AI-Driven Speech Neuroprostheses for Restoring Naturalistic Communication and Embodiment

Kaylo Littlejohn

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2025-147
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2025/EECS-2025-147.html

August 7, 2025

AI-Driven Speech Neuroprostheses for Restoring Naturalistic Communication and Embodiment

By

Kaylo Littlejohn

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering- Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Gopala Anumanchipalli, Co-chair
Professor Edward Chang, Co-chair
Professor Ren Ng
Professor Preeya Khanna

Summer 2025

AI-Driven Speech Neuroprostheses for Restoring Naturalistic Communication and Embodiment

Kaylo Littlejohn

Abstract

Can we rebuild the bridge between brain and voice, restoring human communication for people with paralysis? This thesis outlines our translational systems that restore speech to individuals with vocal-tract paralysis.

Speech neuroprostheses have the potential to restore communication and embodiment to individuals living with paralysis, but achieving naturalistic speed and expressivity has remained elusive. The advances presented in this thesis enabled a clinical trial participant with severe limb and vocal paralysis to "speak again" for the first time in 18+ years using an AI "brain-to-voice" decoder that restores their pre-injury voice. We use high-density surface recordings of the speech cortex in a participant to achieve high-performance, large-vocabulary, real-time decoding across three complementary speech-related output modalities: text, speech audio, and facial-avatar animation. Leveraging advances in machine learning for automatic speech recognition and synthesis, we trained and evaluated deep-learning models using neural data collected as participants attempted to silently speak a sentence, enabling decoding speeds approaching natural conversational rates. We also demonstrate the control of virtual orofacial movements for speech and non-speech communicative gestures via a high-fidelity "digital talking avatar" controlled by the participant's brain.

Building on the above advances in high-performance brain-to-speech decoding, I outline our findings demonstrating low-latency, continuously streaming brain-to-voice synthesis with neural decoding in 80-ms increments. The recurrent neural network transducer models demonstrated implicit speech detection capabilities and could continuously decode speech indefinitely, enabling uninterrupted use of the decoder and further increasing speed. Our framework also successfully generalized to other silent-speech interfaces, including single-unit recordings and electromyography.

Together, the findings in this thesis introduce a multimodal, low-latency speech-neuroprosthetic approach with substantial promise for restoring full, embodied communication to people with severe paralysis.

A video overview of our brain decoding technique and impact can be found at this link.

To Ronald and Donna Doherty

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank, first and foremost, the participants B1, B2, B3, and their families and caregivers for their tireless contributions to the clinical trial. Their persistent and consistent effort, as well as their selflessness in undertaking research for the benefit of others, was essential to advancing the field of speech BCIs and enabling future clinical neuroprosthetic systems for many who need them. They are the Apollo astronauts of brain-computer interfaces, and I can't thank them enough for their efforts.

Next, I'd like to thank my co-advisor, Dr. Edward Chang, for his visionary efforts in establishing the field of speech-BCIs and the BRAVO clinical trial. Eddie has been a fantastic person to work with over the past five years and has been a great source of inspiration for me and many others. Without Eddie, we'd have no clinical trial or many fundamental contributions in the domain of speech and auditory neuroscience. Thank you for the opportunity to conduct research in the lab, for your pioneering work, and your dedication to BRAVO.

Thank you to my advisor, Gopala K. Anumanchipalli. It has been a pleasure to work with you and grow within the lab. I joined at the same time as Gopala (actually, one semester before), and it's been great to see the group grow over the years. I remember the days when the lab had only two people, then grew to four, then ten, and now has roughly 10 PhD students, triple the number of undergrads, and numerous collaborators (114 on Slack!). You've been extremely helpful in shaping the high-level vision for my PhD, as well as the thought process for research in general, especially encouraging me to focus on the vision of what I want to create in the world. Additionally, many, if not the substantial majority, of the technical ideas in this thesis were a direct result of discussions you've had with me and/or teammates (e.g., CTC and RNN-T for speech decoding). You also gave me a lot of academic freedom to choose any problem I wanted to pursue (within reason) and work at my own pace, provided I achieved results, which helped me throughout my research career at Berkeley.

