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.

Chang, Edward F (2015). "Towards large-scale, human-based, mesoscopic neurotechnologies". *Neuron* 86.1, pp. 68–78.

Chang, Edward F. and Gopala K. Anumanchipalli (2019). "Toward a Speech Neuroprosthesis". en. *JAMA* 323.5, p. 413. ISSN: 0098-7484. DOI: 10.1001/jama.2019.19813.

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.

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.

Nijboer, Femke, Danny Plass-Oude Bos, Yvonne Blokland, et al. (Jan. 2014). "Design requirements and potential target users for brain-computer interfaces – recommendations from rehabilitation professionals". en. *Brain-Computer Interfaces* 1.1, pp. 50–61. ISSN: 2326-263X, 2326-2621. DOI: 10.1080/2326263X.2013.877210.

Nijboer, Femke, Eric W Sellers, Jürgen Mellinger, et al. (2008). "A P300-based brain–computer interface for people with amyotrophic lateral sclerosis". *Clinical neurophysiology* 119.8, pp. 1909–1916.

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.

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.

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.

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 50-word classification detection models and performed decoding performance, electrode contribution, and model stability analyses. With David A. Moses and Jessie. R. Liu, 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 and Roth 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 and Ney 1995; Chen and Goodman 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 and Krogh 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.

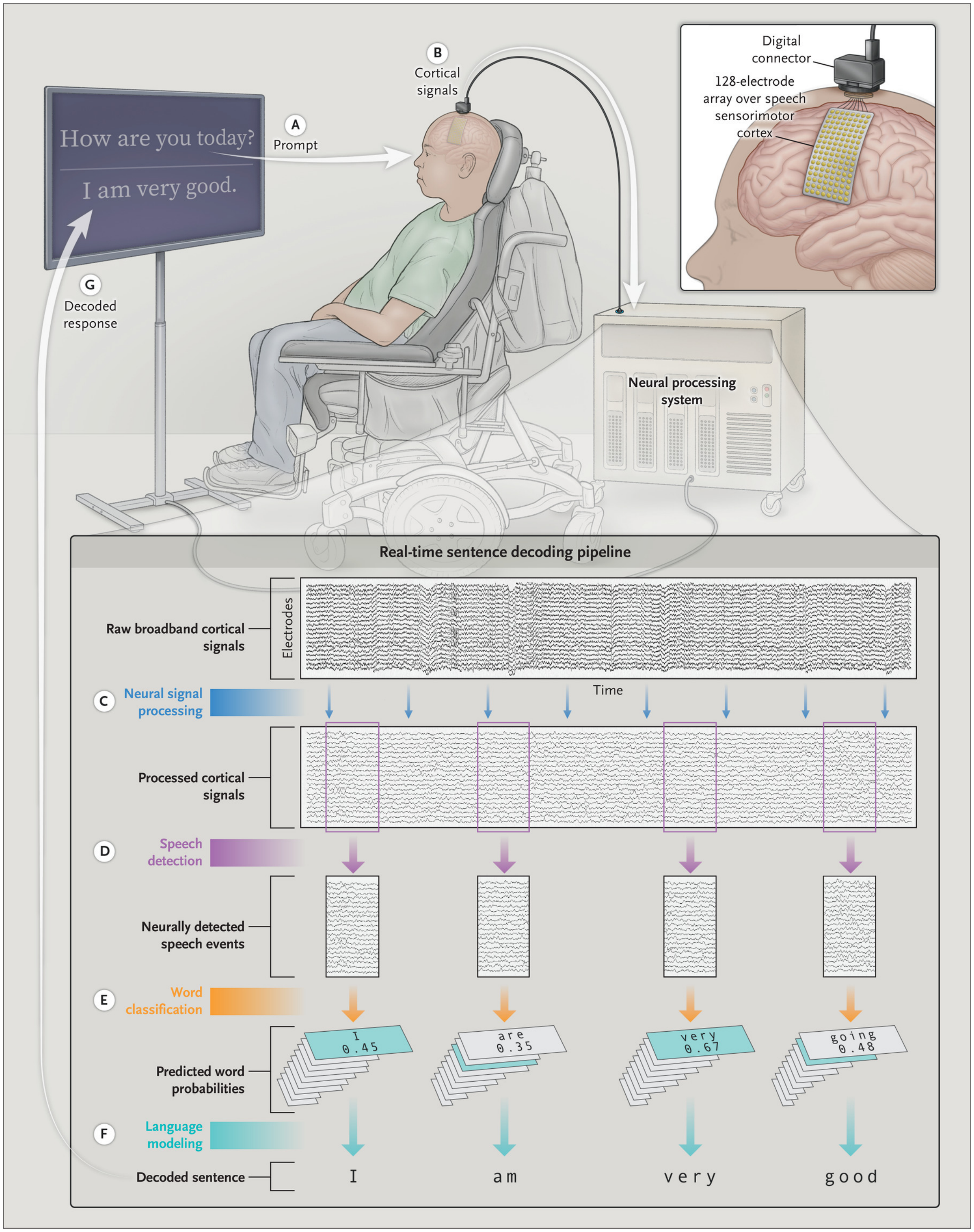
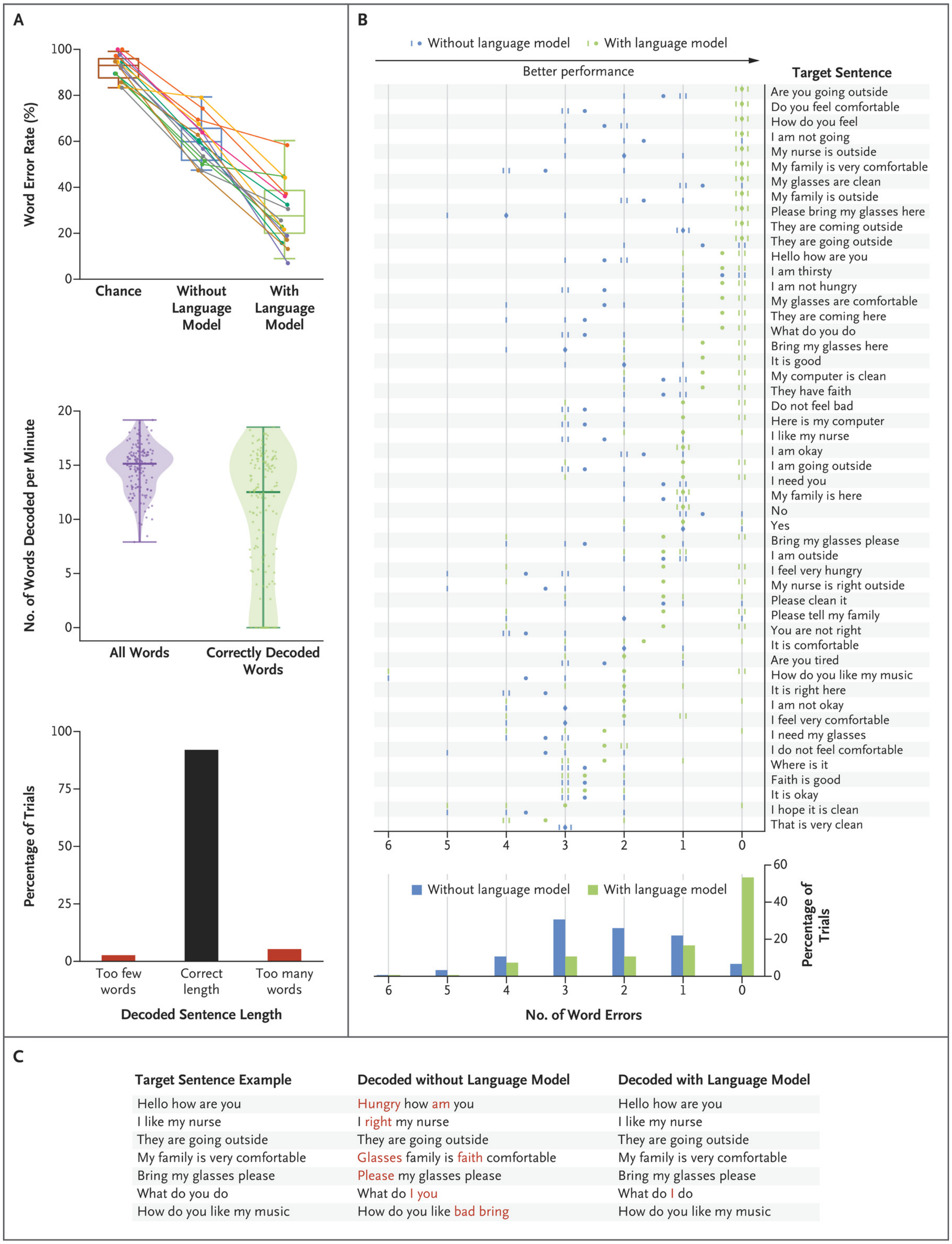


Figure 1.1. Schematic depiction of the decoding pipeline. (continued on next page).

(Previous page.) **Figure 1.1. Schematic depiction of the decoding 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.

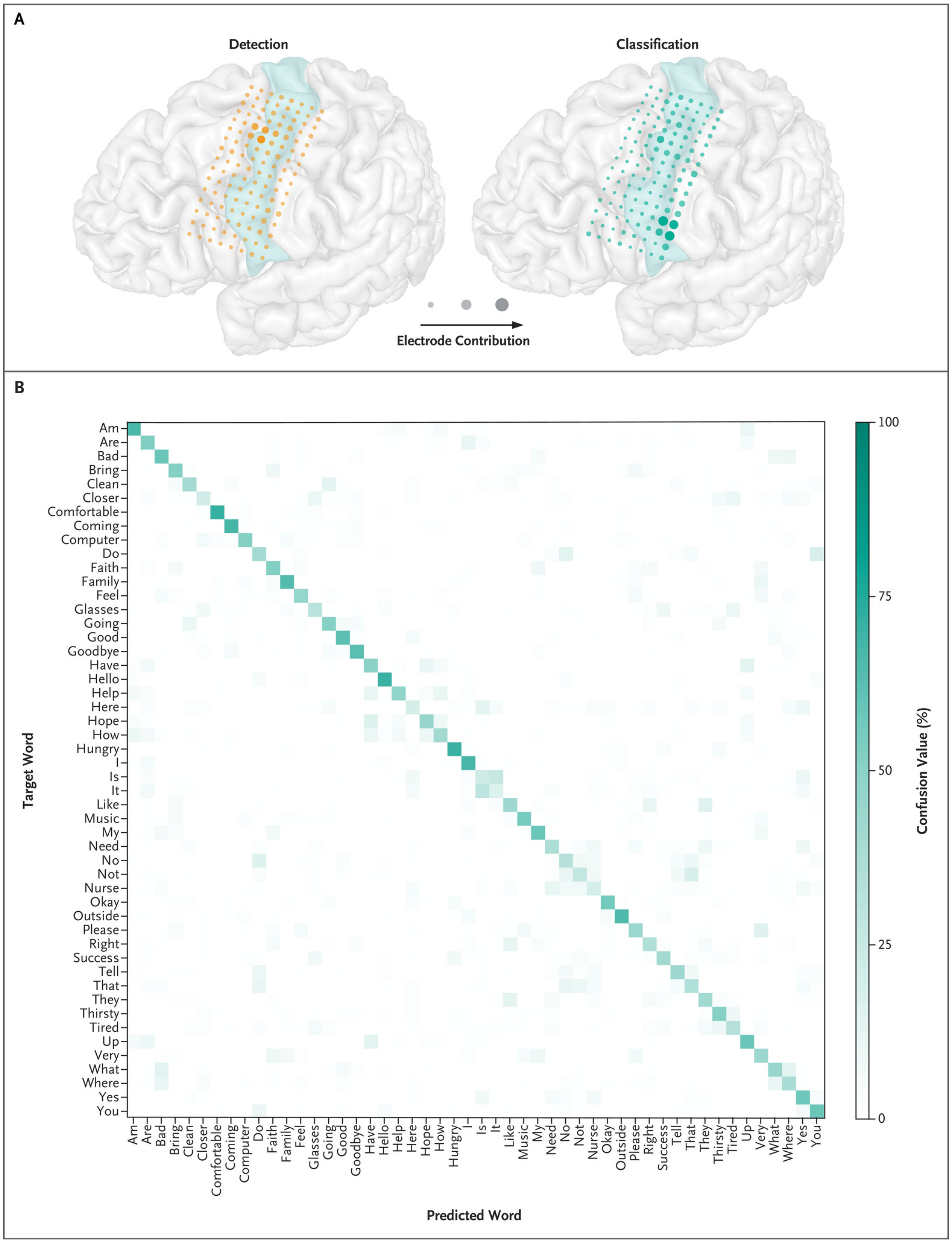


Figure 1.3. Distinct Neural Activity Patterns during Word-Production Attempts. (continued on next page).

(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.

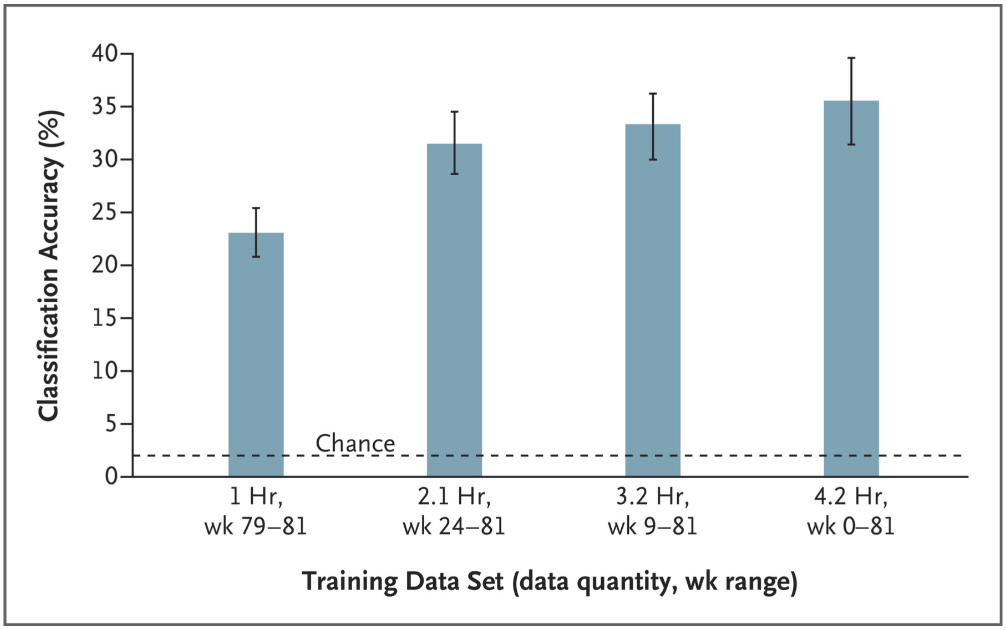


Figure 1.4. Signal Stability and Long-Term Accumulation of Training Data to Improve Decoder Performance. Each bar depicts the mean classification accuracy (the percentage of trials in which the target word was correctly predicted) from isolated-word data sampled from the final weeks of the study period (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.

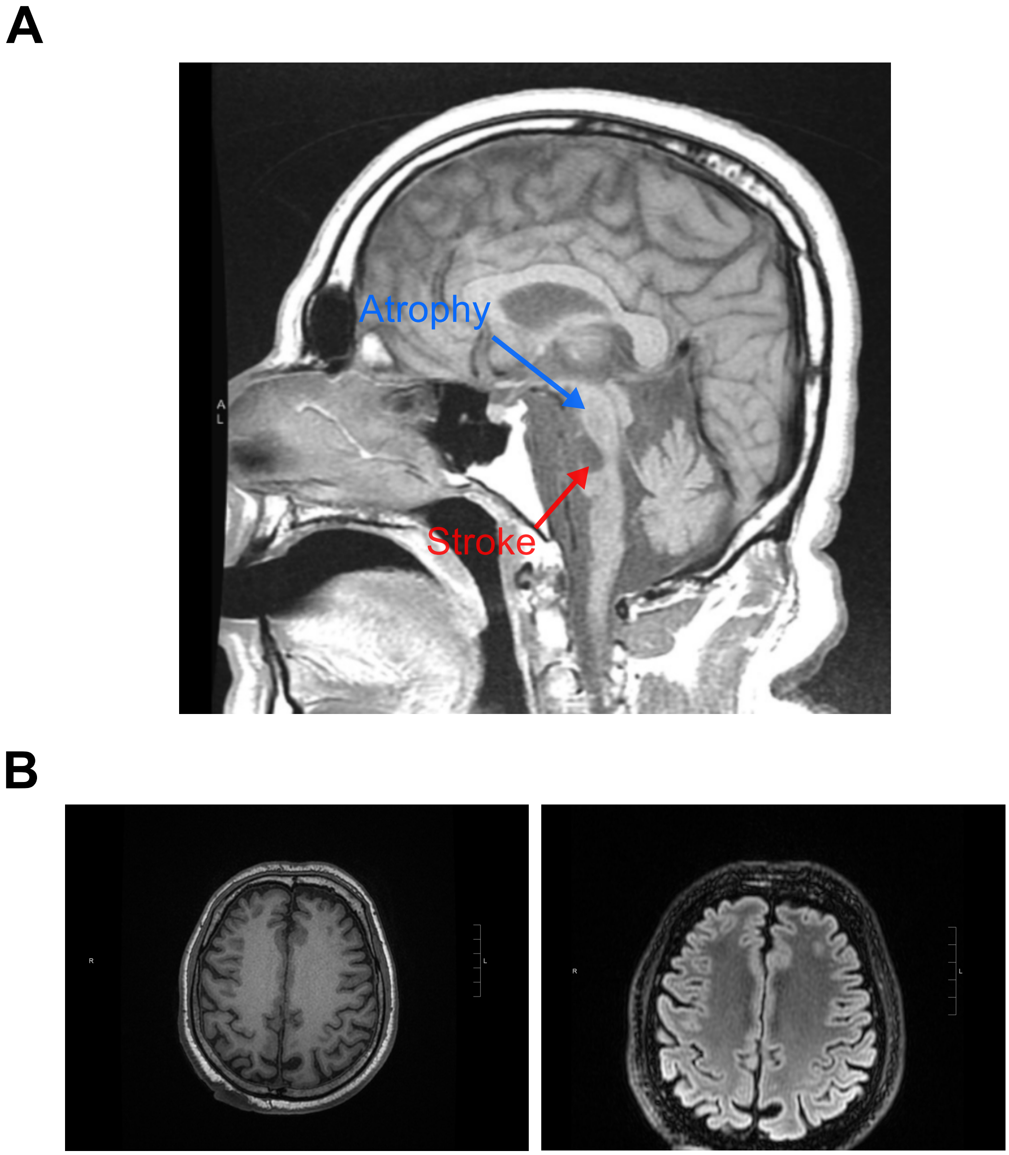


Figure 1.5. MRI results for the participant Panel A shows a sagittal MRI for the participant, who has encephalomalacia and brain-stem atrophy (labeled in blue) caused by pontine stroke (labeled in red). Panel B shows two additional MRI scans that indicate the absence of cerebral atrophy, suggesting that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

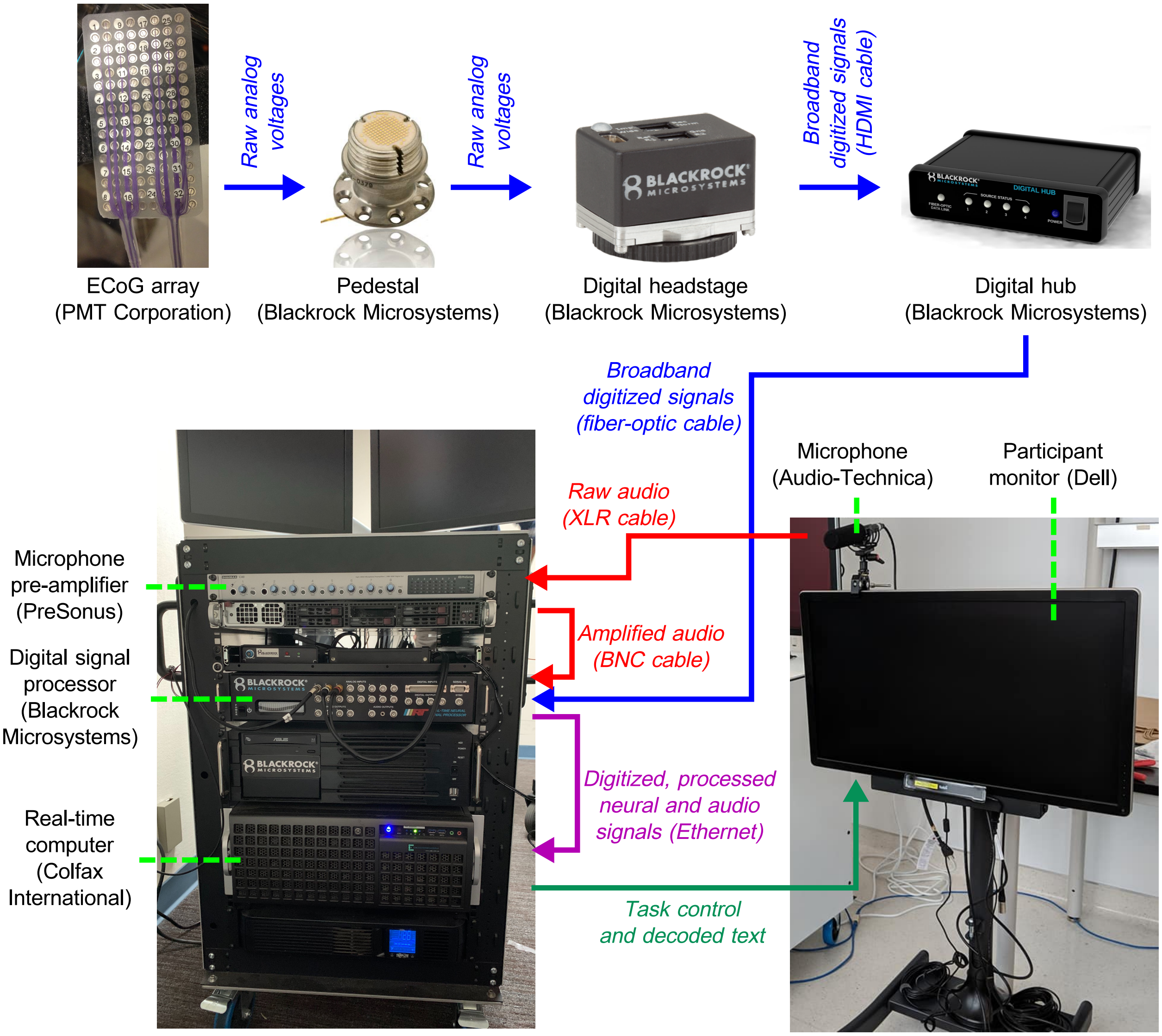


Figure 1.6. Real-time neural data acquisition hardware infrastructure Electrocorticography (ECoG) data acquired from the implanted array and percutaneous pedestal connector are processed and transmitted to the Neuroport digital signal processor (DSP). Simultaneously, microphone data are acquired, amplified, 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).

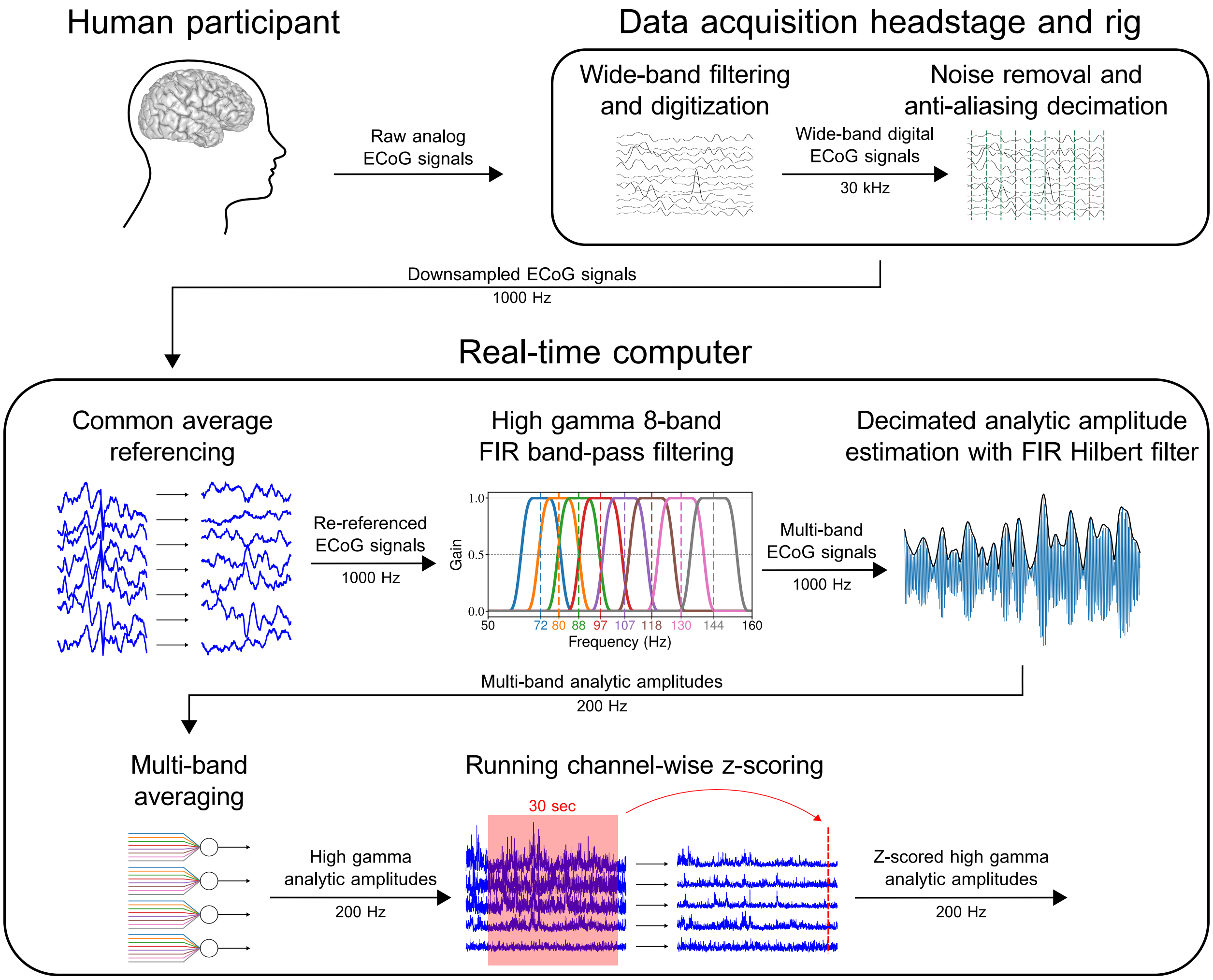

Human participant
Data acquisition headstage and rig
Wide-band filtering and digitization
Noise removal and anti-aliasing decimation
Raw analog ECoG signals
Wide-band digital ECoG signals
30 kHz
Downsampled ECoG signals
1000 Hz
Real-time computer
Common average referencing
High gamma 8-band FIR band-pass filtering
Decimated analytic amplitude estimation with FIR Hilbert filter
Re-referenced ECoG signals
1000 Hz
Gain
1.0
0.5
0.0
50
72
80
88
97
107
118
130
144
160
Frequency (Hz)
Multi-band ECoG signals
1000 Hz
Multi-band analytic amplitudes
200 Hz
Multi-band averaging
Running channel-wise z-scoring
30 sec
High gamma analytic amplitudes
200 Hz
Z-scored high gamma analytic amplitudes
200 Hz

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.