Thanks Sean, Alex, David, Margaret, Jessie, Max, Sam, Irina, Ran, Cady, Vanessa, Karunesh, and Josh for your contributions to the Chang Lab at UCSF and the BRAVO trial. It's been a great deal of fun working with you all to collect data and stay up late at night trying to train the models and prepare for demos and decoding. There have been many ups and downs, as well as exciting discoveries, along the way, and I genuinely miss the days we spent together trying to solve challenging problems in BCI, working alongside the participants. Together, we achieved something truly special for the world by designing a 0-to-1 technology that decodes transcription, voice, and animation from the brain, and I'm incredibly proud to be one of your teammates in achieving this significant milestone. The impact of the series of 5-8 papers produced within BRAVO cannot be understated. It seems so clear to us why this is possible based on prior articulatory coding papers, but it is truly like something out of a movie. Thanks to the broader Chang Lab as well for your feedback and support during this

process.

Thanks, Cheol Jun, Adit, Anshul, Inga, and the rest of the team at Berkeley Speech Group and Berkeley AI Research, for your technical contributions and for being a fantastic group of engineers who push the boundaries of spoken language modeling and bringing significant machine learning expertise to the clinical trial. Thanks to Speech Graphics for collaborating to bring brain-to-avatar technology into a reality.

Thank you to my dissertation committee members, Ren and Preeya, for your feedback throughout my PhD, and thank you to Bruno, Yubei, Michael Chang, Paul Sajda, Alex Krem, and Tyler Bonnen for your mentorship. Thank you to the BCI Society and BCI community (industry and academia alike) for creating such an amazing environment of innovation and for your support and camaraderie in pushing the field forward.

I would also like to thank my Mom, Grandmother, Savier, Abby, Jacob, Kemar, Jeniesha, Serenity, Ocean, Adam, Jen, and Uncle Jeff for their support throughout my PhD. Thank you to Mitch Markowitz and Coach Patrick Kelly for your mentorship. Thank you to my fiancée, Trang, and her family, Hieu, Kiet, Hao, and the extended family, for their support.

1 Introduction

1.1 Speech Neuroprosthesis Motivation and Paradigm

Loss of speech after paralysis is devastating, but circumventing motor-pathway injury by directly decoding speech from intact cortical activity has the potential to restore natural communication and self-expression. Recent discoveries have defined how key features of speech production are facilitated by the coordinated activity of vocal-tract articulatory and motor-planning cortical representations (Bouchard et al., 2013; Chartier et al., 2018). Successful speech decoding was performed first in individuals implanted with intracranial electrodes for clinical epilepsy monitoring (Anumanchipalli et al., 2019) and subsequently in individuals with paralysis as part of early feasibility clinical trials to restore speech (Moses et al., 2021; Metzger et al., 2023; Littlejohn et al., 2024), the latter two citations being the primary subject of this thesis. Although restoring natural speech is a long-term goal, speech neuroprostheses already have performance levels that surpass communication rates offered by current assistive-communication technology (Silva et al., 2024a). Losing the ability to speak drastically hinders communication and, as a result, substantially reduces quality of life. Diseases that cause injury to descending motor-neuron tracts in the brainstem, such as amyotrophic lateral sclerosis and brainstem stroke, can leave individuals paralysed, with little-to-no voluntary muscle control. In some cases, this can result in incomplete or total locked-in syndrome, in which almost all forms of natural communication are precluded.

Augmentative and alternative communication (AAC) devices leverage residual voluntary motor function, such as eye or head movements, to allow individuals with paralysis to spell out intended messages, albeit through slow and effortful interfaces (Chang & Anumanchipalli, 2020). Communication neuroprostheses that decode cortical activity into attempted cursor movements (Pandarinath et al., 2017) or handwriting (Willett et al., 2023b) have made strides in achieving faster spelling in individuals with paralysis. However, communication rates for