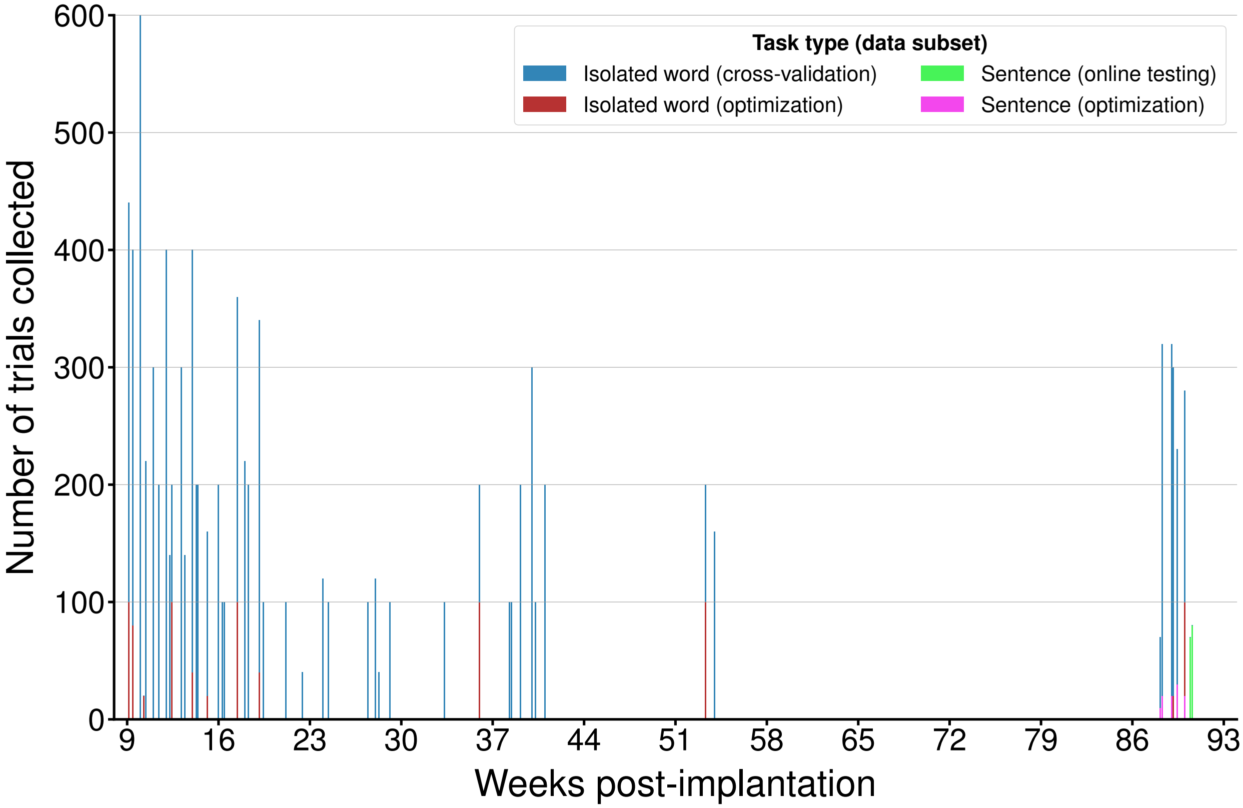


Figure 1.8. Data collection timeline Bars are stacked vertically if more than one data type was collected in a day (the height of the stacked bars for any given day is equal to the total number of trials collected that day). The irregularity of the data collection schedule was influenced by external 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.

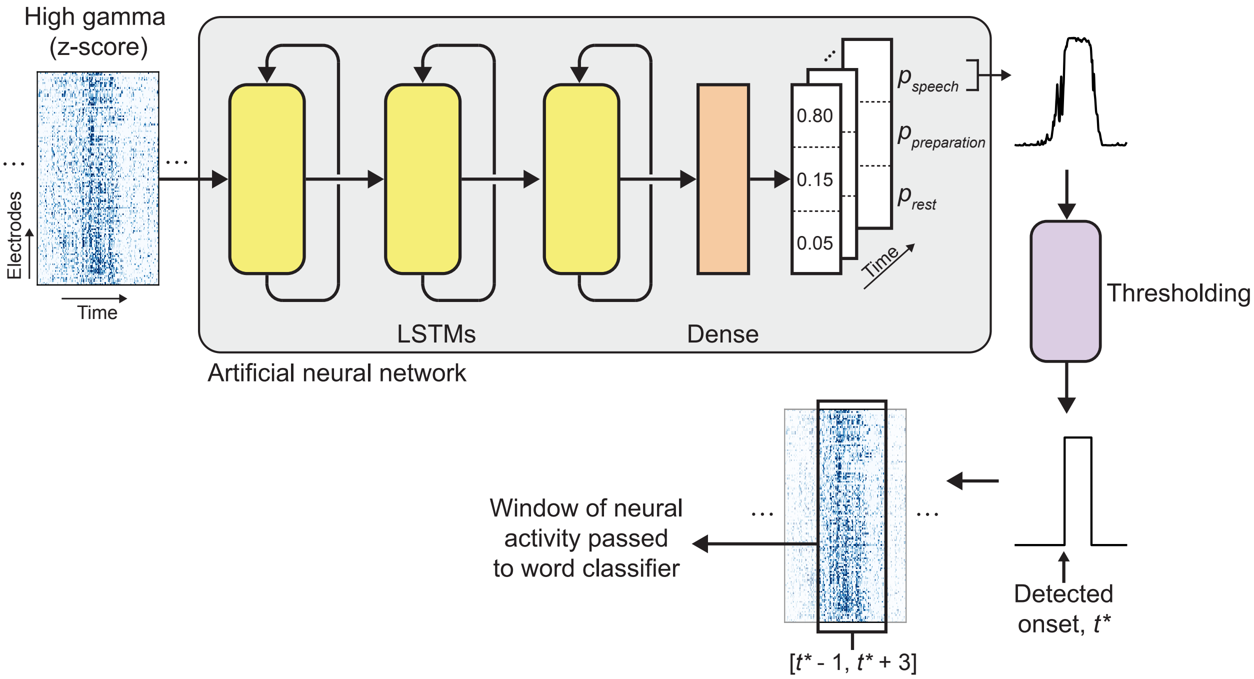


Figure 1.9. Speech detection model schematic The z-scored high gamma activity across all electrodes is processed time point by time point by an artificial neural network consisting of a stack of three long short-term memory layers (LSTMs) and a single dense (fully connected) layer. The dense layer projects 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”.

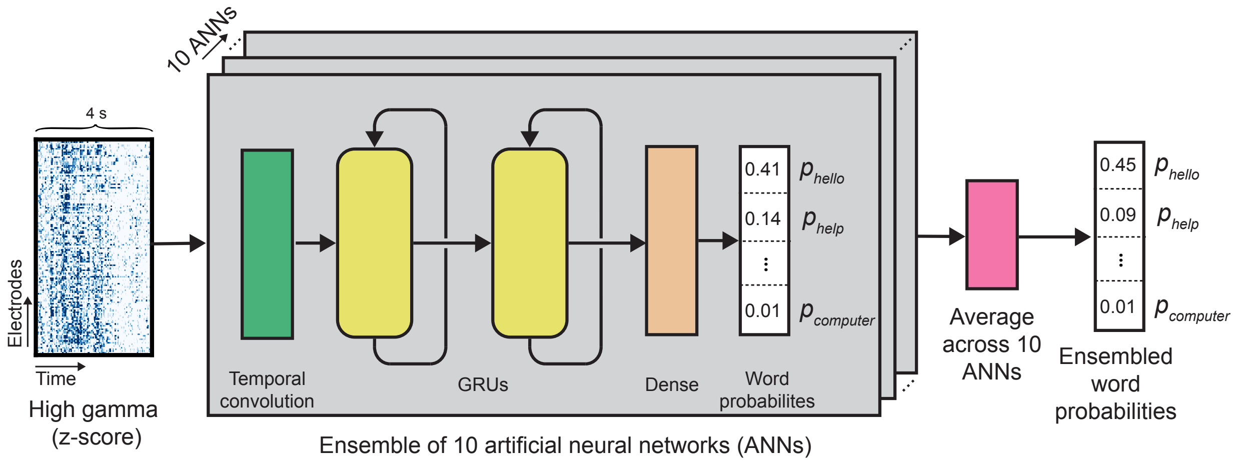


Figure 1.10. Word classification model schematic For each classification, a 4-second time window of high gamma activity is processed by an ensemble of 10 artificial neural network (ANN) models. Within each ANN, the high gamma activity is processed by a temporal convolution followed by two bidirectional 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.

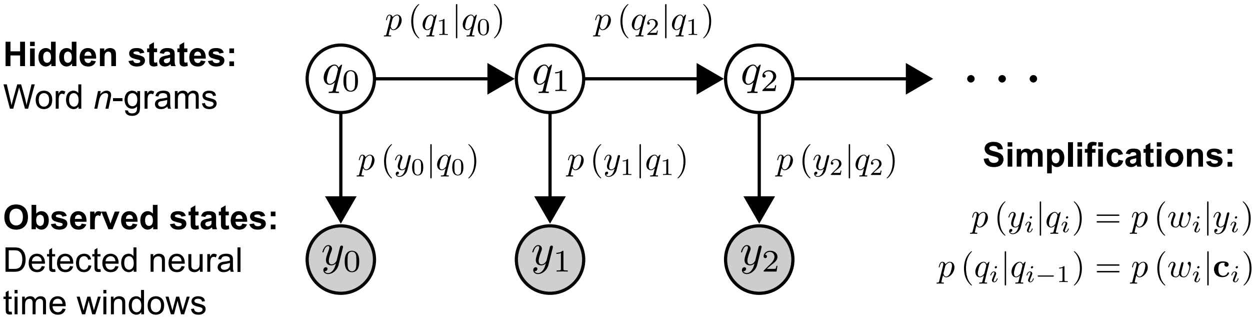


Figure 1.11. Sentence decoding hidden Markov model This hidden Markov model (HMM) describes the relationship between the words that the participant attempts to produce (the hidden states q_i) and the associated detected time windows of neural activity (the observed states y_i). The HMM emission probabilities $py_0|q_0$ 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).

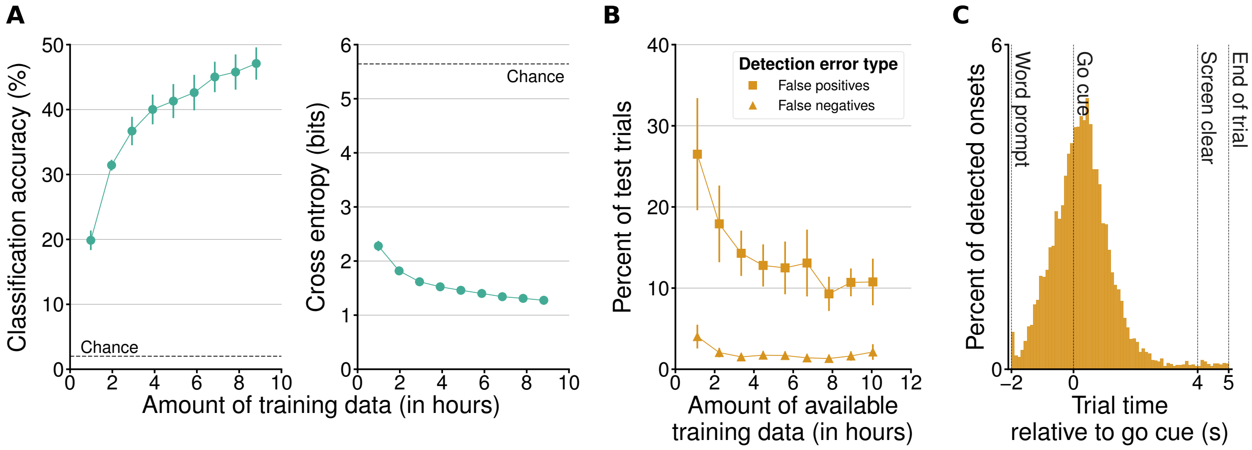


Figure 1.12. Auxiliary modeling results with isolated word data Panel A shows the effect of the amount of training data on word classification accuracy (left) and cross-entropy loss (right) using cortical activity recorded during the participant's isolated word production attempts. Lower cross entropy 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).

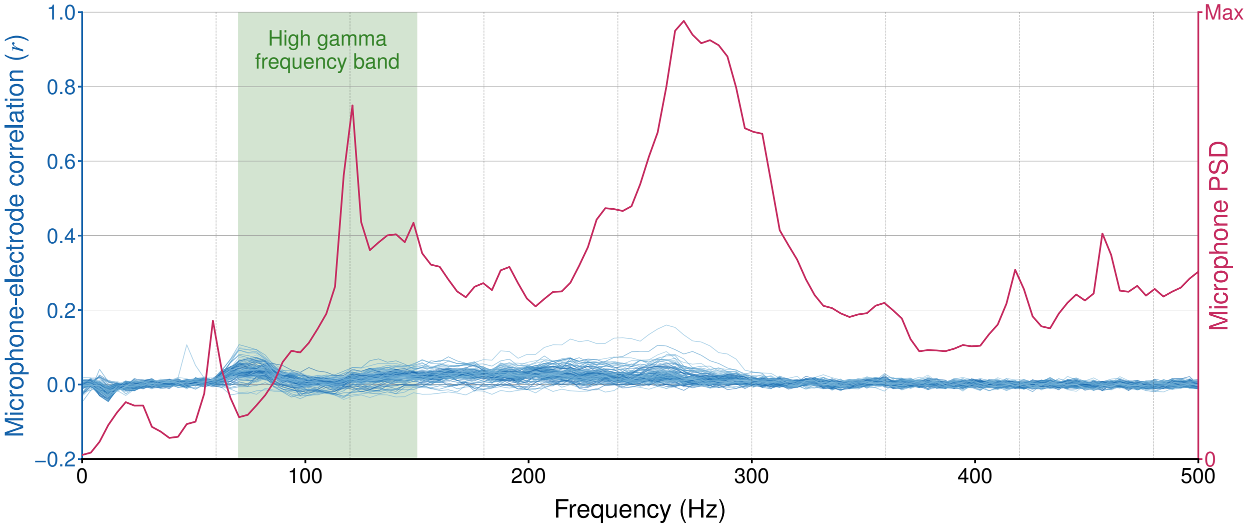


Figure 1.13. Acoustic contamination investigation Each blue curve depicts the average correlations between the spectrograms from a single electrode and the corresponding spectrograms from the time-aligned microphone signal as a function of frequency. The red curve depicts the average power spectral density (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.

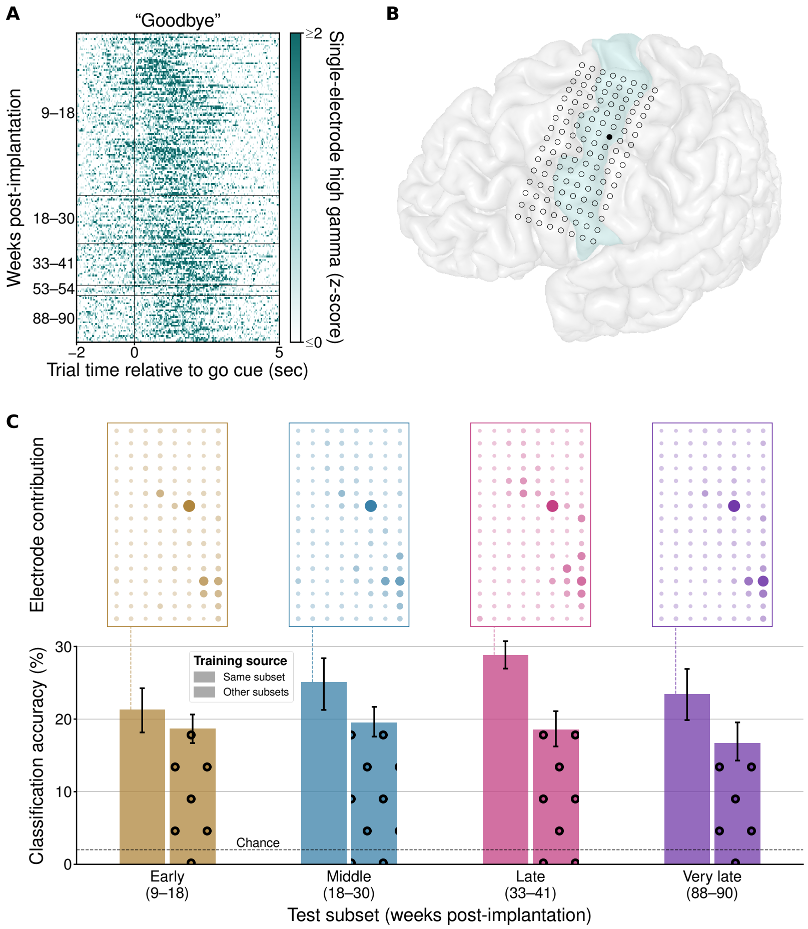


Figure 1.14. Long-term stability of speech-evoked signals Panel A shows neural activity from a single electrode across all of the participant's attempts to say the word "Goodbye" during the isolated word task, spanning 81 weeks of recording. Panel B shows the participant's brain reconstruction overlaid 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.

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 NATO and hand-motor classification model, developed the beam-search and langauge model scoring processes, and did all decoder, feature, vocabulary, and paradigm ablations. I also contributed to the neural-feature analyses and the nearest-neighbor analyses. With Jessie R. Lou, 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 decoding 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 and Schmitt 2003) and enable effective day-to-day use of assistive-communication technology (Tilborg and Deckers 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 displayed 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 = 8.08×10^{-12} and z = 7.09, P = 8.08×10^{-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 = 8.08×10^{-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 = 6.37×10^{-6} and z = 6.61, P = 1.19×10^{-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 = 6.34×10^{-8} and z = 6.11, P = 2.01×10^{-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 ($z = 3.78$, $P = 1.57 \times 10^{-4}$, 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 = 2.85×10^{-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 messages 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 robust 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 and Rodgers 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 articulatory 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 and Hickok 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 and McClellan 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 and Ba 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 previous 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 × 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 × 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 and Franz 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.

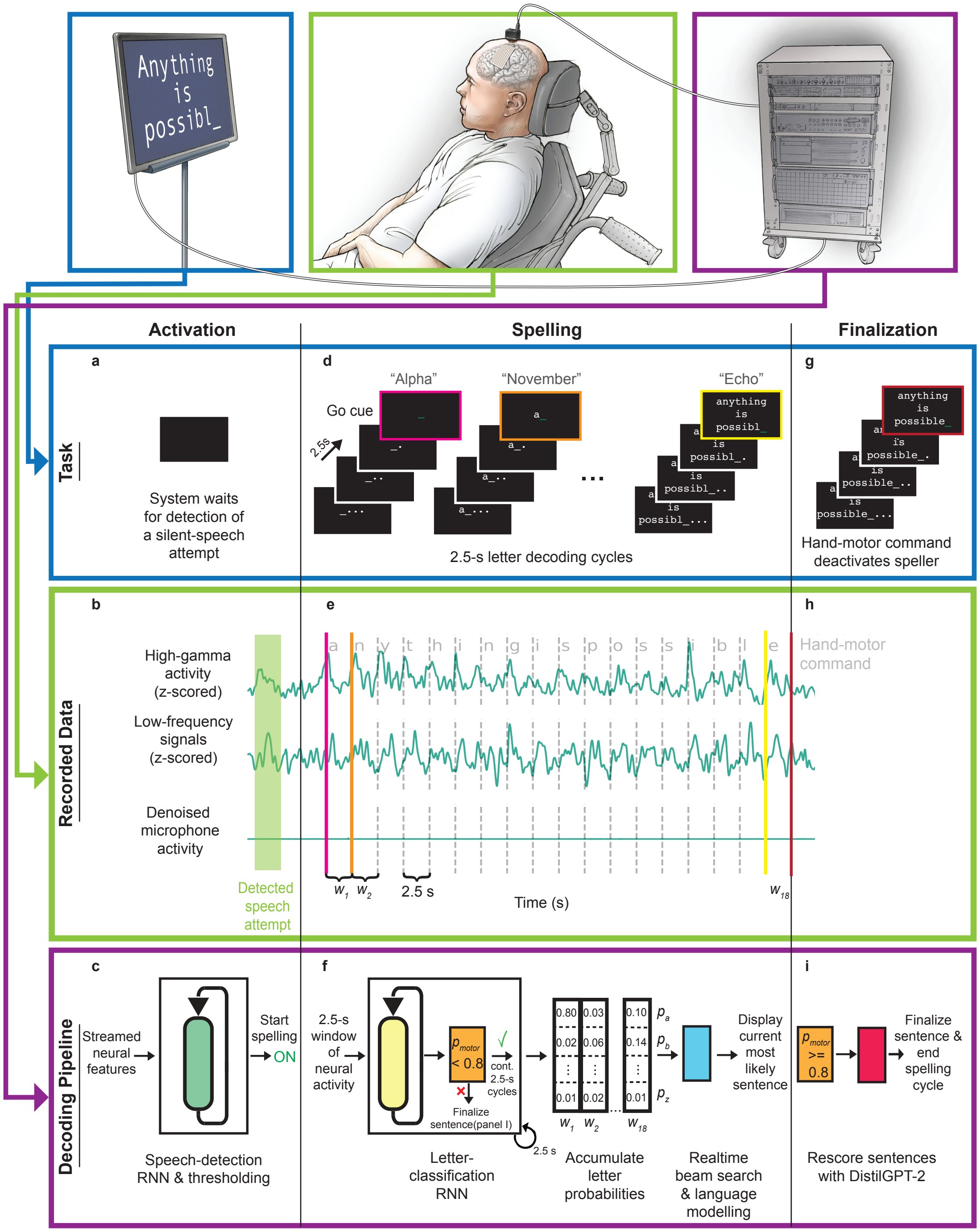
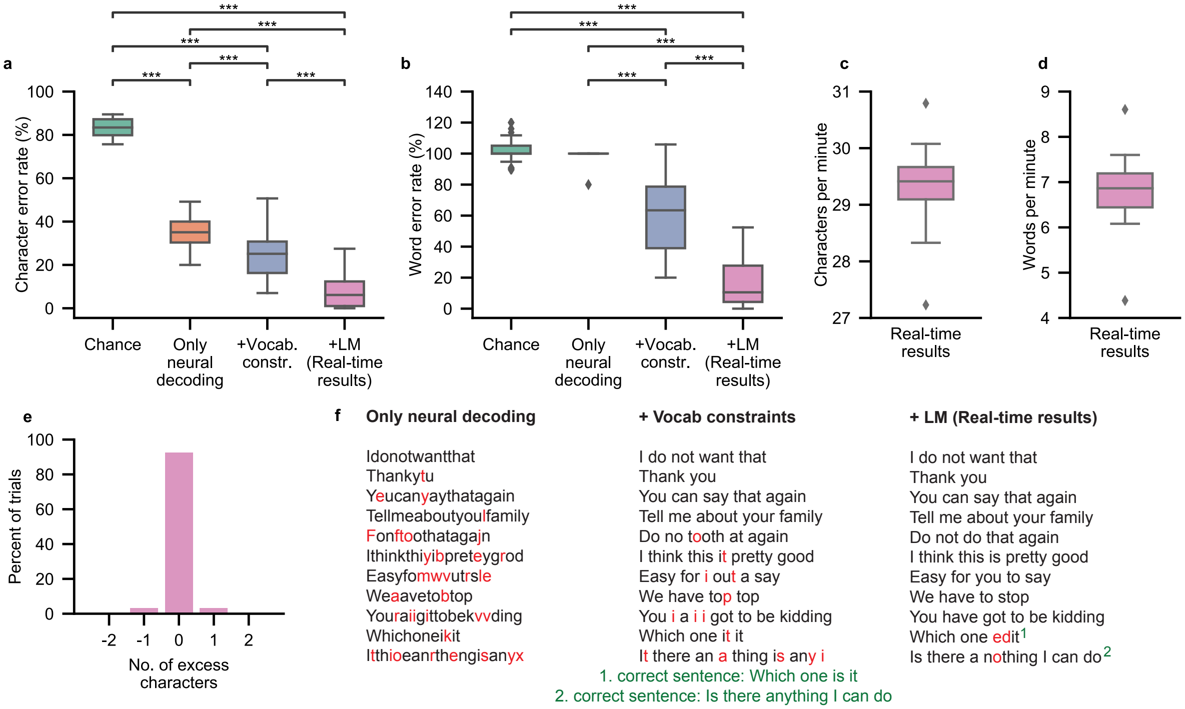


Figure 2.1. Schematic depiction of the spelling pipeline. (continued on next page).

(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.



Only neural decoding	+ Vocab constraints	+ LM (Real-time results)
Idonotwantthat	I do not want that	I do not want that
Thankytu	Thank you	Thank you
Yeucanyaythatagain	You can say that again	You can say that again
Tellmeaboutyoulfamily	Tell me about your family	Tell me about your family
Fonftoothatagajn	Do no tooth at again	Do not do that again
Ithinkthiyibpreteygrod	I think this it pretty good	I think this is pretty good
Easyfomwvutrsle	Easy for i out a say	Easy for you to say
Weaavetobtop	We have top top	We have to stop
Youraiigittobekvvding	You i a i i got to be kidding	You have got to be kidding
Whichoneikit	Which one it it	Which one edit[1]
Itthioeanrthengisanyx	It there an a thing is any i	Is there a nothing I can do[2]

1. correct sentence: Which one is it
2. correct sentence: Is there anything I can do

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.

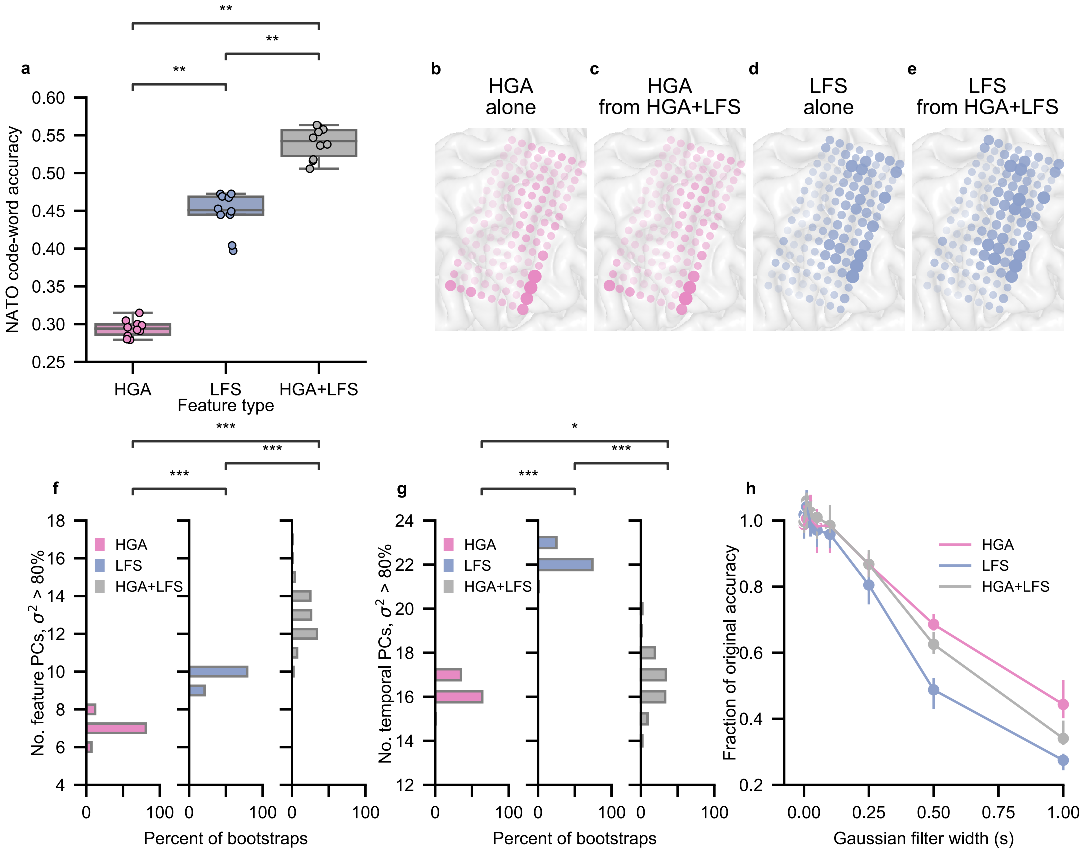


Figure 2.3. Characterization of high-gamma activity (HGA) and low-frequency signals (LFS) during silent-speech attempts. (continued on next page).

(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.

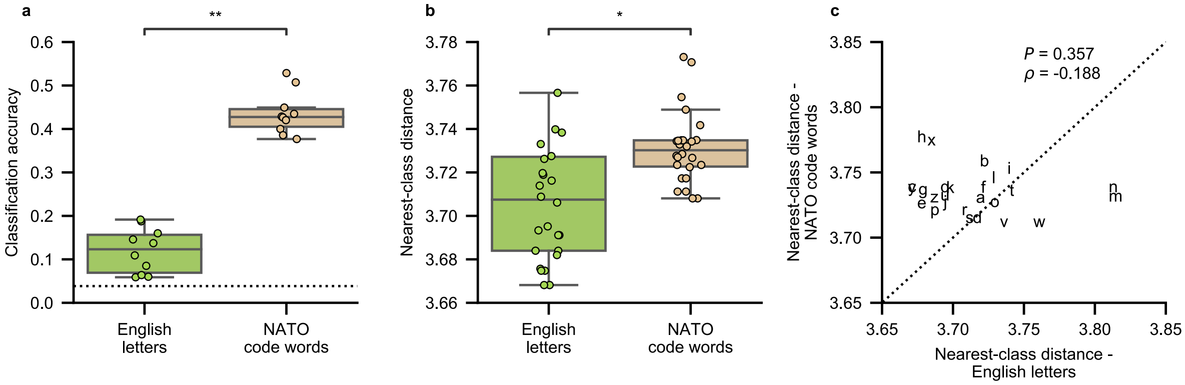


Figure 2.4. Comparison of neural signals during attempts to silently say English letters and NATO code words. **a** Classification accuracy (across n=10 cross-validation folds) using models trained with HGA+LFS features is significantly higher for NATO code words than for English letters (**P=1.57 × 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 × 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.

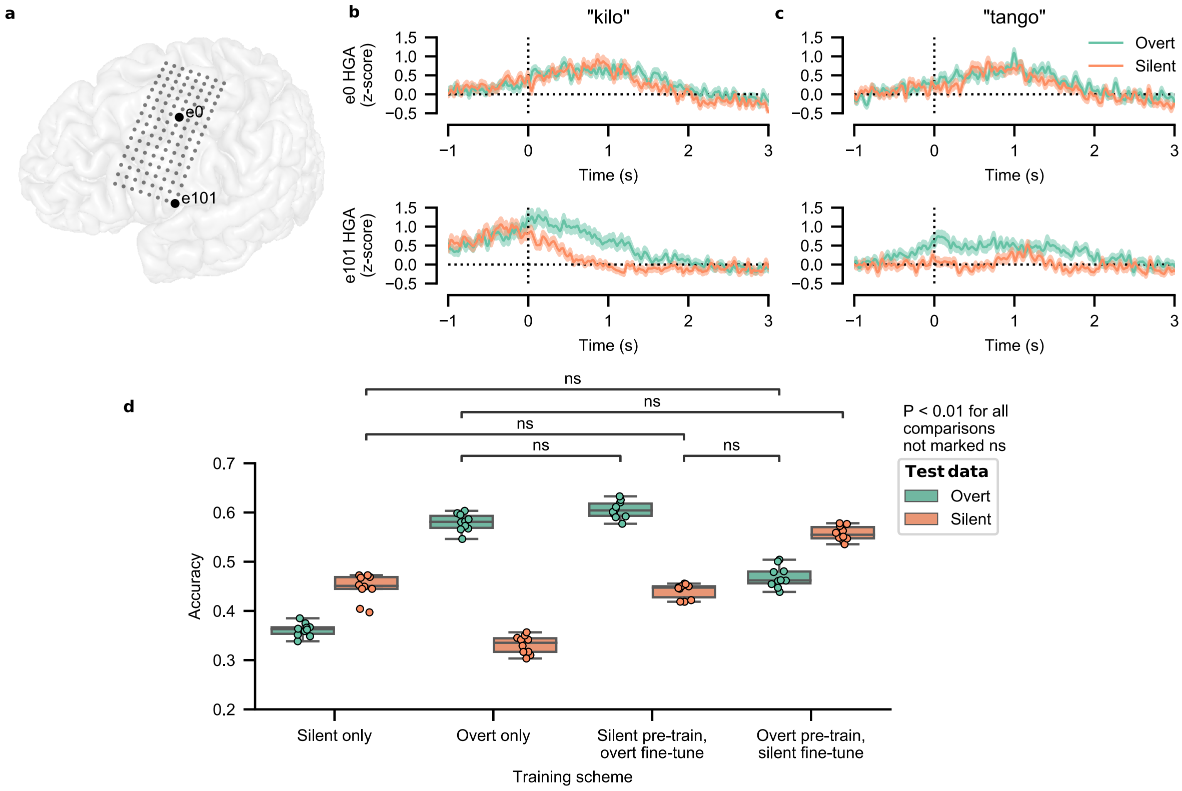
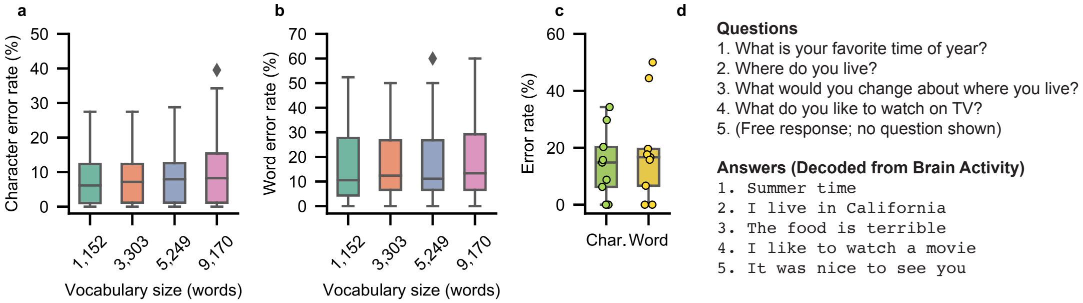


Figure 2.5. Differences in neural signals and classification performance between overt- and silent-speech attempts. a MRI reconstruction of the participant's brain overlaid with implanted electrode locations. The locations of the electrodes used in **b** and **c** are bolded and numbered in the overlay. **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** Code-word 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%.



Questions

1. What is your favorite time of year?
2. Where do you live?
3. What would you change about where you live?
4. What do you like to watch on TV?
5. (Free response; no question shown)

Answers (Decoded from Brain Activity)

```
1. Summer time
2. I live in California
3. The food is terrible
4. I like to watch a movie
5. It was nice to see you
```

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.

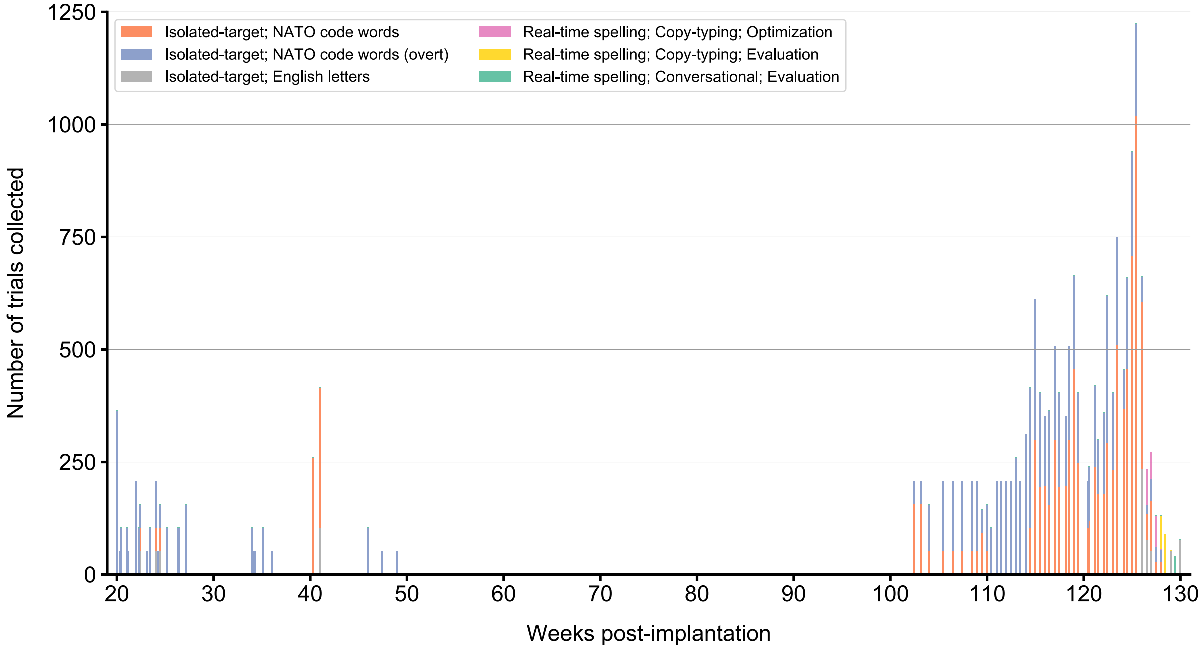


Figure 2.7. Data collection timeline Each bar depicts the total number of trials collected on each day of recording. The participant and implant date are the same as in our previous work (David A. Moses, Metzger, et al. 2021). If more than one type of dataset was collected in a single day, the bar is 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.

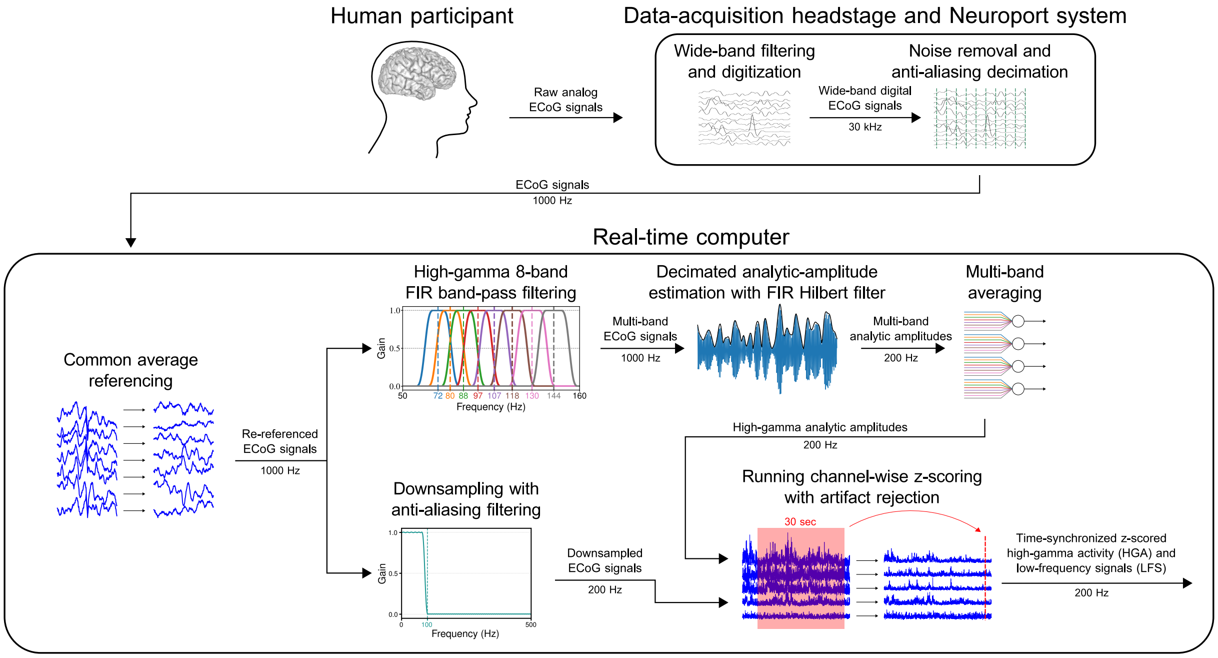

Human participant
Data-acquisition headstage and Neuroport system
Raw analog ECoG signals
Wide-band filtering and digitization
Wide-band digital ECoG signals
30 kHz
Noise removal and anti-aliasing decimation
ECoG signals
1000 Hz
Real-time computer
Common average referencing
Re-referenced ECoG signals
1000 Hz
High-gamma 8-band FIR band-pass filtering
Gain
1.0
0.5
0.0
50
72 80 88 97 107 118 130 144
160
Frequency (Hz)
Multi-band ECoG signals
1000 Hz
Decimated analytic-amplitude estimation with FIR Hilbert filter
Multi-band analytic amplitudes
200 Hz
Multi-band averaging
High-gamma analytic amplitudes
200 Hz
Downsampling with anti-aliasing filtering
Gain
1.0
0.5
0.0
0
100
500
Frequency (Hz)
Downsampled ECoG signals
200 Hz
Running channel-wise z-scoring with artifact rejection
30 sec
Time-synchronized z-scored high-gamma activity (HGA) and low-frequency signals (LFS)
200 Hz

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^{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^{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 and Jovanovic 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^{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.

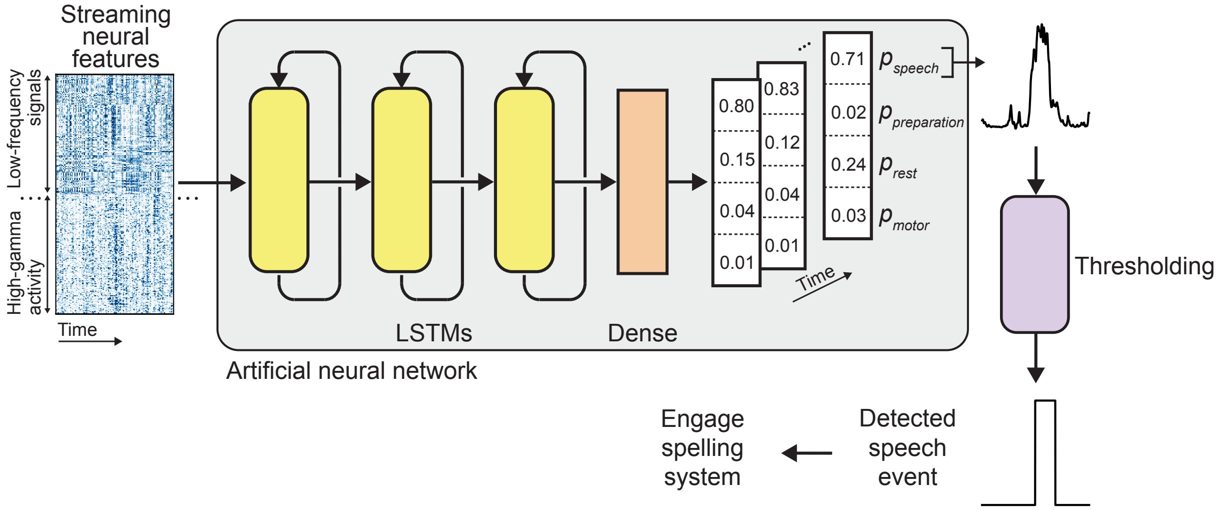


Figure 2.9. Speech-detection model schematic To detect silent-speech attempts from the participant's neural activity during real-time sentence spelling, first the z-scored low-frequency signals (LFS) and high-gamma activity (HGA) for each electrode are processed continuously by a stack of 3 long short-term 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.

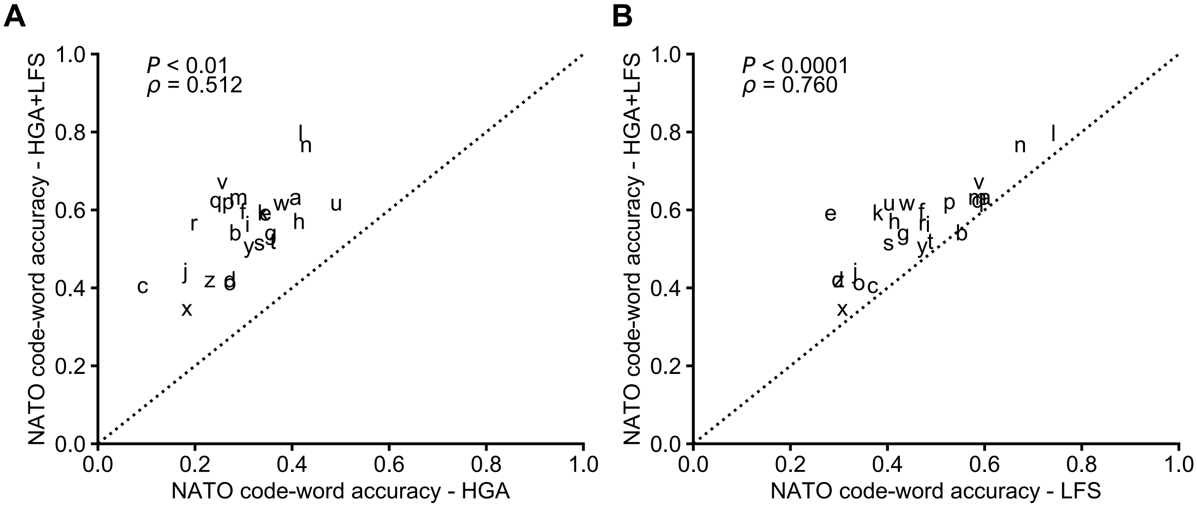


Figure 2.10. Effects of feature selection on code-word classification accuracy A. Classification accuracy improves for each code word when using high-gamma activity (HGA) and low-frequency signals (LFS) together (the combined HGA+LFS feature set) instead of only HGA features. The accuracies are significantly 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.

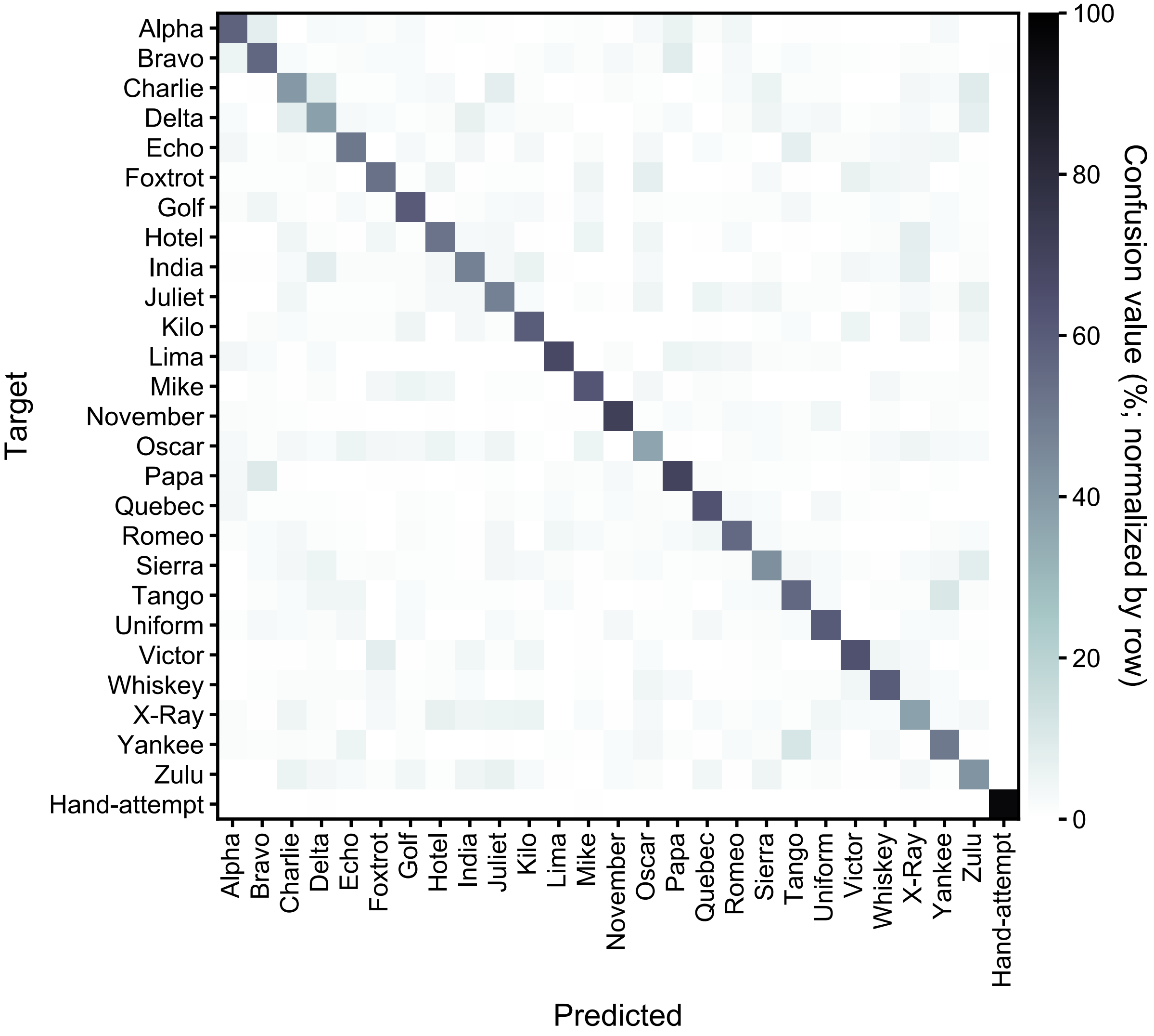


Figure 2.11. Confusion matrix from isolated-target trial classification using HGA and LFS Confusion values, computed during offline classification of neural data (using both high-gamma activity and low-frequency signals) recorded during isolated-target trials, are shown for each NATO code word and the 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 $[54.3, 58.2]$.

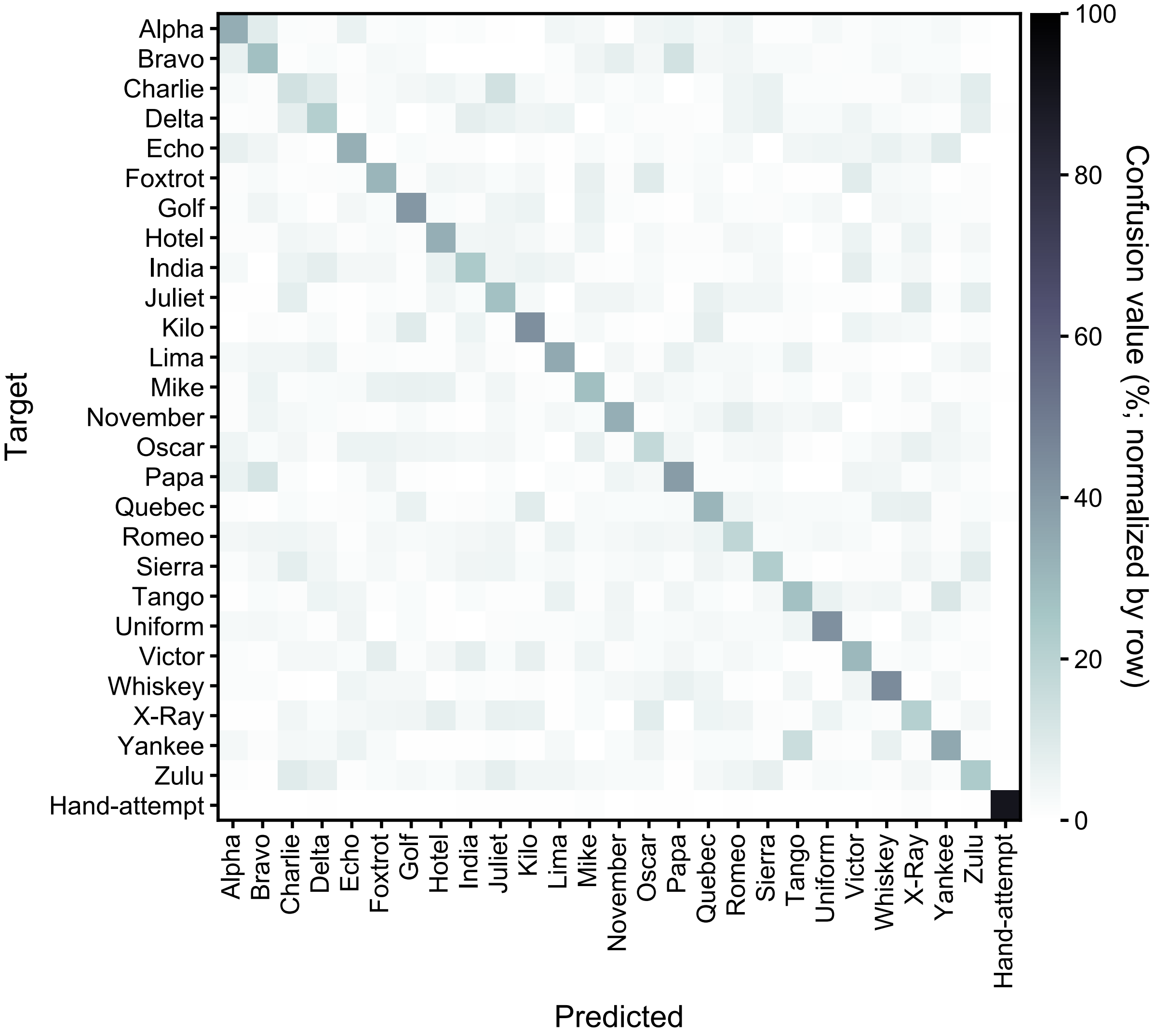


Figure 2.12. Confusion matrix from isolated-target trial classification using only HGA Confusion values, computed during offline classification of neural data (using only high-gamma activity) recorded during isolated-target trials, are shown for each NATO code word and the attempted hand squeeze. Each 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 $[32.0, 33.6]$.

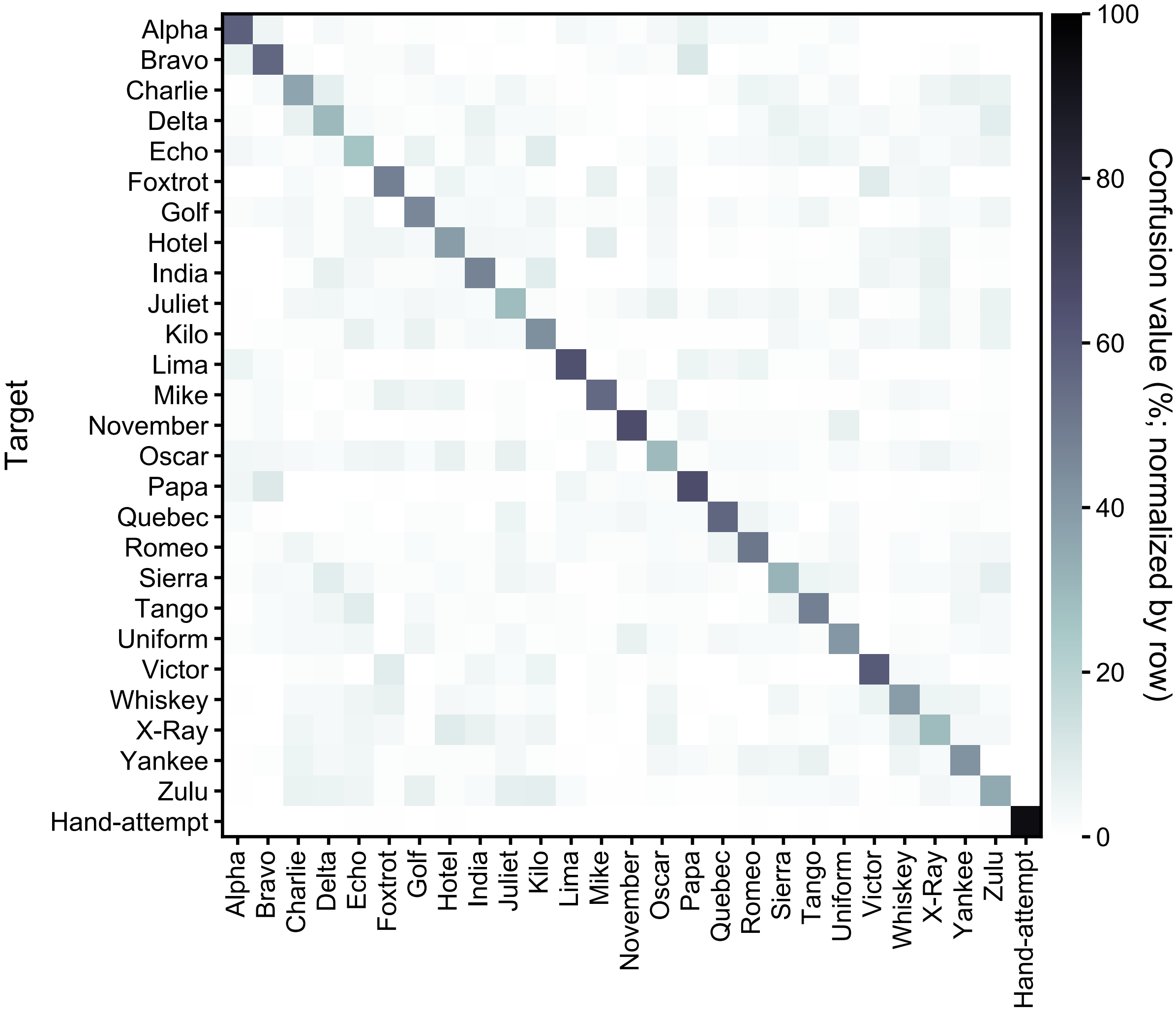


Figure 2.13. Confusion matrix from isolated-target trial classification using only LFS Confusion values, computed during offline classification of neural data (using only low-frequency signals) recorded during isolated-target trials, are shown for each NATO code word and the attempted hand squeeze. Each 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 $[42.9, 49.7]$.

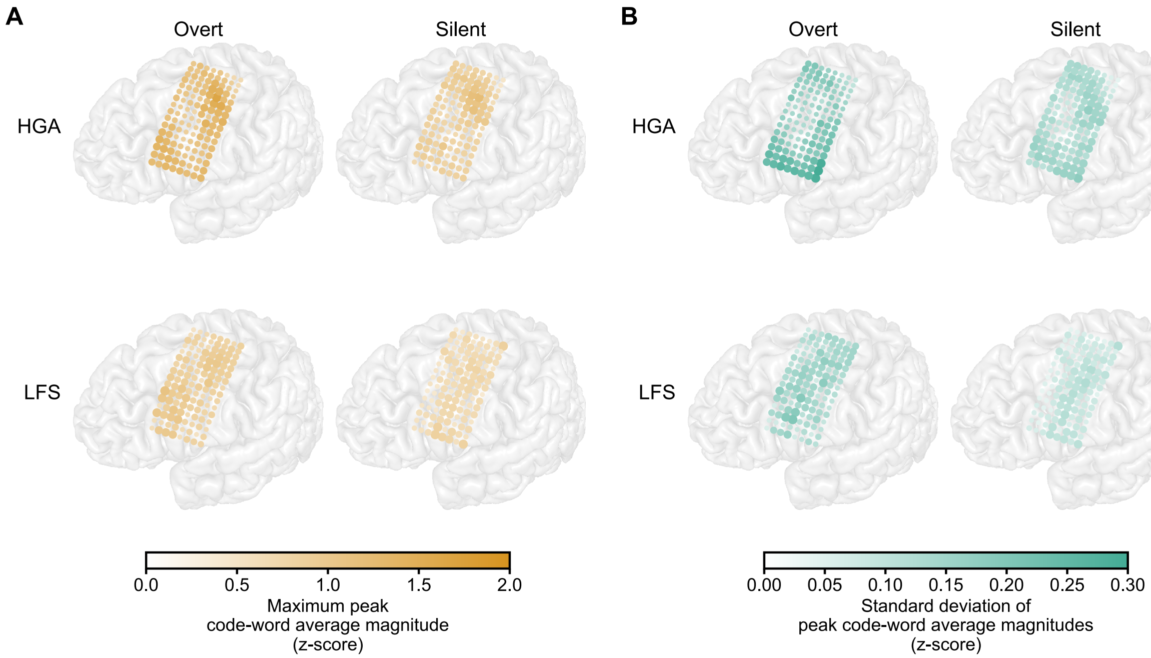


Figure 2.14. Neural-activation statistics during overt- and silent-speech attempts A. Each image shows an MRI reconstruction of the participant's brain overlaid with electrode locations and the maximum neural activations for each electrode, type of speech attempt (overt or silent), and feature type (high-gamma 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]	$\mid z$-value $\mid$	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]	$\mid z$-value $\mid$	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]	$\mid z$-value $\mid$	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		Group 2			P-value
Train	Test	Train	Test	$\mid z\text{-value} \mid$	(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

A high-performance neuroprosthesis for speech decoding and avatar control

Disclaimer: This chapter contains material from the following publication: **Sean L. Metzger***, Kaylo T. Littlejohn*, Alex B. Silva*, David A. Moses*, Margaret P. Seaton* et. al. (2023). A high performance neuroprosthesis for speech decoding and avatar control. *Nature*, 1-10. doi: 10.1038/s41586-023-06443-4

Personal contributions:

I designed and constructed the text decoding models and analyses with help from Alex Silva. I also developed the analyses showing persistent somatotopy in figure 5 with Alex Silva. With Kaylo Littlejohn, I trained and developed neural network models and analyses for avatar decoding (my models were the continuous articulatory gesture decoding, and discretized articulatory gesture decoding models). I also led the design and prioritization of the corpora and tasks. Kaylo Littlejohn, David Moses, and I developed the real-time decoding pipelines, and I also contributed to data collection. All co-first authors contributed to writing and editing the manuscript and the development of figures.

3.1 Abstract

Speech neuroprostheses have the potential to restore communication to people living with paralysis, but naturalistic speed and expressivity are elusive (David A. Moses et al. 2021). Here, we use high-density surface recordings of the speech cortex in a clinical-trial participant with severe limb and vocal paralysis to achieve high-performance real-time decoding across three complementary speech-related output modalities: text, speech audio, and facial-avatar animation. We trained and evaluated deep-learning models using neural data collected as the participant attempted to silently speak sentences. For text, we demonstrate accurate and rapid large-vocabulary decoding with a median rate of 78 words per minute and median word error rate of 25%. For speech audio, we demonstrate intelligible and rapid speech synthesis of high-utility phrases, in the participant's own voice, with human listeners achieving a median perceptual word error rate of 28%. For facial avatar, we demonstrate the control of virtual orofacial movements for speech and non-speech communicative gestures. The decoders reached high performance with fewer than two weeks of training. Our findings introduce a new multimodal speech-neuroprosthetic approach that has significant promise to restore full, embodied communication to people living with severe paralysis.

3.2 Main

Speech is the ability to express thoughts and ideas through spoken words. Speech loss after neurological injury is devastating because it significantly impairs communication and causes social isolation (Peters et al. 2015). Previous demonstrations have shown that it is possible to decode speech from the brain activity of a person with paralysis, but only in the form of text and with limited speed and vocabulary (David A. Moses et al. 2021; Metzger et al. 2022). A compelling goal is to both enable faster large-vocabulary text-based communication and restore the produced speech sounds and facial movements related to speaking. While text outputs are good for basic messages, speaking has rich prosody, expressiveness, and identity that can enhance embodied communication beyond what can be conveyed in text alone. To address this, we designed a multimodal speech neuroprosthesis that uses broad-coverage, high-density electrocorticography to decode text and audio-visual speech outputs from articulatory vocal-tract representations distributed throughout the sensorimotor cortex. Due to severe paralysis caused by a basilar artery brainstem stroke that occurred over 18 years ago, our 47-year-old participant cannot speak or vocalize speech sounds given the severe weakness of her orofacial and vocal muscles (anarthria) and cannot type given the weakness in her arms and hands (quadriplegia). Instead, she uses commercial head tracking assistive technology to communicate slowly to select letters at up to 14 words per minute. Here, we demonstrate flexible, real-time decoding of brain activity into text, speech sounds, and both verbal and non-verbal orofacial movements. Additionally, we show that decoder performance is driven by broad coverage of articulatory representations distributed throughout the sensorimotor cortex that have persisted after years of paralysis.

3.3 Overview of multimodal speech-decoding system

We designed a speech-decoding system that enabled a clinical-trial participant (ClinicalTrials.gov; NCT03698149) with severe paralysis and anarthria to communicate by decoding intended sentences from signals acquired by a 253-channel high-density electrocorticography (ECoG) array implanted over speech cortical areas of the sensorimotor cortex and superior temporal gyrus (Figure 3.1a-c). The array was positioned over cortical areas relevant for orofacial movements, and simple movement tasks demonstrated differentiable activations associated with attempted movements of the lips, tongue, and jaw (Figure 3.1d).

For speech decoding, the participant was presented with a sentence as a text prompt on a screen and was instructed to silently attempt to say the sentence after a visual go cue. Specifically, she attempted to silently speak the sentence without vocalizing any sounds, which differs from imagined or inner speech because she was trying to engage her articulators to the best of her ability. It also differs from mouthing words as she has significant orofacial weakness and we did not want her to struggle or be constrained by these limitations. Meanwhile, we processed neural signals recorded from all 253 ECoG electrodes to extract high-gamma activity (HGA; between 70–150 Hz) and low-frequency signals (LFS; between 0.3–17 Hz) (Metzger et al. 2022). We trained deep-learning models to learn mappings between these ECoG features and phones, speech-sound features, and articulatory gestures, which we then used to output text, synthesize speech audio, and animate a virtual avatar, respectively (Figure 3.1a).

We evaluated our system using three custom sentence sets containing varying amounts of unique words and sentences named “50-phrase-AAC,” “529-phrase-AAC,” and “1024-word-General.” The first two sets closely mirror corpora preloaded on commercially available augmentative and alternative communication (AAC) devices, designed to let patients express basic concepts and caregiving needs (Beukelman et al. 2007). We chose these two sets to assess our ability to decode high-utility sentences at a limited and expanded vocabulary level. The 529-phrase-AAC set contained 529 sentences composed of 372 unique words, and we selected 50 high-utility sentences composed of 119 unique words to create the 50-phrase-AAC set. To evaluate how well our system performed with a larger vocabulary containing common English words, we created the 1024-word-General set, containing 9,655 sentences composed of 1,024 unique words sampled from Twitter and movie transcriptions. We primarily used this set to assess how well our decoders could generalize to sentences that the participant did not attempt to say during training with a vocabulary size large enough to facilitate general-purpose communication.

To train our neural-decoding models prior to real-time testing, we recorded ECoG data as the participant silently attempted to speak individual sentences. Learning statistical mappings between the ECoG features and the sequences of phones and speech sound features in the sentences was challenged by the absence of clear timing information in the attempted speech. To overcome the inability to definitively know when the phones and speech units began and ended, we used a connectionist temporal classification (CTC) loss function during training of our neural decoders, which is commonly used in automatic speech recognition to infer sequences of sub-word units (such as phones or letters) from speech waveforms when precise time alignment between the units and the waveforms is unknown (Graves et al. 2006). We used CTC loss during training of the text-decoding, speech-synthesis, and articulatory-decoding models to enable prediction of phone probabilities, discrete speech-sound units, and discrete articulator movements, respectively, from the ECoG signals.

3.4 Text decoding

Text-based communication is an important modality for facilitating messaging and interaction with technology. Initial efforts to decode text from the brain activity of a person with anarthria during attempted speech had various limitations, including slow decoding rates and small vocabulary sizes (David A. Moses et al. 2021; Metzger et al. 2022). Here, we address these limitations by implementing a flexible approach via phone decoding, enabling decoding of arbitrary phrases from large vocabularies while approaching naturalistic speaking rates.

To evaluate real-time performance, we decoded text as the participant attempted to silently say 249 randomly selected sentences from the 1024-word-General set that were not used during model training (Figure 3.2a). To decode text, we streamed features extracted from ECoG signals starting 500 ms prior to the go cue into a bidirectional recurrent neural network (RNN). Prior to testing, we trained the RNN to predict the probabilities of 39 phones and silence at each time step. A CTC beam search then determined the most likely sentence given these probabilities. First, it created a set of candidate phone sequences that were constrained to form valid words within the 1,024-word vocabulary. Then, it evaluated candidate sentences by combining each candidate's underlying phone probabilities with its linguistic probability using a natural-language model.

To quantify text-decoding performance, we used standard metrics in automatic speech recognition: word error rate (WER), phone error rate (PER), character error rate (CER), and words per minute (WPM). WER, PER, and CER measure the percentage of decoded words, phones, and characters, respectively, that were incorrect.

During real-time evaluation sentences then computed error rates across sequential pseudo-blocks of 10-sentence segments (and one pseudo-blocks of 9 sentences due to an error trail). We achieved a median PER of 18.5% (99% CI [14.1, 28.5]; Figure 3.2b), a median WER of 25.5% (99% CI [19.3, 34.5]; Figure 3.2c), and a median CER of 19.9% (99% CI [15.0, 30.1]; Figure 3.2d; see Table 1 for example decodes; see Figure 3.6 for the relationship between decoded PER and WER). For all metrics, performance was better than chance, which we computed by re-evaluating performance after using temporally shuffled neural data as the input to our decoding pipeline ($P < 0.0001$ for all three comparisons, two-sided Wilcoxon rank-sum tests with 5-way Holm-Bonferroni correction). The average WER passes the 30% threshold below which speech-recognition applications generally become useful 6 while providing access to a large vocabulary of over 1,000 words, indicating that our approach may be viable in clinical applications.

To probe whether decoding performance was dependent on the size of the vocabulary used to constrain model outputs and train the language model, we measured decoding performance in offline simulations using log-spaced vocabulary sizes ranging from 1,506 to 39,378 words. We created each vocabulary using the by augmenting the 1024-word-General vocabulary with the N-1,024 most frequently occurring words outside this set in large-scale corpora, where N is the size of the vocabulary. Then, for each vocabulary, we retrained the natural-language model to incorporate the new words and enabled the model to output any word from the larger vocabulary, and then performed decoding with the real-time evaluation trials. We observed robust decoding performance as vocabulary size grew (Figure 3.2g, See Figure 3.7 for CER and PER). With a vocabulary of 39,378 words, we achieved a median offline WER of 27.6% (99% CI [20.0, 34.7]).

We verified that our system remained functional in a freeform setting in which the participant volitionally and spontaneously attempted to silently say unprompted sentences, with the neural data aligned to speech onsets detected directly from the neural features instead of to go cues.

We observed a median real-time decoding rate of 78.3 WPM (99% CI [75.5, 79.4]; Figure 3.2f). This decoding rate exceeds our participant's typical communication rate using her previous assistive device (14.2 WPM) and is closer to naturalistic speaking rates than has been previously reported with communication neuroprostheses (David A. Moses et al. 2021; Metzger et al. 2022; Vansteensel et al. 2016; Pandarinath et al. 2017; Willett et al. 2021a).

To assess how well our system could decode phones in the absence of a language model and constrained vocabulary, we evaluated performance using just the RNN neural-decoding model (using the most likely phone prediction at each time step) in an offline analysis. This yielded a median PER of 29.4% (99% CI [26.2, 32.8.]; Figure 3.2b) which is only 10.9 percentage points higher than the full model, demonstrating that the primary contributor to phone-decoding performance was the neural-decoding RNN model and not the CTC beam search or language model ($P < 0.0001$ for all comparisons to chance and to the full model, two-sided Wilcoxon signed-rank tests with 5-way Holm-Bonferroni correction; Extended Data Table 1).

We also characterized the relationship between quantity of training data and text-decoding performance in offline analyses. For each day of data collection, we trained 5 models with different random initializations on all the data collected on or before that date, then simulated performance on the real-time blocks. We observed steadily declining error rates over the course of 13 days of training-data collection (Figure 3.2f), during which we collected 9,506 sentence trials corresponding to about 1.6 hours of training data per day. These results show that functional speech-decoding performance can be achieved after a relatively short period of data collection compared to our prior work 1,3 and is likely to continue to improve with more data.

To assess signal stability, we measured real-time classification performance during a separate NATO-motor task that we collected during each research session with our participant. In each trial of this task, we prompted the participant to attempt to either silently say one of the 26 code words from the NATO phonetic alphabet (alpha, bravo, charlie, and so forth) or attempt one of four hand movements (described and analyzed in a later section). We trained a neural-network classifier to predict the most likely NATO code word from a 4-second window of ECoG features (aligned to the task go cue), and we evaluated real-time performance with the classifier during the NATO-motor task (Figure 3.2g). We continued to retrain the model using available data prior to real-time testing until day 40, at which point we froze the classifier after training it on data from the 1,196 available trials . Across 19 sessions after freezing the classifier, we observed a mean classification accuracy of 96.8% (99% CI [94.5, 98.6]), with accuracies of 100% obtained on 8 of these sessions. Accuracy remained high after a 61 day pause in recording for the participant to travel. These results illustrate the stability of the cortical-surface neural interface without requiring recalibration and demonstrate that high performance can be achieved with relatively few training trials.

To evaluate model performance on pre-defined sentence sets without any pausing between words, we trained text-decoding models on neural data recorded as the participant attempted to silently say sentences from the 50-phrase-AAC and 529-phrase-AAC sets, then simulated offline text decoding with these sets. With the 529-phrase-AAC set, we observed a median WER of 17.1% across sentences (99% CI [8.89%, 28.9%]), with a median decoding rate of 89.9 WPM (99% CI [83.6, 93.3]). With the 50-phrase-AAC set, we observed a median WER of 4.92% (99% CI [3.18, 14.04]) with median decoding speeds of 101 WPM (99% CI: [95.6, 103]). PERs and CERs for each set are given in Figure 3.8 and Figure 3.9. These results illustrate extremely rapid and accurate decoding for finite, pre-defined sentences that could be used frequently by users.

3.5 Personalized speech synthesis

An alternative approach to text decoding is to synthesize speech sounds directly from recorded neural activity, which could offer a pathway towards more naturalistic and expressive communication for someone who is unable to speak. Previous work in speakers with intact speech has demonstrated that intelligible speech can be synthesized from neural activity during vocalized or mimed speech (Angrick et al. 2019; Anumanchipalli et al. 2019) but this has not been shown with someone who is paralyzed.

We performed real-time speech synthesis by transforming the participant's neural activity directly into audible speech as she attempted to silently speak during the audio-visual task condition (Figure 3.3a). To synthesize speech, we passed time windows of neural activity around the go cue into a bidirectional RNN. Prior to testing, we trained the RNN to predict the probabilities of 100 discrete speech units at each time step. To create the reference speech-unit sequences for training, we used HuBERT, a self-supervised speech-representation learning model (Hsu et al. 2021a) which encodes a continuous speech waveform into a temporal sequence of discrete speech units that captures latent phonetic and articulatory representations (Cho et al. 2022). Because our participant cannot speak, we acquired reference speech waveforms from a recruited speaker for the AAC sentence sets or via a text-to-speech algorithm for the 1024-word-General set. We used a CTC loss function during training to enable the RNN to learn mappings between the ECoG features and speech units derived from these reference waveforms without alignment between our participant's silent speech attempts and the reference waveforms. After predicting the unit probabilities, we passed the most likely unit at each time step into a pre-trained unit-to-speech model that first generated a mel spectrogram and vocoded this mel spectrogram into an audible speech waveform in real time (Lakhotia et al. 2021; Prenger et al. 2019). Offline, we used a voice-conversion model trained on a brief segment of the participant's speech (recorded before her injury) to process the decoded speech into the participant's own personalized synthetic voice.

We qualitatively observed that spectrograms decoded in real-time shared both fine-grained and broad time-scale information with corresponding reference spectrograms (Figure 3.3b). To quantitatively assess the quality of the decoded speech, we used the mel-cepstral distortion (MCD) metric, which measures the similarity between two sets of mel-cepstral coefficients (which are speech-relevant acoustic features) and is commonly used to evaluate speech-synthesis performance (Yamagishi et al. 2010). Lower MCD indicates stronger similarity. We achieved mean MCDs of 3.45 (99% CI [3.25, 3.82]), 4.49 (99% CI [4.07, 4.67]), and 5.21 (99% CI [4.74, 5.51]) dB for the 50-phrase-AAC, 529-phrase-AAC, and 1024-word-General sets, respectively (Figure 3.3c). We observed similar MCD performance on the participant's personalized voice. Performance increased as the number of unique words and sentences in the sentence set decreased but was always better than chance (all $P < 0.0001$, two-sided Wilcoxon rank-sum tests with 19-way Holm-Bonferroni correction; chance MCDs were measured using waveforms generated by passing temporally shuffled ECoG features through the synthesis pipeline). Furthermore, these MCDs are comparable to those observed with text-to-speech synthesizers 16 and better than prior neural-decoding work with participants that were able to speak naturally 11.

Human-transcription assessments are a standard method to quantify the perceptual accuracy of synthesized speech 17. To directly assess the intelligibility of our synthesized speech waveforms, crowd-sourced evaluators listened to the synthesized speech waveforms and then transcribed what they heard into text. We then computed perceptual WER and CERs by comparing these transcriptions to the ground-truth sentence texts. We achieved median WERs of 8.20% (99% CI [3.28, 14.5]), 28.2% (99% CI [18.6, 38.5]), 54.4% (99% CI [50.5, 65.2]) and median CERs of 6.64% (99% CI [2.71, 10.6]), 26.3% (99% CI [15.9, 29.7]), and 45.7% (99% CI [39.2, 51.6]) across test trials for the 50-phrase-AAC, 529-phrase-AAC, and 1024-word-General sets, respectively (Figure 3.3d and 3e; see Figure 3.11 for correlations between WER and MCD). Similar to the MCD results,WERs and CERs improved as the number of unique words and sentences in the sentence set decreased (all $P < 0.0001$, two-sided Wilcoxon rank-sum tests with 19-way Holm-Bonferroni correction; chance measured by shuffling the mapping between the transcriptions and the ground-truth sentence texts). Together, these results demonstrate that it is possible to synthesize intelligible speech from the brain activity of a person with paralysis.

3.6 Facial-avatar decoding

Face-to-face audio-visual communication offers multiple advantages over solely audio-based communication. Previous studies show that non-verbal facial gestures often account for a significant portion of the perceived feeling and attitude of a speaker (Mehrabian 1981; Jia et al. 2012) and that face-to-face communication enhances social connectivity (Sadikaj and Moskowitz 2018) and intelligibility (Sumby and Pollack 1954). Therefore, animation of a facial avatar to accompany synthesized speech and further embody the user is a promising means toward naturalistic communication, and it may be possible via decoding of articulatory and orofacial representations in the speech-motor cortex (Chartier et al. 2018; Bouchard et al. 2013; Carey et al. 2017; Mugler et al. 2018). To this end, we developed a facial-avatar BCI to decode neural activity into articulatory speech gestures and render a dynamically moving virtual face during the audio-visual task condition (Figure 3.4a).

To synthesize the avatar's motion, we used an avatar-animation system designed to transform speech signals into accompanying facial-movement animations for applications in games and film (Speech Graphics Ltd, Edinburgh, Scotland). This technology uses speech-to-gesture methods that predict articulatory gestures from sound waveforms then synthesizes the avatar animation from these gestures (Berger et al. 2011). We designed a 3-D virtual environment to display the avatar to our participant during testing. Before testing, the participant selected an avatar from multiple potential candidates.

We implemented two approaches for animating the avatar: a direct approach and an acoustic approach. We used the direct approach for offline analyses to evaluate if articulatory movements could be directly inferred from neural activity without the use of a speech-based intermediate, which has implications for potential future uses of an avatar that are not based on speech representations, including non-verbal facial expressions. We used the acoustic approach for real-time audio-visual synthesis because it provided low-latency synchronization between decoded speech audio and avatar movements.

For the direct approach, we trained a bidirectional RNN with CTC loss to learn a mapping between ECoG features and reference discretized articulatory gestures. These articulatory gestures were obtained by passing the reference waveforms through the animation system's speech-to-gesture model. We then discretized the articulatory gestures using a vector quantized variational autoencoder (VQ-VAE) (A. v. d. Oord et al. 2017). During testing, we used the RNN to decode the discretized articulatory gestures from neural activity and then dequantized them into continuous articulatory gestures using the VQ-VAE's decoder. Finally, we used the gesture-to-animation subsystem to animate the avatar face from the continuous gestures.

We found that the direct approach produced articulatory gestures that were strongly correlated with reference articulatory gestures across all datasets, highlighting the system's ability to decode articulatory information from brain activity.

We then evaluated direct-decoding results by measuring the perceptual accuracy of the avatar. Here, we used a forced-choice perceptual assessment to test whether the avatar animations contained visually salient information about the target utterance. Crowd-sourced evaluators watched silent videos of the decoded avatar animations and were asked to identify to which of two sentences each video corresponded. One sentence was the ground-truth sentence while the other was randomly selected from the set of test sentences. We used the median bootstrapped accuracy across six evaluators to represent the final accuracy for each sentence. We obtained median accuracies of 85.7% (99% CI [79.0, 92.0]), 87.7% (99% CI [79.7, 93.7]), and 74.3% (99% CI [66.7, 80.8]) across the 50-phrase-AAC, 529-phrase-AAC, and 1024-word-General sets, demonstrating that the avatar conveyed perceptually meaningful speech-related facial movements (Figure 3.4b).

Next, we compared the facial-avatar movements generated during direct decoding with real movements made by healthy speakers. We recorded videos of eight healthy volunteers as they read aloud sentences from the 1024-word-General set. We then applied a facial-keypoint recognition model (dlib) (Davis E King n.d.) to avatar and healthy-speaker videos to extract trajectories important for speech; jaw opening, lip aperture, and mouth width. For each pseudo-block of 10 test sentences, we computed the mean correlations across sentences between the trajectory values for each possible pair of corresponding videos (36 total combinations with one avatar and eight healthy-speaker videos). Prior to calculating correlations between two trajectories for the same sentence, we applied dynamic time warping (DTW) to account for variability in timing. We found that the jaw opening, lip aperture, and mouth width of the avatar and healthy speakers were well correlated with median values of 0.733 (99% CI [0.711, 0.748]), 0.690 (99% CI [0.663, 0.714]), and 0.446 (99% CI [0.417, 0.470]) respectively (Figure 3.4c). Although correlations amongst pairs of healthy speakers were higher than between the avatar and healthy speakers (all P $<$ 0.0001, two-sided Mann-Whitney U-test with 9-way Holm-Bonferroni correction) there was a large degree of overlap between the two distributions, illustrating that the avatar reasonably approximated the expected articulatory trajectories relative to natural variances between healthy speakers. Correlations for both distributions were significantly above chance, which was calculated by temporally shuffling the human trajectories and then recomputing correlations with DTW (all P $<$ 0.0001, two-sided Mann-Whitney U-test with 9-way Holm-Bonferroni correction).

Avatar animations rendered in real time using the acoustic approach also exhibited strong correlations between decoded and reference articulatory gestures, high perceptual accuracy, and visual facial-landmark trajectories that were closely correlated with healthy-speaker trajectories. These findings emphasize the strong performance of the speech-synthesis neural decoder when used with the speech-to-gesture rendering system, although this approach cannot be used to generate meaningful facial gestures in the absence of a decoded speech waveform.

In addition to articulatory gestures to visually accompany synthesized speech, a fully embodying avatar BCI would also enable the user to portray non-speech orofacial gestures, including movements of particular orofacial muscles and expressions that convey emotion (Salari et al. 2020). To this end, we collected neural data from our participant as she performed two additional tasks: an articulatory-movement task and an emotional-expression task. In the articulatory-movement task, the participant attempted to produce 6 orofacial movements (Figure 3.4d). In the emotional-expression task, the participant attempted to produce 3 types of expressions — happy, sad, and surprised — with either low, medium, or high intensity, resulting in 9 unique expressions. Offline, for the articulatory-movement task we trained a small feed-forward neural-network model to learn the mapping between the ECoG features and each of the targets. For the articulatory-movement task, we observed a median classification accuracy of 87.8% (99% CI [85.1, 90.5]; across n=10 cross-validation folds; Figure 3.4d) when classifying between the 6 movements. For the emotional-expression task, we trained a small RNN to learn the mapping between ECoG features and each of the expression targets. We observed a median classification accuracy of 74.0% (99% CI [70.8, 77.1]; across n=15 cross-validation folds; Figure 3.4e) when classifying between the 9 possible expressions and a median classification accuracy of 96.9% (99% CI [93.8,100]) when only considering the classifier's outputs for the strong-intensity versions of the 3 expression types. In separate, qualitative task blocks, we showed that the participant could control the avatar BCI to portray the articulatory movements and strong-intensity emotional expressions, illustrating the potential of multimodal communication BCIs to restore the ability to express meaningful orofacial gestures.

3.7 Articulatory representations drive decoding

In healthy speakers, neural representations in the sensorimotor cortex (SMC, comprising the precentral and postcentral gyri) encode articulatory movements of the orofacial musculature (Chartier et al. 2018; Carey et al. 2017; Eichert et al. 2020). With the implanted electrode array centered over the SMC of our participant, we hypothesized that articulatory representations persisting after paralysis underlied speech-decoding performance. To assess this, we fit a linear temporal receptive-field encoding model to predict HGA for each electrode from the phone probabilities computed by the text decoder during the 1024-word-General text task condition. For each speech-activated electrode, we calculated the maximum encoding weight for each phone, yeilding a phonetic-tuning space where each electrode had an associated vector of phone-encoding weights. Within this space, we determined if phone clustering was organized by the primary orofacial articulator of each phone (place of articulation, POA; Figure 3.5a), which has been shown in prior studies with healthy speakers (Chartier et al. 2018; Bouchard et al. 2013). We parceled phones into four POA categories: labial, vocalic, back tongue, and front tongue. Hierarchical clustering of phones revealed grouping by POA ($P < 0.0001$ compared to chance, one-tailed permutation test; Figure 3.5b). We observed a variety of tunings across the electrodes, with some electrodes exhibiting tuning to single POA categories and others to multiple categories (such as both front-tongue and back-tongue phones or both labial and vocalic phones; Figure 3.5c). We visualized the phonetic tunings in a two-dimensional space, revealing separability between labial and non-labial consonants (Figure 3.5d) and between lip-rounded and non-lip-rounded vowels (Figure 3.5e).

Next, we investigated whether these articulatory representations were arranged somatotopically (with ordered regions of cortex preferring single articulators), which is observed in healthy speakers (Bouchard et al. 2013). Because the dorsal-posterior corner of our ECoG array provided coverage of the hand cortex, we also assessed how neural activation patterns related to attempted hand movements fit into the somatotopic map, using data collected during the NATO-motor task containing four finger flexion targets (either thumb or simultaneous index- and middle-finger flexion for each hand). We visualized the grid locations of the electrodes that most strongly encoded the vocalic, front-tongue, and labial phones as well attempted hand movement (the top 30% of electrodes having maximal tuning for each condition; Figure 3.5f). Kernel density estimates revealed a somatotopic map with encoding of attempted hand movements, labial phones, and front-tongue phones organized along a dorsal-ventral axis. The relatively anterior localization of the vocalic cluster in the precentral gyrus is likely associated with the laryngeal motor cortex, consistent with previous investigations in healthy speakers (Bouchard et al. 2013; Carey et al. 2017; Breshears et al. 2015).

Next, we assessed whether the same electrodes that encoded POA categories during silent speech attempts also encoded non-speech articulatory-movement attempts. Using the previously computed phonetic encodings and HGA recorded during the articulatory movement task, we found a positive correlation between front-tongue phonetic encoding and HGA magnitude during attempts to raise the tongue ($P < 0.0001$, $r=0.84$, ordinary least squares regression; Figure 3.5g). We also observed a positive correlation between labial phonetic tuning and HGA magnitude during attempts to pucker the lips ($P < 0.0001$, $r = 0.89$, ordinary least squares regression; Figure 3.5h). Although most electrodes were selective to either lip or tongue movements, others were activated by both (Figure 3.5i). Together, these findings suggest that, after 18 years of paralysis, our participant's SMC maintains general-purpose articulatory encoding that is not speech specific and contains representations of non-verbal emotional expressions and articulatory-movements (see Figure 3.4). During the NATO-motor task, electrodes encoding attempted finger flexions were largely orthogonal to those encoding NATO code words, which helped to enable accurate neural discrimination between the four finger-flexion targets and the silent-speech targets (the model correctly classified 569 out of 570 test trials as either finger flexion or silent speech).

To characterize the relationship between encoding strength and importance during decoding, we computed a contribution score for each electrode and decoding modality by measuring the effect of small perturbations to the electrode's activity on decoder predictions, as in previous work (David A. Moses et al. 2021; Metzger et al. 2022; Simonyan et al. 2014) (Figure 3.12a-c). We noted that many important electrodes were adjacent, suggesting sampling of useful, non-redundant information from the cortex despite the electrodes' close proximity. We also observed degraded performance during an offline simulation of low-density sampling, further highlighting the benefit of high-density cortical recording. As we hypothesized, many of the highest-contributing electrodes also exhibited substantial articulatory-feature encoding defined in Figure 3.5 and were similarly important for all three modalities (Figure 3.12e-g). Indeed, the brain areas that most strongly encoded POA, notably the SMC, were the most critical to decoding performance in leave-one-area-out offline analyses (Figure 3.13).

These results are in line with growing evidence for motor-movement encoding in the postcentral gyrus (Umeda et al. 2019; Murray and Coulter 1981; Arce et al. 2013), which is further supported by an analysis of peak-activation times that revealed no significant difference between electrodes in the precentral versus postcentral gyrus during silent attempts to speak ($P > 0.01$ two-sided Mann-Whitney U-test) (Umeda et al. 2019; Murray and Coulter 1981; Arce et al. 2013). Interestingly, temporal-lobe electrodes contributing to decoding were not strongly activated during auditory perception ($r < 0.1$, $P > 0.01$, Pearson correlation permutation test), suggesting they may record cortical activity from the subcentral gyrus (Eichert et al. 2020) or production-specific sites within the temporal lobe (**binder˙current˙2017**).

3.8 Discussion

Faster, more accurate, and more natural communication are among the most desired needs of people who have lost the ability to speak after severe paralysis (Peters et al. 2015; Rousseau et al. 2015; Felgoise et al. 2016; Huggins et al. 2011). Here, we have demonstrated that all of these needs can be addressed with a speech-neuroprosthetic device that decodes articulatory cortical activity into multiple output modalities, including text, speech audio synchronized with a facial avatar, and facial expressions.

During 14 days of data collection shortly after device implantation, we achieved high performance text decoding, exceeding communication speeds of previous brain-computer interfaces (BCIs) by a factor of 4 or more 1,3,9 and expanding the vocabulary size of our prior direct-speech BCI by a factor of 20 1. We also showed for the first time that intelligible speech can be synthesized from the brain activity of a person with paralysis. Finally, we introduced a novel modality of BCI control in the form of a digital "talking face" — a personalized avatar capable of dynamic, realistic, and interpretable speech and non-verbal facial gestures. Together, we believe that these results have surpassed an important threshold of performance, generalizability, and expressivity that could soon have practical benefits to people with speech loss.

The progress here was enabled by several key innovations and findings: 1) Advances in the neural interface, providing denser and broader sampling of the distributed orofacial and vocal-tract representations across the lateral sensorimotor cortex; 2) Highly stable recordings from non-penetrating cortical-surface electrodes, enabling training and testing across days and weeks without requiring day-of recalibration; 3) Custom sequence-learning neural-decoding models, facilitating training without alignment of neural activity and output features 4) Self-supervised learning-derived discrete speech units, serving as effective intermediate representations for intelligible speech synthesis; 5) Control of a virtual face from brain activity to accompany synthesized speech and convey facial expressions; and 6) Persistent articulatory encoding in the SMC of our participant that is consistent with prior intact-speech characterizations despite over 18 years of anarthria, including hand and orofacial-motor somatotopy organized along a dorsal-ventral axis and phonetic tunings clustered by place of articulation.

A limitation of the present proof-of-concept study is that the results shown are from only one participant. An important next step is to validate these decoding approaches in other individuals with varying degrees and etiologies of paralysis, for example patients who are fully locked-in with ALS (Pandarinath et al. 2017; Bruurmijn et al. 2017). Furthermore, while we were able to train decoders without the participant hearing or seeing the targets for synthesis and avatar, providing instantaneous closed-loop feedback during decoding has the potential to improve user engagement, model performance, and neural entrainment 42,43. Also, further advances in electrode interfaces 44 to enable denser and broader cortical coverage should continue to improve accuracy and generalizability towards eventual clinical applications.

The ability to interface with evolving technology to communicate with family and friends, facilitate community involvement and occupational participation, and engage in virtual, internet-based social contexts (such as social media and metaverses) can vastly expand a person's access to meaningful interpersonal interactions and ultimately improve their quality of life (Peters et al. 2015; Felgoise et al. 2016). We show here that BCIs can give this ability back to patients through highly personalizable audio-visual synthesis capable of restoring aspects of their personhood and identity. This is further supported by our participant's feedback on the technology, in which she describes how a multimodal BCI would improve her daily life by increasing expressivity, independence, and productivity. A major goal now is to move beyond these initial demonstrations and build seamless integration with real-world applications.

3.9 Methods

Clinical-trial overview

This study was completed within the BCI Restoration of Arm and Voice (BRAVO) clinical trial (ClinicalTrials.gov; NCT03698149). The primary endpoint of this trial is to assess the long-term safety and tolerability of an electrocorticography (ECoG)-based interface through the measurement of treatment-emergent adverse events. All data presented here are part of the ongoing exploratory clinical trial and do not contribute toward any conclusions regarding the primary safety endpoints of the trial. The clinical trial began in November 2018, with all data in this present work collected in 2022 and 2023. Following the Food and Drug Administration's investigational device exemption approval for the neural-implant device used in this study, the study protocol was approved by the University of California, San Francisco Institutional Review Board. The participant gave her informed consent to participate in this trial following multiple conversations with study investigators where the details of study enrollment, including risks related to the study device, were thoroughly explained to her. The original and current clinical protocols are provided as a supplementary file alongside this article.

Participant

The participant, who was 47 years old at time of enrollment into the study, was diagnosed with quadriplegia and anarthria by neurologists and a speech-language pathologist following a right-pontine infarct in 2005. When the participant was 30 years old and in good health, she experienced sudden onset dizziness, slurred speech, quadriplegia and bulbar weakness. She was found to have a large pontine infarct with left vertebral artery dissection and basilar artery occlusion. During enrollment testing, she scored 29/30 on the Mini Mental State Exam and was only unable to achieve the final point because she could not physically draw a figure due to her paralysis. She can vocalize a small set of monosyllabic sounds, such as ”ah” or ”ooh”, but she is unable to articulate intelligible words. During clinical assessments, a speech-language pathologist prompted her to say 58 words and 10 phrases and also asked her to respond to 2 open-ended questions within a structured conversation. From the resulting audio and video transcriptions of her speech attempts, the speech-language pathologist measured her intelligibility to be 5% for the prompted words, 0% for the prompted sentences, and 0% for the open-ended responses. To investigate how similar her movements during silent speech attempts were relative to neurotypical speakers, we applied a state-of-the-art visual-speech recognition 45 model to videos of the participant's face during imagined, silently attempted, and vocal attempted speech. We found a median WER of 95.8% (99% CI [90.0, 125.0]) for silently attempted speech, which was far higher than the median WER from videos of volunteer healthy speakers, which was 50.0% (99% CI [37.5, 62.5]). Functionally, she cannot use speech to communicate. Instead, she relies on a transparent letter board and a Tobii Dynavox for communication. She used her transparent letter board to provide informed consent to participate in this study and to allow her image to appear in demonstration videos. To sign the physical consent documents, she used her communication board to spell out ”I consent” and directed her spouse to sign the documents on her behalf.

Neural implant

The neural-implant device used in this study featured a high-density ECoG array (PMT) and a percutaneous pedestal connector (Blackrock Microsystems). The ECoG array consists of 253 disk-shaped electrodes arranged in a lattice formation with 3-mm center-to-center spacing. Each electrode has a 1-mm recording-contact diameter and a 2-mm overall diameter. The array was surgically implanted subdurally on the pial surface of the left hemisphere of the brain, covering regions associated with speech production and language perception, including the middle aspect of the superior and middle temporal gyri, the precentral gyrus, and the postcentral gyrus. Pre-operative functional magnetic resonance imaging with standard clinical speech and motor tasks was performed, which also indicated hand-motor encoding in the brain area covered by the dorsal part of the array. The percutaneous pedestal connector, which was secured to the skull during the same operation, conducts electrical signals from the ECoG array to a detachable digital headstage and HDMI cable (CerePlex E256; Blackrock Microsystems). The digital headstage minimally processes and digitizes the acquired cortical signals and then transmits the data to a computer for further signal processing. The device was implanted in September 2022 at UCSF Medical Center with no surgical complications.

Signal processing

We used the same signal-processing pipeline detailed in our previous work (Metzger et al. 2022) to extract high-gamma activity (HGA) and low-frequency signals (LFS) from the ECoG signals at a 200-Hz sampling rate. Briefly, we first apply common average referencing to the digitized ECoG signals and downsample them to 1kHz after applying an anti-aliasing filter with a cutoff of 500Hz. Then we compute HGA as the analytic amplitude of these signals after band-passing them in the high-gamma range (70–150 Hz), then downsample them to 200Hz. For LFS, we only apply a low-pass anti-aliasing filter with a cutoff frequency of 100 Hz, then downsample signals to 200Hz. For data normalization, we applied a 30-second sliding-window z-score in real time to the HGA and LFS features from each ECoG channel.

We performed all data collection and real-time decoding tasks in the common area of the participant's residence. We used a custom Python package named rtNSR, which we created in prior work but have continued to augment and maintain over time (David A. Moses et al. 2021; Metzger et al. 2022; David A Moses et al. 2018), to collect and process all data, run the tasks, and coordinate the real-time decoding processes. After each session, we uploaded the neural data to our lab's server infrastructure, where we analyzed the data and trained decoding models.

Task design

Experimental paradigms

To collect training data for our decoding models, we implemented a task paradigm in which the participant attempted to produce prompted targets. In each trial of this paradigm, we presented the participant with text representing a speech target (for example, “Where was he trying to go?”) or a non-speech target (for example, “Lips back”). The text was surrounded by three dots on both sides, which sequentially disappeared to act as a countdown. After the final dot disappeared, the text turned green to indicate the go cue, and the participant attempted to silently say that target or perform the corresponding action. After a brief delay, the screen cleared and the task continued to the next trial.

During real-time testing, we used three different task conditions: text, audio-visual, and NATO-motor. We used the text task condition to evaluate the text decoder. In this condition, we used the top half of the screen to present prompted targets to the participant, similar to what we used for training. We used the bottom half of the screen to display an indicator (three dots) when the text decoder first predicted a non-silence phone, which we updated to the full decoded text once the sentence was finalized.

We used the audio-visual task condition to evaluate the speech-synthesis and avatar-animation models, including the articulatory-movement and emotional-expression classifiers. In this condition, the participant attended to a screen showing the Unreal Engine environment that contained the avatar. The viewing angle of the environment was focused on the avatar's face. In each trial, speech and non-speech targets appeared on the screen as white text. After a brief delay, the text turned green to indicate the go cue, and the participant attempted to silently say that target or perform the corresponding action. Once the decoding models processed the neural data associated with the trial, the decoded predictions were used to animate the avatar and, if the current trial presented a speech target, play the synthesized speech audio.

We used the NATO-motor task condition to evaluate the NATO code-word classification model and to collect neural data during attempted hand-motor movements. This task contained 26 speech targets (the code words in the NATO phonetic alphabet) and 4 non-speech hand-motor targets (left-thumb flexion, right-thumb flexion, right index- and middle-finger flexion, and left index- and middle-finger flexion). We instructed the participant to attempt to perform the hand-motor movements to the best of her ability despite her severe paralysis. This task condition resembled the text condition, except that the top three predictions from the classifier (and their corresponding predicted probabilities) were shown in the bottom half of the screen as a simple horizontal bar chart after each trial. We used the prompted-target paradigm to collect the first few blocks of this dataset, and then we switched to the NATO-motor task condition to collect all subsequent data and to perform real-time evaluation.

Sentence sets

We used three different sentence sets in this work: "50-phrase-AAC", "529-phrase-AAC", and "1024-word-General." The first two sets contained sentences that are relevant for general dialogue as well as augmentative and alternative communication (AAC) 4. The 50-phrase-AAC set contained 50 sentences composed of 119 unique words, and the 529-phrase-AAC set contained 529 sentences composed of 372 unique words and included all of the sentences in the 50-phrase-AAC set. The 1024-word-General set contained sentences sampled from Twitter and movie transcriptions for a total of 13,463 sentences and 1,024 unique words.

To create the 1024-word-General sentence set, we first extracted sentences from the nltk Twitter corpus (Bird et al. 2009) and the Cornell movies corpus (Danescu-Niculescu-Mizil and L. Lee 2011). We drew 18,284 sentences from this corpora that were composed entirely from the 1,152-word vocabulary from our previous work 3, which contained common English words. We then subjectively pruned out offensive sentences, sentences that grammatically did not make sense, and sentences with overly negative connotation, and kept sentences between 4-8 words, which resulted in 13,463 sentences composed of a total of 1,024 unique words. Partway through training, we removed sentences with syntactic pauses or punctuation in the middle. Of these sentences, we were able to go through 9,506 with our participant for use during the training of text and avatar models. We used 95% of this data to train the models and 5% as a held-out development set to evaluate performance and choose hyperparameters prior to real-time testing. Because the synthesis model required several days to train to convergence, this model only used 6,449 trials for training data since the remaining trials were collected while the model was training. Of these trials, 100 were used as a held-out development set to evaluate performance and choose hyperparameters prior to real-time testing.

We randomly selected 249 sentences from the 1024-word-General set to use as the final test sentences for text decoding. We did not collect training data with these sentences as targets. For evaluation of audio-visual synthesis and avatar, we randomly selected 200 sentences that were not used during training and were not included in the 249 sentences used for text-decoding evaluation. As a result of the prior reordering, the audio-visual synthesis and avatar test sets contained a larger proportion of common words.

For training and testing with the 1024-word-General sentence set, to help the decoding models infer word boundaries from the neural data without forgoing too much speed and naturalness, we instructed the participant to insert small syllable-length pauses (approximately 300–500 ms) between words during her silent speech attempts. For all other speech targets, we instructed the participant to attempt to silently speak at her natural rate.

Text decoding

Phone decoding

For the text-decoding models, we downsampled the neural signals by a factor of 6 (from 200 Hz to 33.33 Hz) after applying an anti-aliasing low-pass filter at 16.67 Hz using the Scipy python package (Virtanen et al. 2020), as in previous work (David A. Moses et al. 2021; Metzger et al. 2022). We then normalized the high-gamma activity and low-frequency signals separately to have an L2-norm of 1 across all time steps for each channel. We used all available electrodes during decoding.

We trained a recurrent neural network (RNN) to model the probability of each phone at each time step, given these neural features. We trained the RNN using the connectionist temporal classification (CTC) loss 5 to account for the lack of temporal alignment between neural activity and phone labels. The CTC loss maximizes the probability of any correct sequence of phone outputs that correspond to the phone transcript of a given sentence. To account for differences in the length of individual phones, the CTC loss collapses over consecutive repeats of the same phone. For example, predictions corresponding to /w ah z/ — the phonetic transcription of "was" — could be a result of the RNN predicting the following valid time series of phones: /w ah z z/, /w w ah ah z/ z/, /w w ah z/, and so forth.

We determined reference sequences using g2p-en K. Park and Kim 2019, a grapheme-to-phoneme model that enabled us to recover phone pronunciations for each word in the sentence sets. We inserted a silence token in between each word and at the beginning and end of each sentence. For simplicity, we used a single phonetic pronunciation for each word in the vocabulary. We used these sentence-level phone transcriptions for training and to measure performance during evaluation.

The RNN itself contained a convolutional portion followed by a recurrent portion, which is a commonly used architecture in automatic speech recognition 52,53. The convolutional portion of our RNN was composed of a 1-D convolutional layer with 500 kernels, a kernel size of 4, and a stride of 4. The recurrent portion was composed of 4 layers of bidirectional gated recurrent units with 500 hidden units. The hidden states of the final recurrent layer were passed through a linear layer and projected into a 41-dimensional space. These values were then passed through a softmax activation function to estimate the probability of each of the 39 phones, the silence token, and the CTC blank token (used in the CTC loss to predict two tokens in a row or to account for silence at each time step) Graves et al. 2006. We implemented these models using the PyTorch Python package (version 1.10.0) Paszke et al. 2019

We trained the RNN to predict phone sequences using an 8-second window of neural activity. To improve the model's robustness to temporal variability in the participant's speech attempts, we introduced jitter during training by randomly sampling a continuous 8-second window from a 9-second window of neural activity spanning from 1 second before to 8 seconds after the go cue, as in previous work David A. Moses et al. 2021; Metzger et al. 2022. During inference, the model used a window of neural activity spanning from 500 ms before to 7.5 seconds after the go cue. To improve communication rates and decoding of variable-length sentences, we terminated trials before a full 8-second window if the decoder determined the participant had stopped attempted speech by using silence detection. Here, we use "silence" to refer to the absence of an ongoing speech attempt; all of the participant's attempts to speak were technically silent, so the "silence" described here can be thought of as idling. To implement this early-stopping mechanism, we performed the following steps: 1) Starting 1.9 s after the go cue and then every 800 ms afterwards, we used the RNN to decode the neural features acquired up to that point in the trial; 2) If the RNN predicted the silence token for the most recent 8 time steps (960 ms) with over 88.8% average probability (or, in

2 out of the 249 real-time test trials, if the 7.5-second trial duration expired), the current sentence prediction was used as the final model output and the trial ended. We attempted a version of the task where the current decoded text was presented to the participant every 800 ms; however, the participant generally preferred only seeing the finalized decoded text.

Beam-search algorithm

We used a CTC beam-search algorithm to transform the predicted phone probabilities into text (Collobert et al. 2016). To implement this CTC beam search, we used the ctc decode function in the torchaudio Python package (Yang et al. 2022). Briefly, the beam search finds the most likely sentence given the phone probabilities emitted by the RNN. For each silent speech attempt, the likelihood of a sentence is computed as the emission probabilities of the phones in the sentence combined with the probability of the sentence under a language-model prior. We used a custom-trained 5-gram language model with Kneser-ney smoothing (Kneser and Ney 1995). We used the KenLM software package (Heafield 2011) to train the 5-gram language model on the full 18,284 sentences that were eligible to be in the 1024-word-General set prior to any pruning. The 5-gram language model is trained to predict the probability of each word in the vocabulary based on the previous words (up to 4). We chose this approach because the linguistic structure and content of conversational tweets and movie lines are more relevant for everyday usage than formal written language commonly used in many standard speech-recognition databases (Panayotov et al. 2015; Ito and Johnson 2017). The beam search also uses a lexicon to restrict phone sequences to form valid words within a limited vocabulary. Here, we used a lexicon defined by passing each word in the vocabulary through a grapheme-to-phoneme conversion module (g2p-en) to define a valid pronunciation for each word. We used a language model weight of 4.5 and a word insertion score of -0.26.

Decoding speed

To measure decoding speed during real-time testing, we used the formula NT, where N is the number of words in the decoded output and T is the time (in minutes) that our participant was attempting to speak. We calculated T by computing the elapsed time between the appearance of the go cue and the time of the data sample which immediately preceded the samples that triggered early stopping, giving the resulting formula: rate $= \frac{N}{t_{silencedetected} - t_{gocue}}$

Here, N remains the number of words in the decoded output. tsilence detected is the time of the data sample which immediately preceded the samples that triggered early stopping, and tgo cue is the time when the go cue appeared.

Error-rate calculation

Word error rate (WER) is defined as the word edit distance, which is the minimum number of word deletions, insertions, and substitutions required to convert the decoded sentence into the target (prompted) sentence, divided by the number of words in the target sentence. Phone error rate (PER) and character error rate (CER) are defined analogously for phones and characters, respectively. When measuring PERs, we ignored the silence token at the start of each sentence, since this token is always present at the start of both the reference phone sequence and the phone decoder's output.

For brain-computer interfaces, error-rate distributions are typically assessed across sets of 5 or more sentences rather than single trials, since single-trial error rates can be noisy and are highly dependent on sentence length (David A. Moses et al. 2021; Metzger et al. 2022; Willett et al. 2021b). Hence, we sequentially parceled sentences into pseudo-blocks of 10 sentences and then evaluated error rates and other metrics across these pseudo-blocks. As in previous work (Metzger et al. 2022; Willett et al. 2021b), this entailed taking the sum of the phone, word, and character edit distances between each of the predicted and target sentences in a given pseudo-block, and dividing it by the total number of phones, words, or characters across all target sentences in the block, respectively. In the single case where a pseudo block contained an invalid trial, that trial was ignored.

Offline simulation of large-vocabulary, 50-phrase-AAC, and 500-phrase-AAC results

To simulate text-decoding results using the larger vocabularies, we used the same neural activity, RNN decoder, and start and end times that were used during real-time evaluation. We only changed the underlying 5-gram language model to be trained on all sentences 4 to 8 words in length in the Twitter and Cornell movies corpora that fell within the desired vocabulary. We evaluated performance using log-spaced vocabulary sizes consisting of 1,506, 2,270, 3,419, 5,152, 7,763, 11,696, 17,621, 11,696, 26,549, and 39,378 words, and also included the real-time results (1024 words). To choose the words at each vocabulary size, with the exception of the already defined vocabulary for the real-time results, we first included all words in the 1024-word-General set. Then, we used a readily available pronunciation dictionary from the Librispeech Corpus Panayotov et al. 2015 to select all words which were present in both the Twitter and Cornell movies corpora and the pronunciation dictionary. The most frequent words which were not in the 1024-word-General set but fell within the pronunciation dictionary were added to reach the target vocabulary size. We then simulated the results on the task with the larger vocabulary and language model.

To simulate text-decoding results on the 50-phrase-AAC and 500-phrase-AAC sentence sets (because we only tested the text decoder in real-time with the 1024-word-General set), we trained RNN decoders on data associated with these two AAC sets. We then simulated decoding using the neural data and go cues from the real-time blocks used for evaluation of the avatar and synthesis methods. We checked for early stopping 2.2 s after the start of the sentence and again every subsequent 350 ms. Once an early stop was detected, or if 5.5 seconds had elapsed since the go cue, we finalized the sentence prediction. During decoding, we applied the CTC beam search using a 5-gram language model fit on the phrases from that set.

Decoding NATO code words and hand-motor movements

We used the same neural-network decoder architecture (but with a modified input and output layer dimensionality to account for differences in the number of electrodes and target classes) as in previous work 3 to output the probability of each of the 26 NATO code words and the four hand-motor targets. To maximize data efficiency, we used transfer learning between our participants; we initialized the decoder using weights from our previous work, and we replaced the first and last layers to account for differences in the number of electrodes and number of classes being predicted, respectively.For the results shown in Figure 3.2h, we computed NATO code-word classification accuracy using a model that was also capable of predicting the motor targets; here, we only measured performance on trials in which the target was a NATO code word, and we deemed incorrect any such trial in which a code-word attempt was misclassified as a hand-motor attempt.

Speech synthesis

Training and inference procedure

We used CTC loss to train an RNN to predict a temporal sequence of discrete speech units extracted using HuBERT (Hsu et al. 2021b) from neural data. HuBERT is a speech-representation learning model that is trained to predict acoustic k-means-cluster identities corresponding to masked timepoints from unlabeled input waveforms. We refer to these cluster identities as discrete speech units, and the temporal sequence of these speech units represents the content of the original waveform.

Because our participant cannot speak, we generated reference sequences of speech units by applying HuBERT to a speech waveform which we refer to as the basis waveform. For the

50-phrase-AAC and 529-phrase-AAC sets, we acquired basis waveforms from a single male speaker (recruited prior to our participant's enrollment in the trial) who was instructed to read each sentence aloud in a consistent manner. Due to the large number of sentences in the 1024-word-General set, we used the Wavenet text-to-speech model (A. v. d. Oord et al. 2016) to generate basis waveforms.

We used HuBERT to process our basis waveforms and generate a series of reference discrete speech units sampled at 50 Hz. We used the base 100-unit, 12-transformer-layer HuBERT trained on 960 hours of LibriSpeech (Panayotov et al. 2015), which is available in the open-source fairseq library (Ott et al. 2019). In addition to the reference discrete speech units, we added the blank token needed for CTC decoding as a target during training.

The synthesis RNN, which we trained to predict discrete speech units from the ECoG features (high-gamma activity and low-frequency signals), consisted of the following layers (in order): 1) A 1-D convolutional layer, with 260 kernels with width and stride of 6; 2) 3 layers of bidirectional gated recurrent units, each with a hidden dimension size of 260; and 3) a 1-D transpose convolutional layer, with a size and stride of 6, that output discrete-unit logits. To improve robustness, we applied data augmentations using the SpecAugment method (D. S. Park et al. 2019) to the ECoG features during training.

From the ECoG features, the RNN predicted the probability of each discrete unit every 5 ms. We only retained the most likely predicted unit at each time step. We ignored time steps where the CTC blank token was decoded, as this is primarily used to adjust for alignment and repeated decodes of discrete units. Next, we synthesized a speech waveform from the sequence of discrete speech units, using a pre-trained unit-to-speech vocoder (A. Lee et al. 2022).

During each real-time inference trial in the audio-visual task condition, we provided the speech-synthesis model with ECoG features collected in a time window around the go cue. This time window spanned from 0.5 seconds before to 4.62 seconds after the go cue for the 50-phrase-AAC the 529-phrase-AAC sentence sets and from 0 seconds before to 7.5 seconds after the go cue for the 1024-word-General sentence set. The model then predicted the most likely sequence of HuBERT units from the neural activity and generated the waveform using the aforementioned vocoder. We streamed the waveform in 5-ms chunks of audio directly to the real-time computer's sound card via the PyAudio Python package.

To decode speech waveforms in the participant's personalized voice (that is, a voice designed to resemble the participant's own voice before her injury), we used YourTTS Casanova et al. 2022, a zero-shot voice conversion model. After conditioning the model on a short clip of our participant's voice extracted from a pre-injury video of her, we applied the model to the decoded waveforms to generate the personalized waveforms (Figure 3.10). To reduce the latency of the personalized speech synthesizer during real-time inference for a qualitative demonstration, we trained a HiFi-CAR convolutional neural network67 to vocode HuBERT units into personalized speech . This model used voice converted LJSpeech (via YourTTS) as training data.

Evaluation

To evaluate the quality of the decoded speech, we computed the mel-cepstral distortion (MCD) between the decoded and reference waveforms (y and y, respectively) Kubichek 1993. This is defined as the squared error between dynamically time warped sequences of mel cepstra (mcd where d is the index of the mel cepstra) extracted from the target and decoded waveforms and is commonly used to evaluate the quality of synthesized speech:

$$MCD(\hat{y}, y) = \frac{10}{\log(10)}\sqrt{(mc_d^y + mc_d^{\hat{y}})}$$

We excluded silence time points at the start and end of each waveform during MCD calculation. For each pseudo-block, we combined the MCD of 10 individual trials by taking their mean.

We designed a perceptual assessment using a crowd-sourcing platform (Amazon Mechanical Turk), where each test trial was assessed by 12 evaluators (except for 3 of the 500 trials, in which only 11 workers completed their evaluations). In each evaluation, the evaluator listened to the decoded speech waveform and then transcribed what they heard. For each sentence, we then computed the WER and CER between the evaluator's transcriptions and the ground-truth transcriptions. To control for outlier evaluator performance, for each trial, we used the median WER and CER across evaluators as the final accuracy metric for the decoded waveform. We reported metrics across pseudo-blocks of 10 sentences to be consistent with text-decoding evaluations and calculated WER across each pseudo-block in the same manner as for text decoding

Avatar

Articulatory-gesture data

We used a dataset of articulatory gestures for all sentences from the 50-phrase-AAC, 529-phrase-AAC, and 1024-word-general datasets provided by Speech Graphics. We generated these articulatory gestures from reference waveforms using Speech Graphics' speech-to-gesture model, which was designed to animate avatar movements given a speech waveform. For each trial, articulatory gestures consisted of 16 individual gesture time series corresponding to jaw, lip, and tongue movements.

Offline training and inference procedure for the direct avatar-animation approach

To perform direct decoding of articulatory gestures from neural activity (the direct approach for avatar animation), we first trained a vector-quantized variational autoencoder (VQ-VAE) to encode continuous Speech Graphics' gestures into discrete articulatory-gesture units (A. v. d. Oord et al. 2017). A VQ-VAE is composed of an encoder network that maps a continuous feature space to a learned discrete codebook and a decoder network that reconstructs the input using the encoded sequence of discrete units. The encoder was composed of 3 layers of 1-D convolutional units with 40 filters, a kernel size 4, and a stride of 2. Rectified linear unit (ReLU) activations followed the second and third of these layers. After this step, we applied a 1-D convolution, with 1 filter and a kernel size and stride of 1, to generate the predicted codebook embedding. We then used nearest-neighbor lookup to predict the discrete articulatory-gesture units. We used a codebook with 40 different 1-D vectors, wherein the index of the codebook entry with the smallest distance to the encoder's output served as the discretized unit for that entry. We trained the VQ-VAE's decoder to convert discrete sequences of units back to continuous articulatory gestures by associating each unit with the value of the corresponding continuous 1-D codebook vector. Next, we applied a 1-D convolution layer, with 40 filters and a kernel size and stride of 1, to increase the dimensionality. Then, we applied 3 layers of 1-D transpose convolutions, with 40 filters, a kernel size of 4, and a stride of 2, to upsample the reconstructed articulatory gestures back to their original length and sampling rate. ReLU activations followed the first and second of these layers. The final 1-D transpose convolution had the same number of kernels as the input signal (16). We used the output of the final layer as the reconstructed input signal during training.

To encourage the VQ-VAE units to decode the most critical gestures (such as jaw opening) rather than focusing on those that are less important (such as nostril flare), we weighted the mean-squared error (MSE) loss for the most important gestures more highly. We upweighted the jaw opening's MSE loss by a factor of 20, and the gestures associated with important tongue movements (tongue body raise, tongue advance, tongue retraction, tongue tip raise) and lip movements (rounding and retraction) by a factor of 5. We trained the VQ-VAE using all of the reference articulatory-gestures from the 50-phrase-AAC, 529-phrase-AAC, and 1024-word-General sentence sets. We excluded from VQ-VAE training any sentence that was used during the evaluations with the 1024-word-General set.

To create the CTC decoder, we trained a bidirectional RNN to predict reference discretized articulatory-gesture units given neural activity. We first downsampled the ECoG features by a factor of 6 to 33.33 Hz. We then normalized these features to have an L2-norm of 1 at each time point across all channels. We used a time window of neural activity spanning from 0.5 s before to 7.5 seconds after the go cue for the 1024-word-General set and from 0.5 s before to 5.5 seconds after for the 50-phrase-AAC and 529-phrase-AAC sets. The RNN then processed these neural features using the following components: 1) A 1-D convolution layer, with 256 filters with kernel size and stride of 2; 2) 3 layers of gated recurrent units, each with a hidden dimension size of 512; and 3) A dense layer, which produced a 41-dimensional output. We then used the softmax activation function to output the probability of the 40 possible discrete units (determined by the VQ-VAE) as well as the CTC blank token. The model hyperparameters stated here are for the 1024-word-General sentence set.

During inference, the RNN yielded a predicted probability of each discretized articulatory-gesture unit at every 60 ms. To transform these output probabilities into a sequence of discretized units, we only retained the most probable unit at each time step. We used the decoder module of the frozen VQ-VAE to transform collapsed sequences of predicted discrete articulatory units (here, “collapsed” means that consecutive repeats of the same unit were removed) into continuous articulatory gestures.

Real-time acoustic avatar-animation approach

During real-time testing, we animated the avatar using avatar-rendering software (referred to as SG Com; provided by Speech Graphics). This software converts a stream of speech audio into synchronized facial animation with a latency of 50 ms. It performs this conversion in two steps: First, it uses a custom speech-to-gesture model to map speech audio to a time series of articulatory-gesture activations; Then, it performs a forward mapping from articulatory-gesture activations to animation parameters on a 3-D MetaHuman character created by Epic Games, Inc (Cary, North Carolina). The output animation was rendered using Unreal Engine 4.26.

For every 10 ms of input audio, the speech-to-gesture model produces a vector of articulatory-gesture activation values, each between 0 and 1 (where 0 is fully relaxed and 1 is fully contracted). The forward mapping converts these activations into deformations, simulating the effects of the articulatory gestures on the avatar face. Because each articulatory gesture approximates the superficial effect of some atomic action, such as opening the jaw or pursing the lips, the gestures are analogous to the Action Units of the Facial Action Coding System 70, a well-known method for taxonomizing human facial movements. However, these articulatory gestures from Speech Graphics are more oriented toward speech articulation and also include tongue movements, containing 16 speech-related articulatory gestures (10 for lips, 4 for tongue, 1 for jaw, and 1 for nostril). The system does not generate values for aspects of the vocal tract that are not externally visible, such as the velum, pharynx, or larynx.

To provide avatar feedback to the participant during real-time testing in the audio-visual task condition, we streamed 10-ms chunks of decoded audio over an Ethernet cable to a separate machine running the avatar processes to animate the avatar in synchrony with audio synthesis. We imposed a 200-ms delay on the audio output in real-time to improve perceived synchronization with the avatar.

The avatar-rendering system also generates non-verbal motion, such as emotional expressions, head motion, eye blinks, and eye darts. These are synthesized using a superset of the articulatory gestures involving the entire face and head. These non-verbal motions are used during the audio-visual task condition and emotional-expression real-time decoding.

Speech-related animation evaluation

To evaluate the perceptual accuracy of the decoded avatar animations, we used a crowdsourcing platform (Amazon Mechanical Turk) to design and conduct a perceptual assessment of the animations. Each decoded animation was assessed by 6 unique evaluators. Each evaluation consisted of playback of the decoded animation (with no audio) and textual presentation of the target (ground-truth) sentence and a randomly chosen other sentence from the same sentence set. Evaluators were instructed to identify the phrase that they thought the avatar was trying to say. We computed the median accuracy of the evaluations across evaluators for each sentence and treated that as the accuracy for a given trial and then computed the final accuracy distribution using the pseudo-block strategy described above.

Separately, we used the dlib software package Davis E. King 2009 to extract 72 facial keypoints for each frame in avatar-rendered and healthy-speaker videos (sampled at 30 frames per second). To obtain videos of healthy speakers, we recorded video and audio of 8 volunteers as they produced the same sentences used during real-time testing in the audio-visual task condition. We normalized the keypoint positions relative to other keypoints to account for head movements and rotation: We computed jaw movement as the distance between the keypoint at the bottom of the jaw and the nose, lip aperture as the distance between the keypoints at the top and bottom of the lips, and mouth width as the distance between the keypoints at either corner of the mouth. To compare avatar keypoint movements to those for healthy speakers, and to compare amongst healthy speakers, we first applied dynamic time warping to the movement time series and then computed the Pearson's correlation between the pair of warped time series. We held out 10 of 200 1024-word-General avatar videos from final evaluation since they were used to select parameters to automatically trim the dlib traces to speech onset and offset. We did this because our automatic segmentation method relied on the acoustic onset and offset, which is absent from direct avatar decoding videos.

Articulatory-movement decoding

To collect training data for non-verbal orofacial-movement decoding, we used the articulatory-movement task. Prior to data collection, the participant viewed a video of an avatar performing the following 6 movements: open mouth, pucker lips, lips back (smiling or lip retraction), raise tongue, lower tongue, and close mouth (rest or idle). Then, the participant performed the prompted-target task containing these movements as targets (presented as text). We instructed the participant to smoothly transition from neutral to the peak of the movement and then back to neutral, all within approximately 2 seconds starting at the go cue.

To train and test the avatar-movement classifier, we used a window of neural activity spanning from 1 second before to 3 seconds after the go cue for each trial. We first downsampled the ECoG features (high-gamma activity and low-frequency signals) by a factor of 6 to 33.33 Hz. We then normalized these features to have an L2-norm of 1 at each time point across all channels separately for the low-frequency and high-gamma signals. Next, we extracted the mean, minimum, maximum, and standard deviation across the first and second halves of the neural time window for each feature. These features were then stacked to form a 4048-dimensional neural-feature vector (the product of 256 electrodes, 2 feature sets, 4 statistics, and 2 data halves) for each trial. We then trained a multi-layer perceptron consisting of 2 linear layers with 512 hidden units and ReLU activations between the first and second layers. The final layer projected the output into a 6-dimensional output vector. We then applied a softmax activation to get a probability for each of the 6 different gestures. We evaluated the network using 10-fold cross-validation.

Emotional-expression decoding

To collect training data for non-verbal emotional-expression decoding, we used the emotional-expression task. Using the prompted-target task paradigm, we collected neural data as the participant attempted to produce three emotions (sad, happy, and surprised) at three intensity levels (high, medium, and low) for a total of 9 unique expressions. The participant chose her three base emotional expressions from a list of 30 options per emotion, and the animations corresponding to the three intensity levels were generated from these chosen base expressions. We instructed the participant to smoothly transition from neutral to the peak of the expression and then back to neutral, all within approximately 2 seconds starting at the go cue. We used the same data-windowing and neural-processing steps as for the articulatory-movement decoding. We used the same model architecture and training procedure as for the NATO-and-hand-motor classifier and our previous work 3. We initialized the expression classifier with a pre-trained NATO-and-hand-motor classifier (trained on 1,222 trials of NATO-motor task data collected prior to the start of collection for the emotional-expression task) and fine-tuned the weights on neural data from the emotional-expression task.

We evaluated the expression classifier using 15-fold cross-validation. Within the training set of each cross-validation fold, we fit 10 unique models to ensemble predictions on the held-out test set. We applied hierarchical agglomerative clustering to the 9-way confusion matrix in Figure 3.4e using SciPy Virtanen et al. 2020.

Articulatory-encoding assessments

To investigate the neural representations driving speech decoding, we assessed the selectivity of each electrode to articulatory groups of phones. Specifically, we fit a linear receptive-field encoding model to predict each electrode's high gamma activity (HGA) from phone-emission probabilities predicted by the text-decoding model during 10-fold cross-validation with data recorded with the 1024-word-General sentence set. We first decimated the HGA by a factor of 24, from 200 Hz to 8.33 Hz, to match the sampling rate of the phone-emission probabilities. Then, we fit a linear receptive-field model to predict the HGA at each electrode, using the phone-emission probabilities as time-lagged input features (39 phones and 1 aggregate token representing both the silence and CTC-blank tokens). We used a +/- 4-sample (480-ms) receptive-field window, allowing for slight misalignment between the text decoder's bidirectional-RNN phone-emission probabilities and the underlying HGA. We fit an independent model for each electrode. The true HGA, HGA(t), is modeled as a weighted linear combination of phone-emission probabilities (indexed by p) in the overall emissions matrix (X) over a +/- 4-sample window around each time point. This resulted in a learned weight matrix w(d,p) in which each phone, p, has temporal coefficients d1. . . D, where d1 is -4 and D is 4. During training the squared error between the predicted HGA, HGA*(t), and the true HGA, HGA(t), is minimized, Using the following formulas:

$$HGA^{*}(t) = \sum_{d=1}^{D} \sum_{p=1}^{P} w(d,p) * X(p, t-d)$$

$$\min \sum_{t} [HGA * (t) - HGA(t)]^{2}$$

We implemented the model with the MNE toolbox's receptive-field ridge regression in Python (Gramfort et al. 2013). We used cross-validation to select the alpha ridge-regression parameter by sweeping over the values [1e-1, 1e0, 1e1, ... 1e5], using 10% of our total data as a held-out tuning set. We used the remaining 90% of our total data for 10-fold cross-validation with the alpha parameter found optimal on the tuning set. We averaged the coefficients for the model across the 10 folds and collapsed across time samples for every phone using the maximum magnitude weight. The sign of the weight could be positive or negative. This yielded a single vector for each electrode, where each element in each vector was the maximum encoding of a given phone. Next, we pruned any electrode channels that were not significantly modulated by silent speech attempts. For each electrode, we computed the mean HGA magnitudes in the 1 s intervals immediately before and after the go cue for each NATO code-word trial in the NATO-motor task. If an electrode did not have significantly increased HGA after the go cue compared to before, it was excluded from the remainder of this analysis (significant modulation determined using one-sided Wilcoxon signed-rank tests with an alpha level of 0.00001 after applying 253-way Holm-Bonferroni correction). We then applied a second pruning step to exclude any electrodes that had encoding values (r) less than or equal to 0.2. We applied the centroid clustering method, a hierarchical, agglomerative clustering technique, to the encoding vectors using the SciPy Python package (Virtanen et al. 2020). We performed clustering along both the electrode and phone dimensions.

To assess any relationships between phone encodings and articulatory features, we assigned each phone to a place-of-articulation (POA) feature category, similar to previous work (Chartier et al. 2018; Bouchard et al. 2013). Specifically, each phone was either primarily articulated at the lips (labial), the front tongue, the back tongue, or larynx (vocalic). To quantify whether the unsupervised phone-encoding clusters reflected grouping by POA, we tested the null hypothesis that the observed parcellation of phones into clusters was not more organized by POA category than by chance. To test this null hypothesis, we used the following steps: 1) Compute the POA linkage distances by clustering the phones by Euclidean distance into F clusters, where F=4 is the number of POA categories; 2) Randomly shuffle the mapping between the phone labels and the phonetic encodings; 3) For each POA category, compute the maximum number of phones within that category that appear within a single cluster; 4) Repeat steps 2 and 3 over a total of 10,000 bootstrap runs; 5) Compute the pairwise Euclidean distance between all combinations of the 10,000 bootstrap results; 6) Repeat step 3 using the true unsupervised phone ordering and clustering; 7) Compute the pairwise Euclidean distances between the result from step 6 and each bootstrap from step 4; 8) Compute the one-tailed Wilcoxon rank-sum test between the results from step 7 and step 5. The resulting P-value is the probability of the aforementioned null hypothesis.

To visualize population-level (across all electrodes that were not pruned from the analysis) encoding of POA features, we first computed the mean encoding of each electrode across the 4 POA feature groups (vocalic, front tongue, labial, and back tongue). We then z-scored the mean encodings for each POA feature and then applied multidimensional scaling (MDS) over the electrodes to visualize each phone in a 2-dimensional space. We implemented this using the scikit-learn Python package (Pedregosa et al. 2011).

To measure somatotopy, we computed kernel density estimations of the locations of top electrodes (the 30% of electrodes with the strongest encoding weights) for each POA category along anterior-posterior and dorsal-ventral axes (Figure 3.2f). To do this, we used the seaborn Python package (Waskom 2021), Gaussian kernels, and Scott's Rule.

To quantify the magnitude of activation in response to non-verbal orofacial movements, we took the median of the evoked response potential to each action over the time window spanning from 1 s before to 2 s after the go cue. From this, we subtracted the same metric computed across all actions to account for electrodes that were non-differentially task-activated. For each action, we then normalized values across electrodes to be between 0 and 1. We used ordinary least squares linear regression, implemented by the statsmodels Python package (Seabold and Perktold 2010), to relate phone-encoding weights with activation to attempted motor movements.

To assess whether postcentral responses largely reflected sensory feedback, we compared the time to activation between precentral and postcentral electrodes. For each speech responsive electrode (see above), we averaged the HGA across trials (ERP) of each of the 26 NATO code-words. For each electrode, we found the time at which each code-word ERP reached its peak. Given that electrodes may have strong preferences for groups of phones (Figure 3.5), we took the minimum time-to-peak across code-word ERPs for further analysis. For each electrode's optimal code-word ERP, we also calculated the time-to-onset, defined as the earliest time point at which the HGA was statistically significantly greater than 0. We measured this with Wilcoxon Rank-sum tests at a significance level of 0.05, similar to prior work (Cheung et al. 2016).

Exclusion analyses

We assigned each electrode to an anatomical region and visualized all electrodes on the pial surface using the same methods described in our previous work 77. For the exclusion analyses, we tested the phone-based text-decoding model on the real-time evaluation trials in the text task condition with the 1024-word-General sentence set. We did not use early stopping for these analyses; we used the full 8-s time windows of neural activity for each trial. For the synthesis and avatar direct-decoding models, we tested on the real-time synthesis evaluation trials from the 1024-word-General set, and evaluation remained consistent with other analyses. Also, we tested the NATO code-word classifier by training and testing on NATO code-word trials recorded during the NATO-motor task. We used all of the NATO-motor task blocks recorded after freezing the classifier (Figure 3.2h), a total of 19 blocks, as the test set.

Electrode contributions

For text, synthesis, and direct avatar decoding models, we measured the contribution of each electrode to the model's predictions. We computed the derivative of each model's loss function with respect to the HGA and LFS features of each electrode across time 32. We then computed the L1 norm of these values across time and averaged across all trials in the corresponding test set for the model. For each electrode, we then summed the resulting contribution for HGA and LFS to obtain one aggregate contribution. For each model, contributions were then normalized to fall between 0 and 1. To compare contributions across decoding modalities, we used ordinary least squares linear regression, implemented by the statsmodels Python package (Seabold and Perktold 2010).

Statistics

Statistical analyses

Statistical tests are fully described in the figure captions and text. To summarize, we used two-sided Mann-Whitney Wilcoxon Rank-sum tests to compare unpaired distributions. Critically, these tests do not assume normally distributed data. For paired comparisons, we used two-sided Wilcoxon signed-rank tests, which also do not assume normally distributed data. When the underlying neural data was not independent across comparisons, we used the Holm-Bonferroni correction for multiple comparisons. P-values < 0.01 were considered statistically significant. 99% confidence intervals were estimated using a bootstrapping approach where we randomly sampled the distribution (e.g. trials or pseudo-blocks) of interest with replacement 2000 or 1000 times and the desired metric was computed. The confidence interval was then computed on this distribution of the bootstrapped metric. P-values associated with the Pearson correlation were computed with a permutation test where data was randomly shuffled 1000 times. To compare success rates of decoding during our freeform demonstration with the main real-time evaluation, we used a t-test.

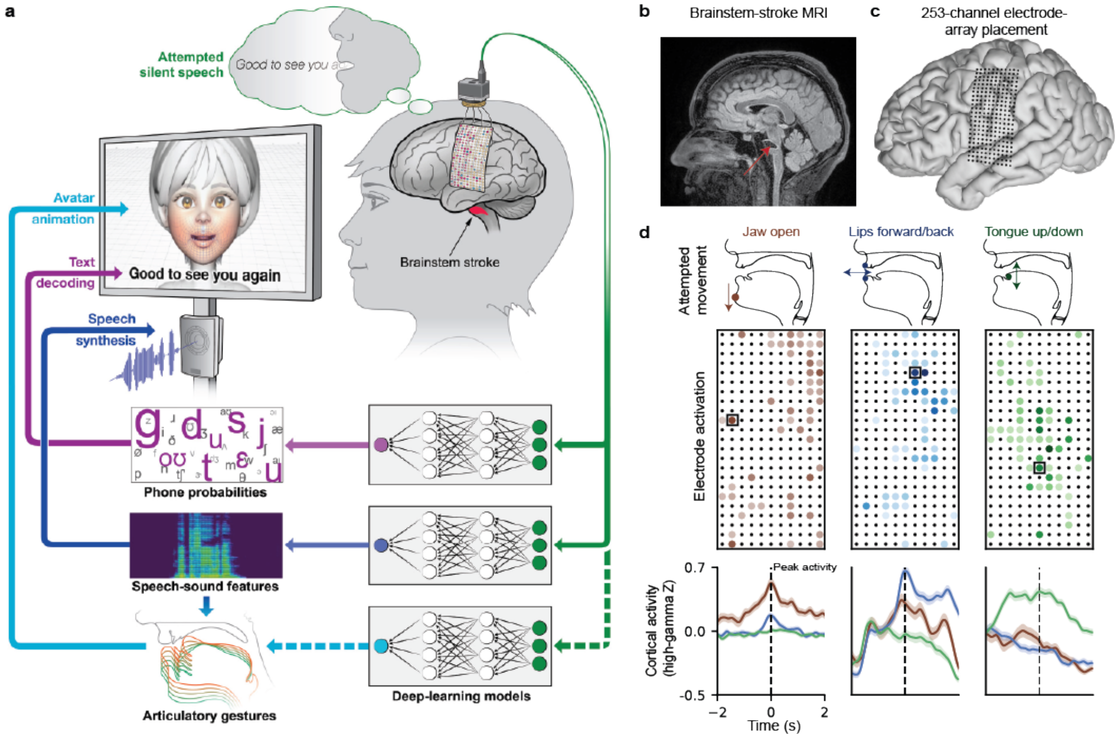


Figure 3.1. Multimodal speech decoding in a participant with vocal-tract paralysis (continued on next page).

(Previous page.) **Figure 3.1. Multimodal speech decoding in a participant with vocal-tract paralysis a** Overview of the speech-decoding pipeline. A brainstem-stroke survivor with anarthria was implanted with a 253-channel high-density electrocorticography (ECoG) array 18 years after injury. Neural activity was processed and used to train deep learning models to predict phone probabilities, speech-sound features, and articulatory gestures. These outputs were used to decode text, synthesize audible speech, and animate a virtual avatar, respectively. **b**, A sagittal MRI showing brainstem atrophy (in the bilateral pons; red arrow) resulting from stroke. **c**, MRI reconstruction of the participant's brain overlaid with the locations of implanted electrodes. The ECoG array was implanted over the participant's lateral cortex, centered on the central sulcus. **d** Top: Simple articulatory movements attempted by the participant. Middle: Electrode-activation maps demonstrating robust electrode tunings across articulators during attempted movements. Only the electrodes with the strongest responses (top 20%) are shown for each movement type. Color indicates the magnitude of the average evoked high-gamma activity (HGA) response with each type of movement. Bottom: Z-scored trial-averaged evoked HGA responses with each movement type for each of the boxed electrodes in the electrode-activation maps. In each plot, each response trace shows mean +/- standard error across trials and is aligned to the peak activation time.

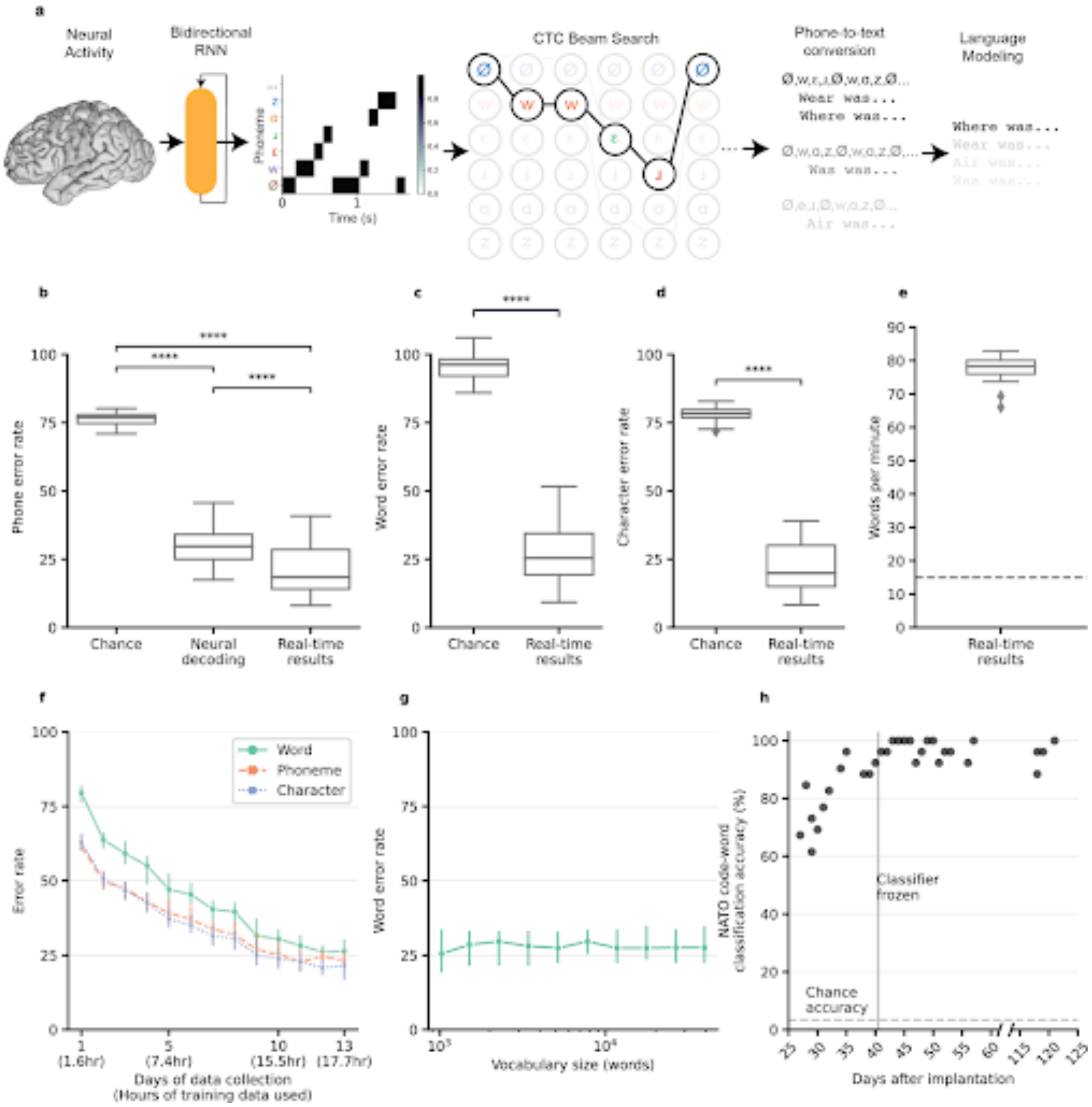


Figure 3.2. High-performance text decoding from neural activity a. During attempts to silently speak, a bidirectional recurrent neural network (RNN) decodes neural features into a time series of phone and silence (denoted as Ø) probabilities. A CTC beam search computes the most likely sequence of phones that can be translated into words in the vocabulary. An n-gram language model rescores sentences created from these sequences to yield the most likely sentence. **b** Median phone error rates, calculated using shuffled neural data (Chance), neural decoding without applying vocabulary constraints or language modeling (Neural decoding only), and the full real-time system (Real-time results) across n=25 pseudo-blocks. **c** Word error rates and d, character error rates for chance and real-time results. In **b–d**, ****P < 0.0001, Two-sided Wilcoxon Signed-Rank test with 5-way Holm-Bonferroni correction for multiple comparisons; P-values and statistics in Extended Data Table 1. **e**, Decoded words per minute. Dashed line denotes previous state-of-the-art speech BCI decoding rate in a person with paralysis David A. Moses et al. 2021 **f**, Offline evaluation of **g**. error rates as a function of data quantity and h. word error rate as a function of the number of words used to apply vocabulary constraints and train the language model. **f-g** Error bars represent 99% CIs of the median, calculated using 1000 bootstraps across n=125 pseudo-blocks **f** and n=25 pseudo-blocks **g** at each point. **h**. Real-time classification accuracy during attempts to silently say 26 NATO code words across many recording days. The vertical line represents when the classifier was no longer re-trained before each session. In **b–g**, results were computed using the real-time evaluation trials with the 1024-word-General sentence set. Box plots in all figures depict median (horizontal line inside box), 25th and 75th percentiles (box), 25th and 75th percentiles +/- 1.5 times the interquartile range (whiskers), and outliers (diamonds).

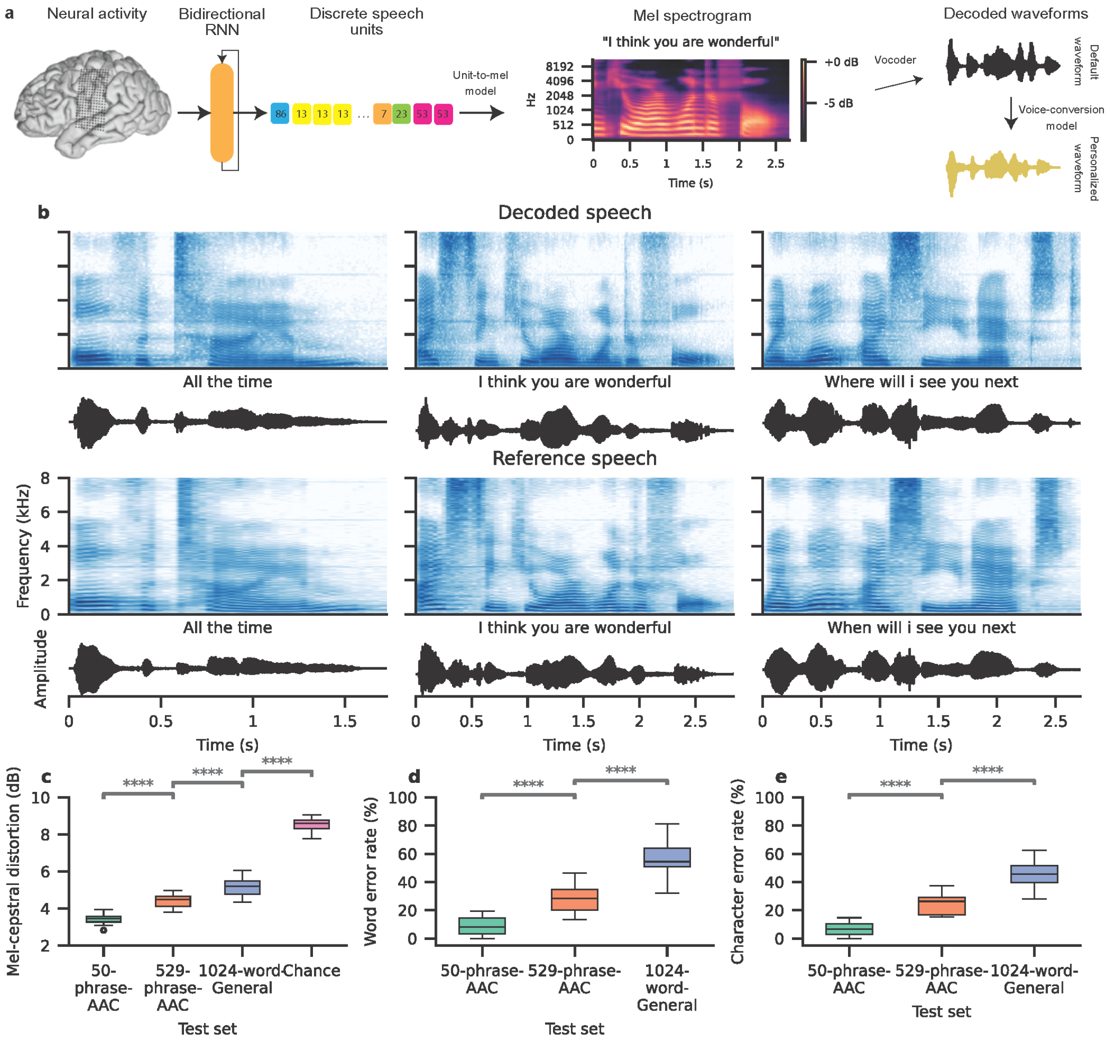


Figure 3.3. Intelligible speech synthesis from neural activity. (continued on next page).

(Previous page.) **Figure 3.3. Intelligible speech synthesis from neural activity. a** Schematic diagram of the speech-synthesis decoding algorithm. During attempts to silently speak, a bidirectional recurrent neural network (RNN) decodes neural features into a time series of discrete speech units. The RNN was trained using reference speech units computed by applying a large pretrained acoustic model (HuBERT) on basis waveforms. Predicted speech units are then transformed into the mel spectrogram and vocoded into audible speech. The decoded waveform is played back to the participant in real time after a brief delay. Offline, the decoded speech was transformed to be in the participant's personalized synthetic voice using a voice-conversion model. **b** Top: Three example decoded spectrograms and waveforms (top) from the 529-phrase-AAC sentence set. Bottom: The corresponding reference spectrograms and waveforms representing the decoding targets. **c** Mel-cepstral distortions (MCDs) for the decoded waveforms. Lower MCD indicates better performance. Chance waveforms were computed by shuffling electrode indices in the test data for the 50-phrase-AAC set with the same synthesis pipeline. **d** Perceptual word error rates from untrained human evaluators via a transcription task. **e** Perceptual character error rates from the same human-evaluation results as d. In **c**–**e**, ****P < 0.0001, Mann-Whitney U-test with 19-way Holm-Bonferroni correction for multiple comparisons; all non-adjacent comparisons were also significant (P < 0.0001; not depicted); n=15 pseudo-blocks for the AAC sets, n=20 pseudo-blocks for 1024-word-General set. P-values and statistics in Extended Data Table 2. In **b**–**e**, all decoded waveforms, spectrograms, and quantitative results use the non-personalized voice .

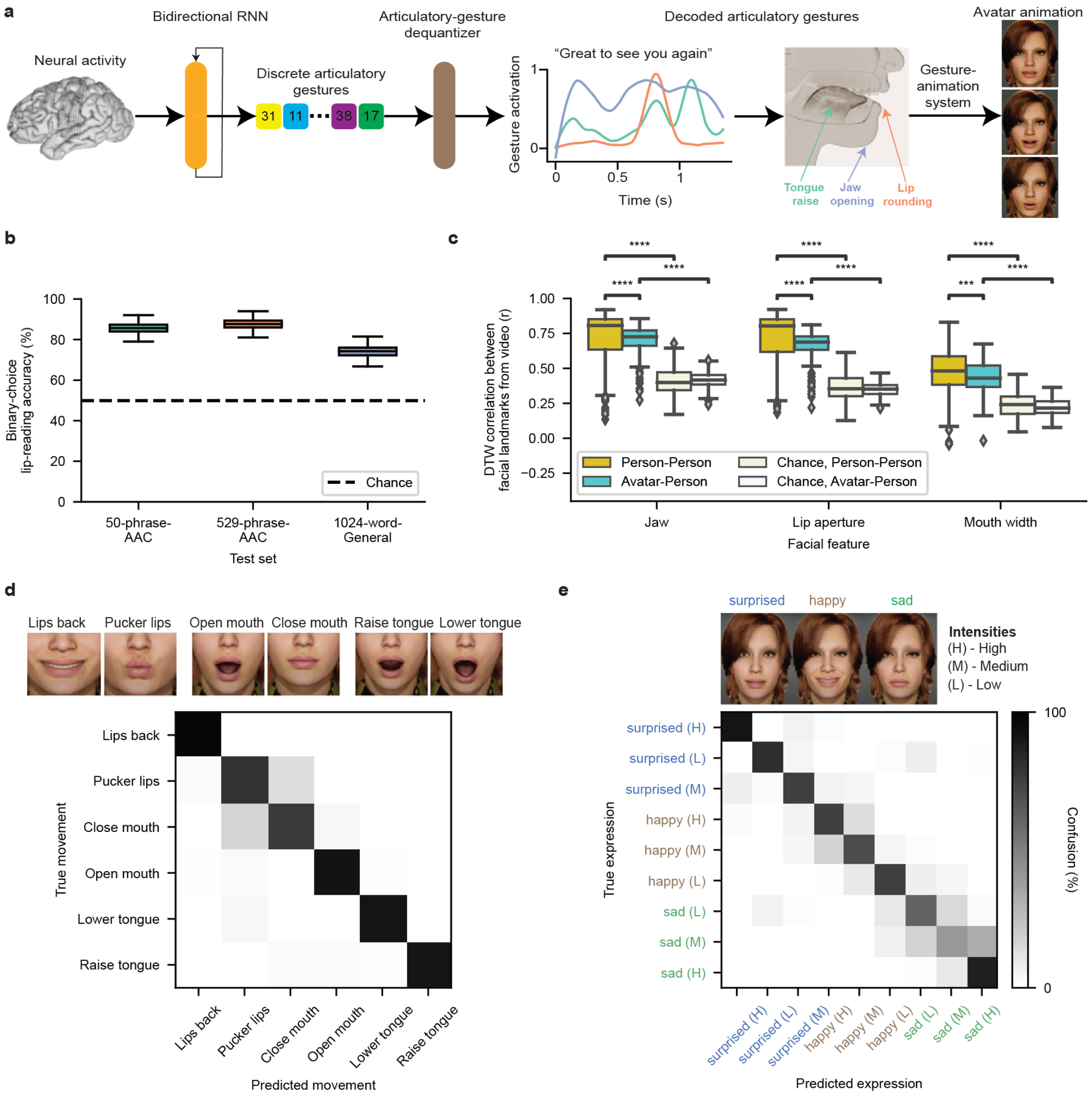


Figure 3.4. Direct decoding of orofacial articulatory gestures from neural activity to drive an avatar (continued on next page).

(Previous page.) **Figure 3.4. Direct decoding of orofacial articulatory gestures from neural activity to drive an avatar a** Schematic diagram of the avatar decoding algorithm. Offline, a bidirectional recurrent neural network (RNN) decodes neural activity recorded during attempts to silently speak into discretized articulatory gestures (quantized via a vector quantised variational autoencoder, abbreviated VQ-VAE). A convolutional neural network de-quantizer (VQ-VAE decoder) is then applied to generate the final predicted gestures, which are then passed through a pre-trained gesture-animation model to animate the avatar in a virtual environment. **b** Binary perceptual accuracies from human evaluators on avatar animations generated from neural activity, n=2000 bootstrapped points. **c** Correlations for jaw, lip, and mouth-width movements between decoded avatar renderings and videos of real human speakers on the 1024-word-General sentence set across all pseudo-blocks for each comparison (n=152 for avatar-person comparison, n=532 for person-person comparisons; ****P < 0.0001, Mann-Whitney U-test with 9-way Holm-Bonferroni correction; p-values and U-statistics in Table S3). A facial-landmark detector (dlib) was used to measure orofacial movements from the videos. **d** Top: Snapshots of avatar animations of 6 non-speech articulatory movements in the articulatory-movement task. Bottom: Confusion matrix depicting classification accuracy across the movements. The classifier was trained to predict which movement the participant was attempting from her neural activity, and the prediction was used to animate the avatar. **e** Top: Snapshots of avatar animations of 3 non-speech emotional expressions in the emotional-expression task. Bottom: Confusion matrix depicting classification accuracy across 3 intensity levels (high, medium, and low) of the 3 expressions, ordered via hierarchical agglomerative clustering on the confusion values. The classifier was trained to predict which expression the participant was attempting from her neural activity, and the prediction was used to animate the avatar.

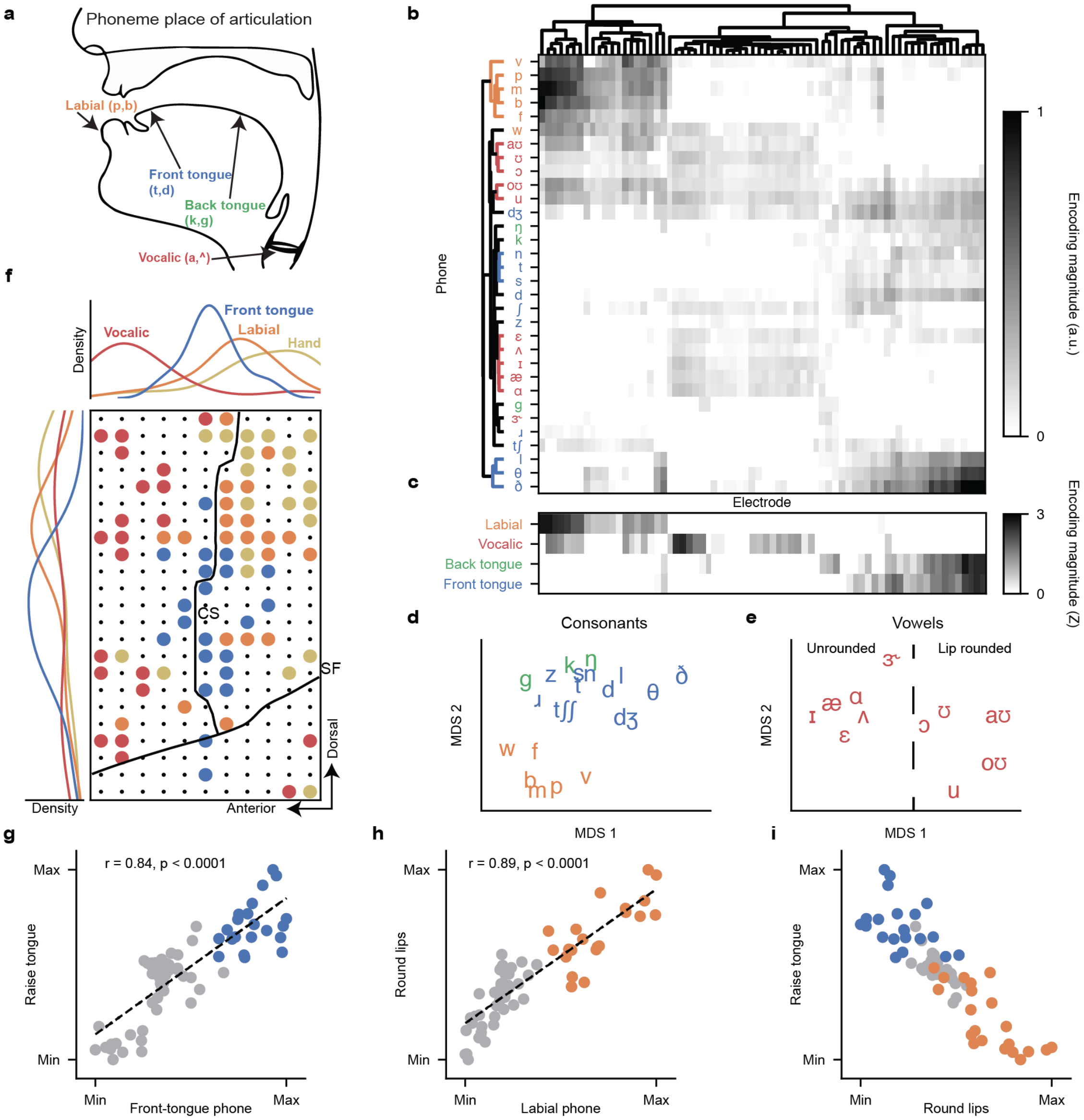


Figure 3.5. Articulatory encodings driving speech decoding (continued on next page).

(Previous page.) **Figure 3.5. Articulatory encodings driving speech decoding** a, Mid-sagittal schematic of the vocal tract with phone place of articulation (POA) features labeled. b, Phone-encoding vectors for each electrode computed by a temporal receptive-field model on neural activity recorded during attempts to silently say sentences from the 1024-word-General set, organized by unsupervised hierarchical clustering. c, Z-scored POA encodings for each electrode, computed by averaging across positive phone encodings within each POA category. Z values are clipped at 0. d,e, Projection of consonant (d) and vowel (e) phone encodings into a two-dimensional space via multidimensional scaling (MDS). f, Bottom-right: Visualization of the locations of electrodes with the greatest encoding weights for labial, front-tongue, and vocalic phones on the electrocorticography array. The electrodes that most strongly encoded finger flexion during the NATO-motor task are also included. Only the top 30% of electrodes within each condition are shown, and the strongest tuning was used for categorization if an electrode was in the top 30% for multiple conditions. Black lines denote the central sulcus (CS) and sylvian fissure (SF). Top and bottom-left: The spatial electrode distributions for each condition along the anterior-posterior and ventral-dorsal axes, respectively. g–i, Electrode-tuning comparisons between front-tongue phone encoding and tongue-raising attempts (g; r=0.84, $P < 0.0001$, ordinary least squares regression), labial phone encoding and lip-puckering attempts (h; r=0.89, $P < 0.0001$, ordinary least squares regression), and tongue-raising and lip-rounding attempts (i). Non-phonetic tunings were computed from neural activations during the articulatory-movement task. Each plot depicts the same electrodes encoding front-tongue and labial phones (from f) as blue and orange dots, respectively; all other electrodes are shown as gray dots.

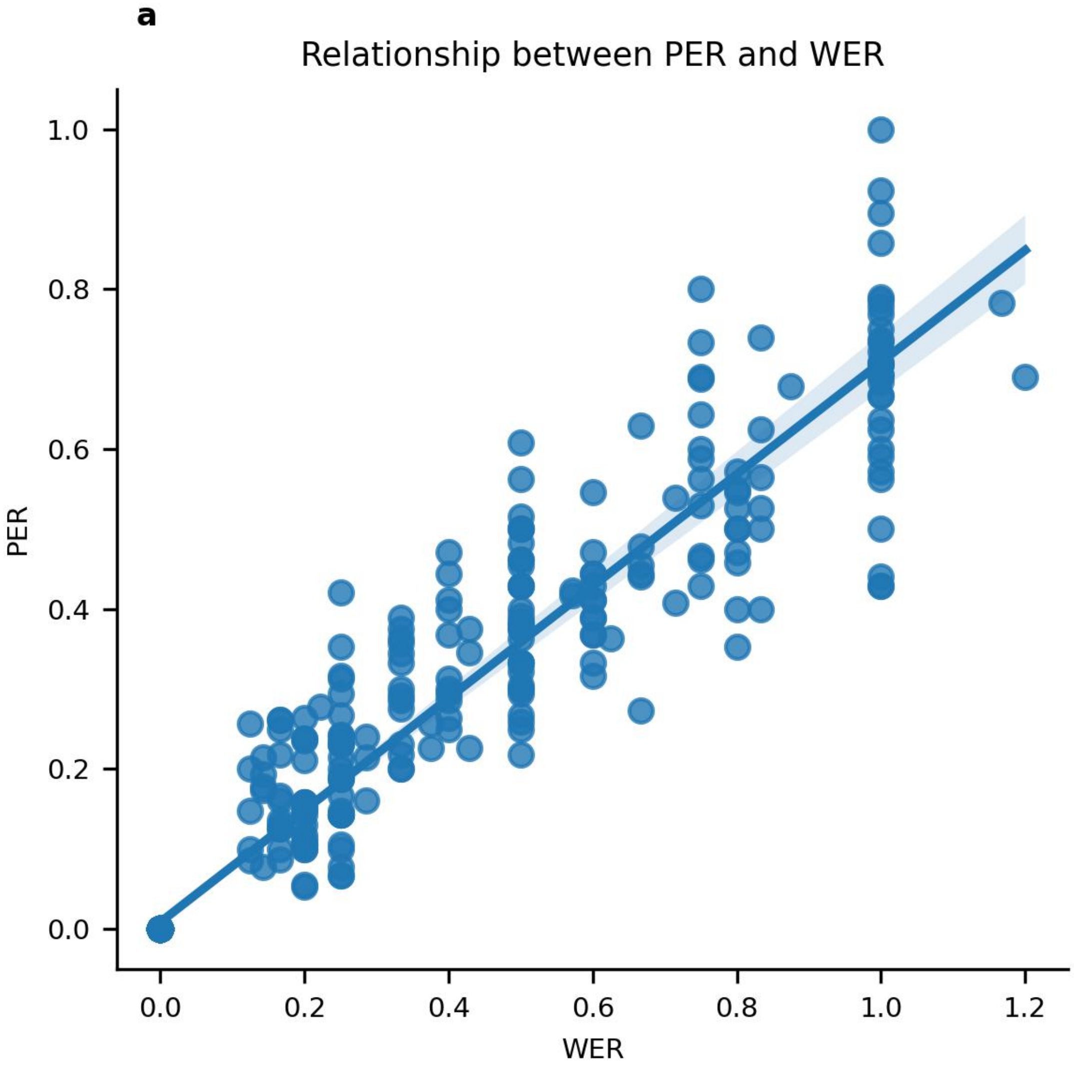


Figure 3.6. Word vs phone error rate Relationship between phone error rate and word error rate across n =549 points. Each point represents the phone and word error rate for all sentences used during model evaluation for all evaluation sets. The points display a linear trend, with the linear equation corresponding with an R^2 of .925. Shading denotes 99% confidence interval which was calculated via bootstrapping over 2000 iterations.

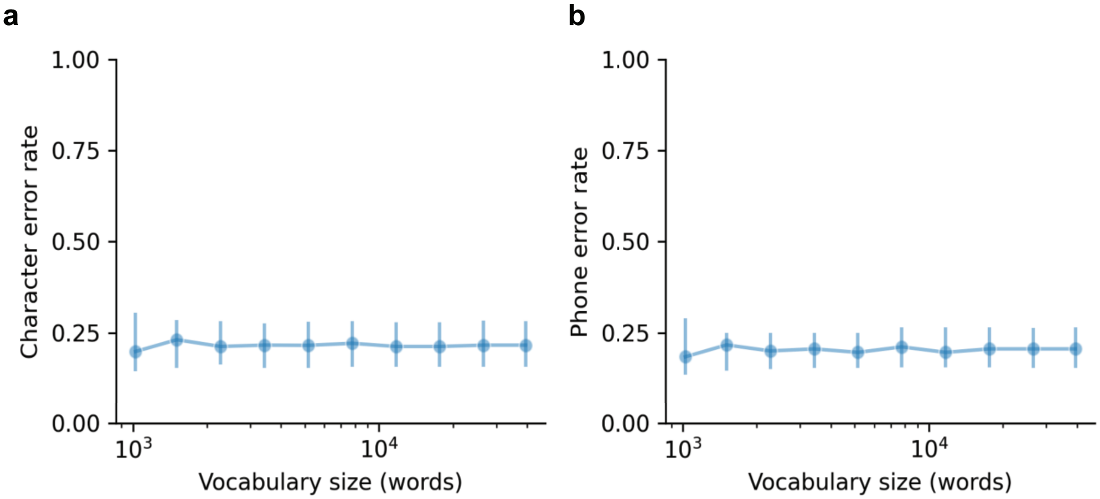


Figure 3.7. Phone and character error rates for simulated text decoding with larger vocabularies We simulated text decoding results with using log-spaced vocabularies of 1,506, 2,269, 3,419, 5,152, 7,763, 11,696, 17,621, 26,549, and 39,378 words, and compared performance to the real-time results using our 1,024 word vocabulary. Each point represents the median **a** character or **b** phone error rate across n=25 real-time evaluation pseudo-blocks, and error bars represent 99% confidence intervals of the median. With our largest 39,378 word vocabulary, we found a median character error rate of 21.7% (99% CI [16.3%, 28.1%]), and median phone error rate of 20.6% (99% CI [15.9%, 26.1%]). We compared the WER, CER, and PER of the simulation with the largest vocabulary size to the real-time results, and found that there was no significant increase in any error rate (P > .01 for all comparisons. Test statistic=48.5, 93.0, 88.0, respectively, p=.342, .342, .239, respectively, Wilcoxon signed-rank test with 3-way Holm-Bonferroni correction).

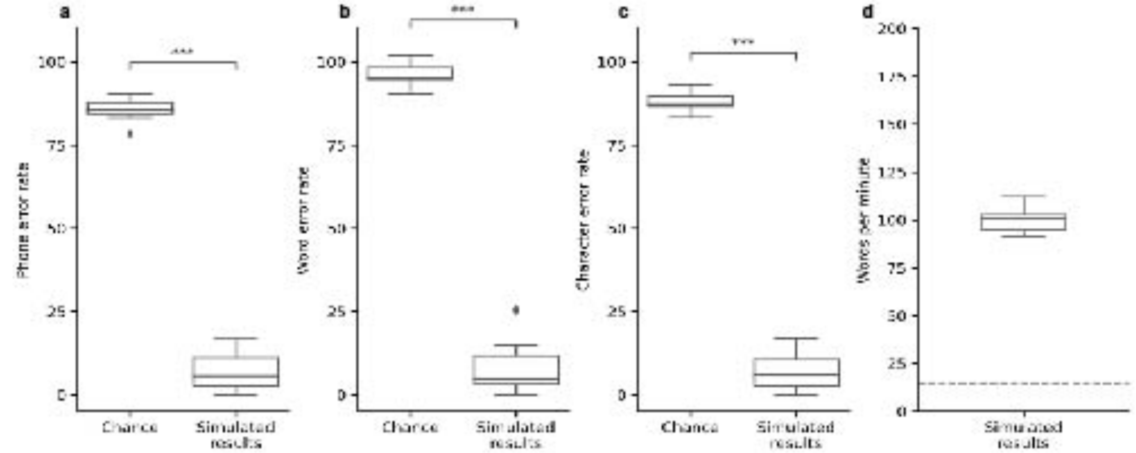


Figure 3.8. Simulated results on the 50-phrase-AAC sentence set We simulated text decoding results on the real-time blocks used for evaluation with the synthesis models. On the 50-phrase-AAC sentence set, we achieved extremely high accuracy. **a** Across $n = 15$ pseudo-blocks, we observed a median PER of 5.63% (99% CI [2.10, 12.0]). **b** Median WER was 4.92% (99% CI [3.18, 14.0]), and **c** median CER was 5.91% (99% CI [2.21, 11.4]). Speech was decoded at high rates with a median WPM of 101 (99% CI [95.6, 103]). The PER, WER, and CER were also significantly better than chance ($P < .001$ for all metrics, Wilcoxon signed-rank test with 3-way Holm-Bonferonni Correction for multiple comparisons). Statistics compare $n = 15$ total pseudo-blocks. For PER: stat=0, P = 1.83e-4. For CER: stat = 0, P=1.83e-4. For WER: stat = 0, P=1.83e-4.

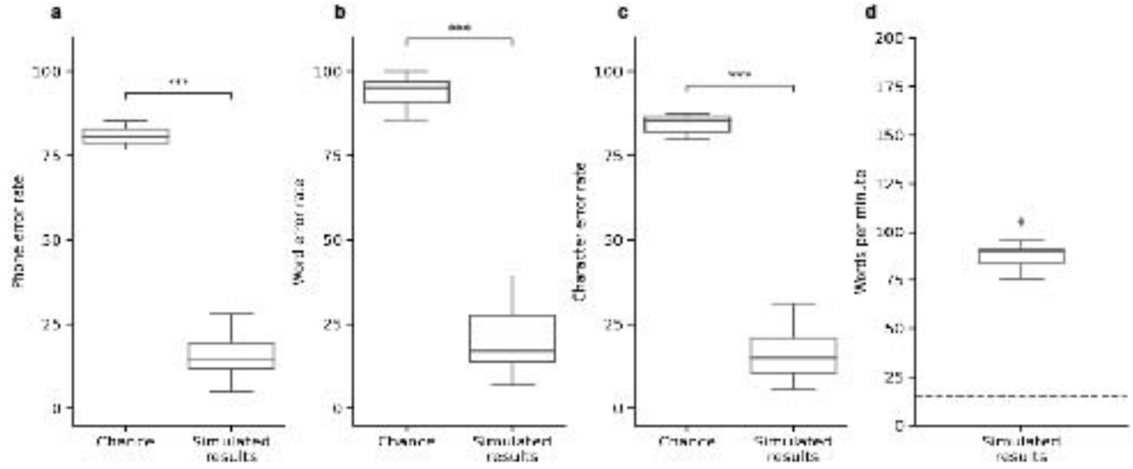


Figure 3.9. Simulated text decoding results on the 529-phrase-AAC sentence set We simulated text decoding results on the real-time blocks used for evaluation of with the synthesis models. We observed a median PER of 17.3 (99% CI [12.6, 20.1]). Median WER was 17.1% (99% CI [8.89, 28.9]), and median CER was 15.2% (99% CI [10.1, 22.7]). Speech was decoded at high rates with a median WPM of 89.9 (99% CI [83.6, 93.3]). The PER, WER, and CER were also significantly better than chance ($P < .001$ for all metrics, two-sided Wilcoxon signed-rank test with 3-way Holm-Bonferonni Correction for multiple comparisons). Statistics compare $n = 15$ total pseudo-blocks. For PER: stat=0, P = 1.83e-4. For CER: stat = 0, P=1.83e-4. For WER: stat = 0, P=1.83e-4.

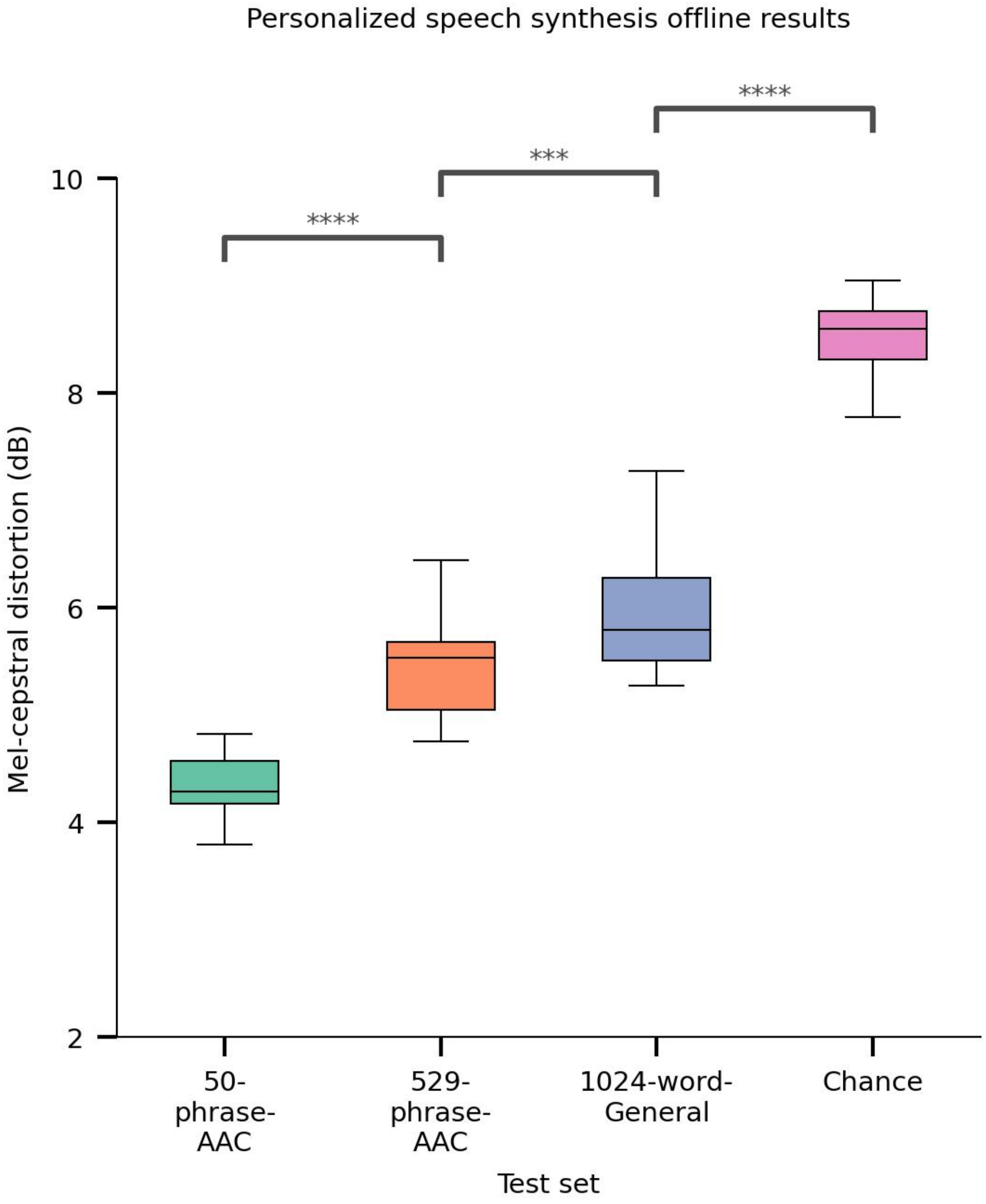


Figure 3.10. Mel-cepstral distortions (MCDs) using a personalized voice tailored to the participant We calculate the Mel-cepstral distortion (MCDs) between decoded speech with the participant's personalized voice and voice-converted reference waveforms for the , , and set. Lower MCD indicates better performance. We achieved mean MCDs of 3.87 (99% CI [3.83, 4.45]), 5.12 (99% CI [4.41, 5.35]), and 5.57 (99% CI [5.17, 5.90]) dB for the (N = 15 pseudo-blocks), (N = 15 pseudo-blocks), and sets (N = 20 pseudo-blocks) Chance MCDs were computed by shuffling electrode indices in the test data with the same synthesis pipeline and computed on the evaluation set. The MCDs of all sets are significantly lower than the chance. vs. $*** = P < 0.001$, otherwise all $**** = P < 0.0001$. Two-sided Wilcoxon rank-sum tests were used for comparisons within-dataset and Mann-Whitney U-test outside of dataset with 9-way Holm-Bonferroni correct.

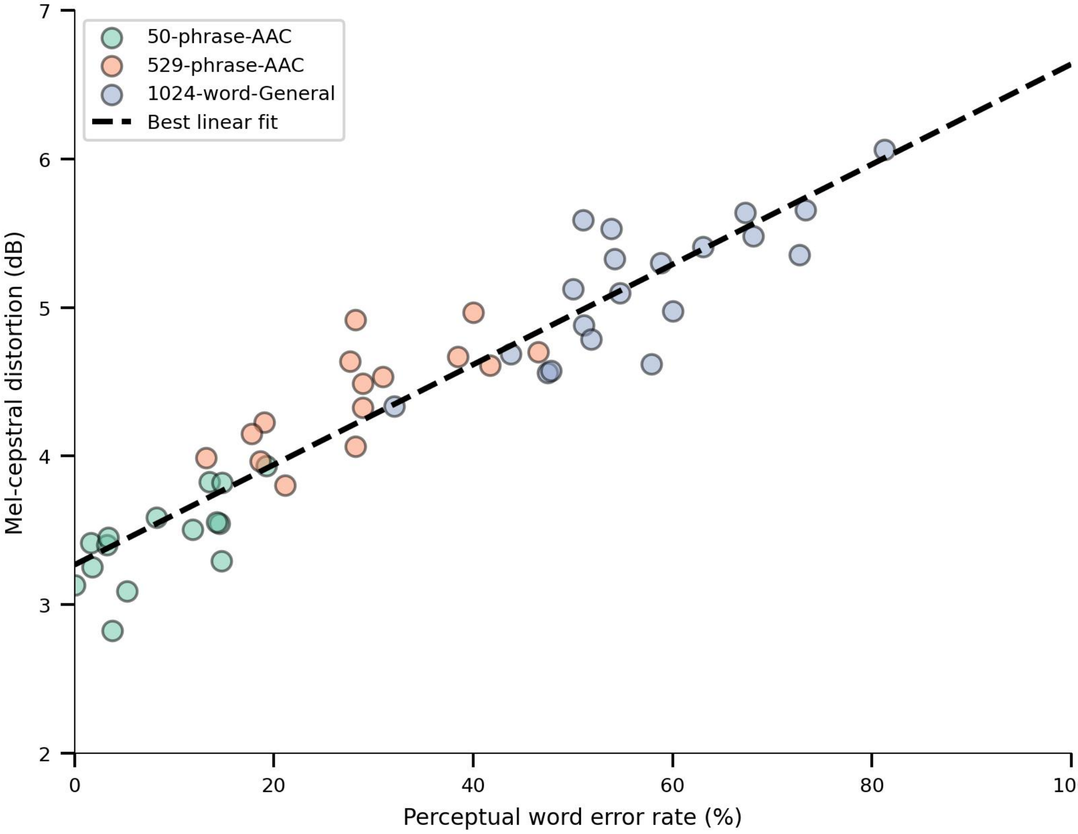


Figure 3.11. Comparison of perceptual word error rate and mel-cepstral distortion Scatter plot illustrating relationship between perceptual word error rate (WER) and mel-cepstral distortion (MCD) for , , . Each data point represents the mean accuracy from a single pseudo-block. A dashed black line indicates the best linear fit to the pseudo-blocks, providing a visual representation of the overall trend. Consistent with expectation, this plot suggests a positive correlation between WER and MCD for our speech-synthesizer.

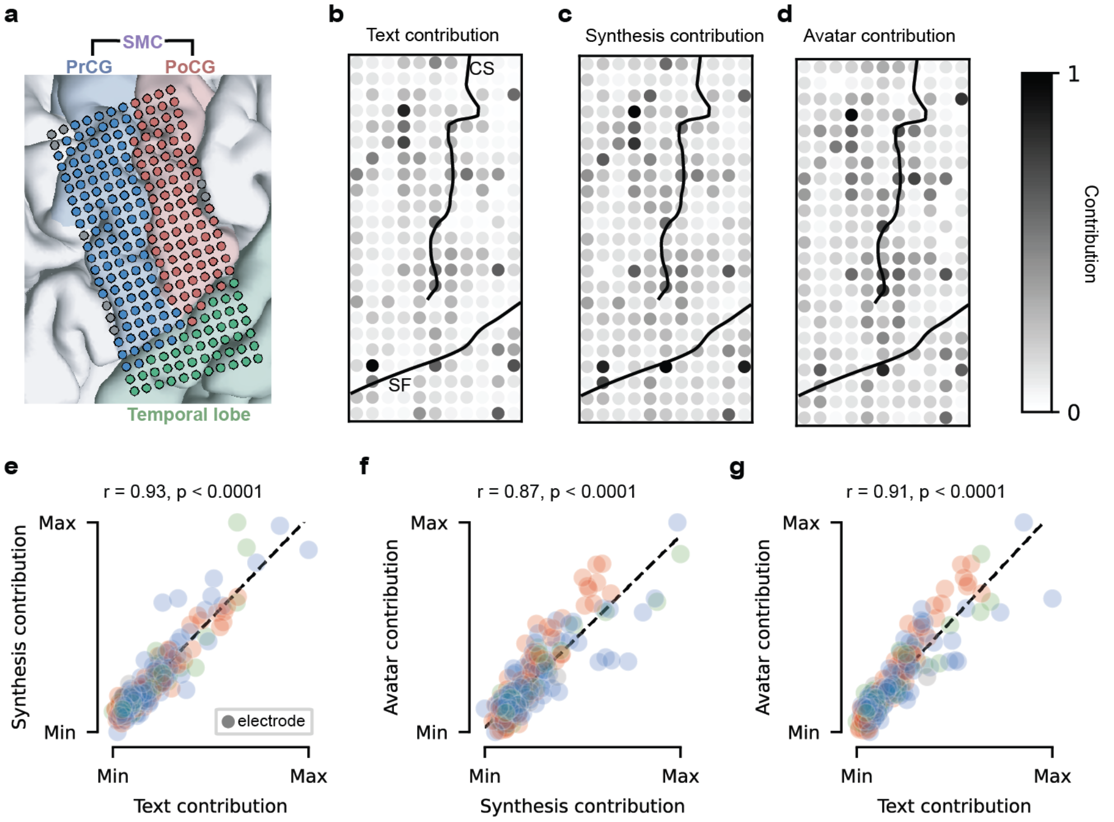


Figure 3.12. Electrode contributions to decoding performance MRI reconstruction of the participant's brain overlaid with the locations of implanted electrodes. Cortical regions and electrodes are colored according to anatomical region (PoCG: postcentral gyrus, PrCG: precentral gyrus, SMC: sensorimotor cortex). **b–d**, Electrode contributions to text decoding **(b)**, speech synthesis **(c)**, and avatar direct decoding **(d)**. Black lines denote the central sulcus (CS) and sylvian fissure (SF). **e-g**, Each plot shows each electrode's contributions to two modalities as well as the Pearson correlation across electrodes and associated p-value.

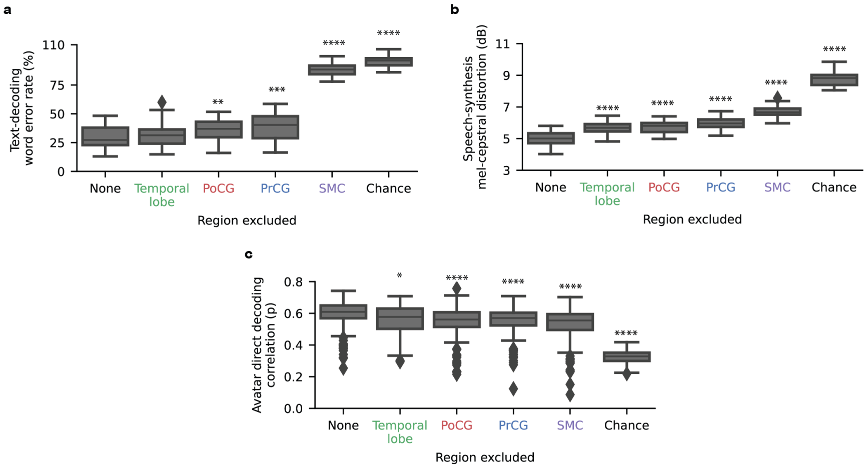


Figure 3.13. Effect of anatomical regions on decoding performance A. a–c, Effect of excluding each region during training and testing on text-decoding word error rate **(a)**, speech-synthesis mel-cepstral distortion **(b)**, and avatar direct-decoding correlation (**c**; average DTW correlation of jaw, lip, and mouth-width landmarks between the avatar and healthy speakers), computed using neural data as the participant attempted to silently say sentences from the 1024-word-General set. Significance markers indicate comparisons against the None condition, which uses all electrodes. *P < 0.01, **P<0.005, ***P<0.001, ****P < 0.0001, two-sided Wilcoxon signed-rank test with 15-way Holm-Bonferroni correction (full comparisons are given in Table S5). Distributions are over 25 pseudo-blocks for text decoding, 20 pseudo-blocks for speech synthesis, and 152 pseudo-blocks (19 pseudo-blocks each for 8 healthy speakers) for avatar direct 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.

Arce, Fritzie I., J.-C. Lee, Callum F. Ross, et al. (June 2013). "Directional information from neuronal ensembles in the primate orofacial sensorimotor cortex". en. *American Journal of Physiology-Heart and Circulatory Physiology*. Publisher: The American Physiological Society. DOI: 10.1152/jn.00144.2013.

Berger, Michael A., Gregor Hofer, and Hiroshi Shimodaira (Sept. 2011). "Carnival—Combining Speech Technology and Computer Animation". *IEEE Computer Graphics and Applications* 31.5. Conference Name: IEEE Computer Graphics and Applications, pp. 80–89. ISSN: 1558-1756. DOI: 10.1109/MCG.2011.71.

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.

Bird, Steven, Ewan Klein, and Edward Loper (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

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.

Breshears, Jonathan D., Annette M. Molinaro, and Edward F. Chang (Aug. 2015). "A probabilistic map of the human ventral sensorimotor cortex using electrical stimulation". eng. *Journal of Neurosurgery* 123.2, pp. 340–349. ISSN: 1933-0693. DOI: 10.3171/2014.11.JNS14889.

Bruurmijn, Mark L C M, Isabelle P L Pereboom, Mariska J Vansteensel, et al. (Dec. 2017). "Preservation of hand movement representation in the sensorimotor areas of amputees". en. *Brain* 140.12, pp. 3166–3178. ISSN: 0006-8950, 1460-2156. DOI: 10.1093/brain/awx274.

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.

Casanova, Edresson, Julian Weber, Christopher Shulby, et al. (Feb. 2022). *YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone*. arXiv:2112.02418 [cs, eess].

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 (2016). "The auditory representation of speech sounds in human motor cortex". *elife* 5, e12577.

Cho, Cheol Jun, Peter Wu, Abdelrahman Mohamed, and Gopala K. Anumanchipalli (Oct. 2022). *Evidence of Vocal Tract Articulation in Self-Supervised Learning of Speech.* arXiv:2210.11723 [cs, eess]. DOI: 10.48550/arXiv.2210.11723.

Collobert, Ronan, Christian Puhrsch, and Gabriel Synnaeve (Sept. 2016). "Wav2Letter: an End-to-End ConvNet-based Speech Recognition System". *arXiv:1609.03193 [cs]*. arXiv: 1609.03193.

Danescu-Niculescu-Mizil, Cristian and Lillian Lee (June 2011). *Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs.* arXiv:1106.3077 [physics]. DOI: 10.48550/arXiv.1106.3077.

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

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.

Gramfort, Alexandre, Martin Luessi, Eric Larson, et al. (2013). “MEG and EEG data analysis with MNE-Python”. *Frontiers in Neuroscience* 7. ISSN: 1662-453X.

Graves, Alex, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber (2006). “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks”. en. *Proceedings of the 23rd international conference on Machine learning - ICML ’06*. Pittsburgh, Pennsylvania: ACM Press, pp. 369–376. ISBN: 978-1-59593-383-6. DOI: `10.1145/1143844.1143891`.

Heafield, Kenneth (2011). “KenLM: Faster and Smaller Language Model Queries”. *Proceedings of the Sixth Workshop on Statistical Machine Translation*. WMT ’11. Association for Computational Linguistics, pp. 187–197. ISBN: 978-1-937284-12-1.

Hsu, Wei-Ning, Benjamin Bolte, Yao-Hung Hubert Tsai, et al. (June 2021a). *HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units*. arXiv:2106.07447 [cs, eess]. DOI: `10.48550/arXiv.2106.07447`.

— (June 2021b). *HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units*. arXiv:2106.07447 [cs, eess].

Huggins, Jane E., Patricia A. Wren, and Kirsten L. Gruis (Sept. 2011). “What would brain-computer interface users want? Opinions and priorities of potential users with amyotrophic lateral sclerosis”. *Amyotrophic Lateral Sclerosis* 12.5, pp. 318–324. ISSN: 1748-2968, 1471-180X. DOI: `10.3109/17482968.2011.572978`.

Ito, Keith and Linda Johnson (2017). *The LJ Speech Dataset*.

Jia, Jia, Xiaohui Wang, Zhiyong Wu, et al. (Dec. 2012). “Modeling the correlation between modality semantics and facial expressions”. *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1–10.

King, Davis E (n.d.). “Dlib-ml: A Machine Learning Toolkit”. en ().

— (2009). “Dlib-ml: A Machine Learning Toolkit”. *Journal of Machine Learning Research* 10, pp. 1755–1758.

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.

Kubichek, R. (1993). “Mel-cepstral distance measure for objective speech quality assessment”. *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing.* Vol. 1. Victoria, BC, Canada: IEEE, pp. 125–128. ISBN: 978-0-7803-0971-5. DOI: 10.1109/PACRIM.1993.407206.

Lakhotia, Kushal, Evgeny Kharitonov, Wei-Ning Hsu, et al. (2021). *Generative Spoken Language Modeling from Raw Audio.* arXiv: 2102.01192 [cs.CL].

Lee, Ann, Peng-Jen Chen, Changhan Wang, et al. (May 2022). “Direct Speech-to-Speech Translation With Discrete Units”. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Dublin, Ireland: Association for Computational Linguistics, pp. 3327–3339. DOI: 10.18653/v1/2022.acl-long.235.

Mehrabian, Albert (1981). *Silent messages: implicit communication of emotions and attitudes.* 2nd ed.

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, 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., 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.

Murray, Elisabeth A. and Joe Dan Coulter (1981). “Organization of corticospinal neurons in the monkey”. en. *Journal of Comparative Neurology* 195.2. _eprint: https://onlinelibrary.wiley.com/doi/p pp. 339–365. ISSN: 1096-9861. DOI: 10.1002/cne.901950212.

Oord, Aaron van den, Oriol Vinyals, and Koray Kavukcuoglu (2017). “Neural Discrete Representation Learning”. *Proceedings of the 31st International Conference on Neural In-*

formation Processing Systems. NIPS'17. event-place: Long Beach, California, USA. Red Hook, NY, USA: Curran Associates Inc., pp. 6309–6318. ISBN: 978-1-5108-6096-4.

Oord, Aaron van den, Sander Dieleman, Heiga Zen, et al. (Sept. 2016). "WaveNet: A Generative Model for Raw Audio". *arXiv:1609.03499 [cs]*. arXiv: 1609.03499.

Ott, Myle, Sergey Edunov, Alexei Baevski, et al. (Apr. 2019). *fairseq: A Fast, Extensible Toolkit for Sequence Modeling*. arXiv:1904.01038 [cs]. DOI: 10.48550/arXiv.1904.01038.

Panayotov, Vassil, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur (Apr. 2015). "Librispeech: An ASR corpus based on public domain audio books". *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ISSN: 2379-190X, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

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.

Park, Daniel S., William Chan, Yu Zhang, et al. (Sept. 2019). "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition". *Interspeech 2019*. arXiv: 1904.08779, pp. 2613–2617. DOI: 10.21437/Interspeech.2019-2680.

Park, Kyubyong and Jongseok Kim (2019). *g2pE*.

Paszke, Adam, Sam Gross, Francisco Massa, et al. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". *Advances in Neural Information Processing*

Systems. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, et al. Vol. 32. Curran Associates, Inc.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, et al. (Nov. 2011). "Scikit-learn: Machine Learning in Python". *The Journal of Machine Learning Research* 12.null, pp. 2825–2830. ISSN: 1532-4435.

Peters, B, G Bieker, SM Heckman, et al. (Mar. 2015). "Brain-Computer Interface Users Speak Up: The Virtual Users' Forum at the 2013 International Brain-Computer Interface Meeting". *Archives of physical medicine and rehabilitation* 96.3 0, S33–S37. ISSN: 0003-9993. DOI: `10.1016/j.apmr.2014.03.037`.

Prenger, Ryan, Rafael Valle, and Bryan Catanzaro (May 2019). "Waveglow: A Flow-based Generative Network for Speech Synthesis". *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brighton, United Kingdom: IEEE, pp. 3617–3621. ISBN: 978-1-4799-8131-1. DOI: `10.1109/ICASSP.2019.8683143`.

Rousseau, Marie-Christine, Karine Baumstarck, Marine Alessandrini, et al. (2015). "Quality of life in patients with locked-in syndrome: Evolution over a 6-year period." *Orphanet journal of rare diseases* 10, pp. 88–88. DOI: `10.1186/s13023-015-0304-z`.

Sadikaj, Gentiana and D. S. Moskowitz (Dec. 2018). "I hear but I don't see you: Interacting over phone reduces the accuracy of perceiving affiliation in the other". en. *Computers in Human Behavior* 89, pp. 140–147. ISSN: 0747-5632. DOI: `10.1016/j.chb.2018.08.004`.

Salari, Efraïm, Zachary V. Freudenburg, Mariska J. Vansteensel, and Nick F. Ramsey (2020). "Classification of Facial Expressions for Intended Display of Emotions Using Brain–Computer Interfaces". en. *Annals of Neurology* 88.3. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ana. pp. 631–636. ISSN: 1531-8249. DOI: 10.1002/ana.25821.

Seabold, Skipper and Josef Perktold (2010). "statsmodels: Econometric and statistical modeling with python". *9th Python in Science Conference*. 9th Python in Science Conference.

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.

Sumby, W. H. and Irwin Pollack (Mar. 1954). "Visual Contribution to Speech Intelligibility in Noise". en. *The Journal of the Acoustical Society of America* 26.2, pp. 212–215. ISSN: 0001-4966. DOI: 10.1121/1.1907309.

Umeda, Tatsuya, Tadashi Isa, and Yukio Nishimura (July 2019). "The somatosensory cortex receives information about motor output". *Science Advances* 5.7. Publisher: American Association for the Advancement of Science, eaaw5388. DOI: 10.1126/sciadv.aaw5388.

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.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, et al. (Mar. 2020). "SciPy 1.0: fundamental algorithms for scientific computing in Python". en. *Nature Methods* 17.3. Number: 3 Publisher: Nature Publishing Group, pp. 261–272. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0686-2.

Waskom, Michael (Apr. 2021). "seaborn: statistical data visualization". en. *Journal of Open Source Software* 6.60, p. 3021. ISSN: 2475-9066. DOI: 10.21105/joss.03021.

Willett, Francis R., Donald T. Avansino, Leigh R. Hochberg, et al. (May 2021a). "High-performance brain-to-text communication via handwriting". en. *Nature* 593.7858. Number: 7858 Publisher: Nature Publishing Group, pp. 249–254. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03506-2.

— (May 13, 2021b). "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.

Yamagishi, Junichi, Bela Usabaev, Simon King, et al. (July 2010). "Thousands of Voices for HMM-Based Speech Synthesis–Analysis and Application of TTS Systems Built on Various ASR Corpora". *IEEE Transactions on Audio, Speech, and Language Processing* 18.5, pp. 984–1004. ISSN: 1558-7916, 1558-7924. DOI: 10.1109/TASL.2010.2045237.

Yang, Yao-Yuan, Moto Hira, Zhaoheng Ni, et al. (May 2022). "Torchaudio: Building Blocks for Audio and Speech Processing". *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ISSN: 2379-190X, pp. 6982–6986. DOI: 10.1109/ICASSP43922.2022.9747236.

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:
Sean Metzger
C7BF35CADA804E9... Author Signature

8/29/2023
Date