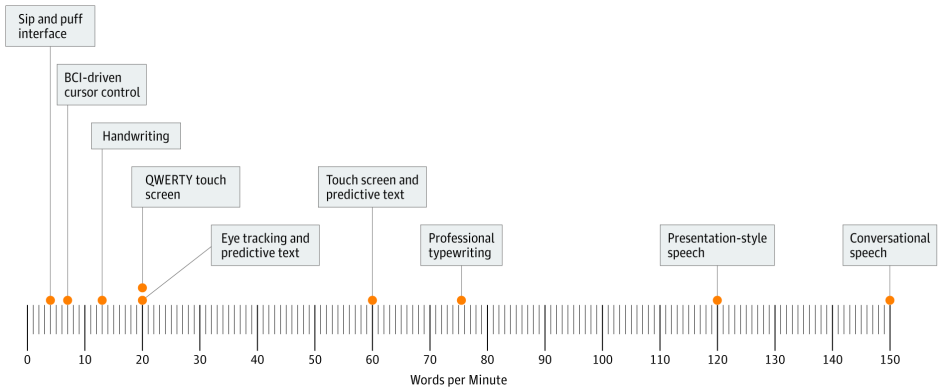

Figure 1.1: **Comparison of Communication Rates Across Various Modalities, Measured as the Average Number of Words per Minute in a Typical Scenario.** BCI indicates brain-computer interface. Assistive communication interfaces are much slower than speech. Adapted from Chang et al. 2020 (Chang & Anumanchipalli, 2020).

neuroprostheses controlled by attempted hand movements remain slower and less expressive than natural speech (Chang & Anumanchipalli, 2020). On the other hand, speech enables efficient and rapid communication at rates of around 150–200 words per minute (WPM) in conversational settings. This is driven by millisecond-level coordination of more than 100 vocal-tract muscles that articulate speech sounds to express language (Figure 1.1).

A speech neuroprosthesis is a device that uses algorithms to translate brain activity during intended speech into communication signals, for example, text (such as words or sentences on a screen), acoustics (such as vocalized sounds or phrases) or facial movements that accompany speech. Speech neuroprostheses have the potential to not only enable more natural communication of words and sentences but also restore other expressive components of communication that convey meaning, such as intonation, loudness and facial gestures (Silva et al., 2024a; Chang & Anumanchipalli, 2020). Advancements in speech neuroscience, neural-interface technology and machine learning have accelerated progress towards the goal of a clinically viable speech neuroprosthesis. Studies furthering our basic understanding of the cortical encoding of speech features, notably motor control of the vocal-tract articulators during speech, laid the groundwork for a speech neuroprosthesis that decoded text in the form of words and sentences from the cortical activity of an individual with incomplete locked-in syndrome who relied on AAC methods to communicate (Bouchard et al., 2013; Moses et al., 2021). Our work presented in this thesis expanded this initial demonstration by directly decoding cortical activity into audible speech (Metzger et al., 2023; Littlejohn et al., 2024) and allowing more generalizable and rapid text decoding (Card et al., 2024).

The core goal of a speech neuroprosthesis is to transform neural activity during intended speech into communication units, such as text, audible sounds or orofacial movements. Text

and synthesized speech are two outputs commonly used in speech-decoding systems (Moses et al., 2021; Metzger et al., 2022; Anumanchipalli et al., 2019; Willett et al., 2023b; Wairagkar et al., 2023; Littlejohn et al., 2024; Silva et al., 2024b; Angrick et al., 2023). For text decoding, decoders are trained to predict a discrete set of linguistic features - such as characters, phonemes, words or sentences - that correspond to the intended speech of the user. For speech synthesis, acoustic features, such as the mel-spectrogram and pitch, are decoded and mapped to a final acoustic waveform that can be played back to the user. Thirdly, articulatory gestures may be decoded and used drive the animation of a live-action avatar (Metzger et al., 2023), which we demonstrate in this thesis (Figure 1.2).

1.2 Text Decoding

A first approach to text decoding is to classify neural activity from isolated, intended speech as a word or sentence in a predefined vocabulary. In able speakers, single-word classification has been highly successful using predefined and restricted vocabularies (Berezutskaya et al., 2022; Martin et al., 2018). To scale to longer segments of speech, pre-defined sentence sets have been targeted instead of single words. Successful classification of a limited set of produced sentences in the setting of question-and-answer dialogue has been demonstrated, where incorporating context from the predicted question improved classification of the answer. To build on this result and move beyond direct classification of words or sentences, Makin et al. 2020 (Makin et al., 2020) applied a recurrent neural network (RNN) encoder-decoder framework to encode temporal patterns in neural activity during sentence production into an abstract representation that was then decoded word-by-word into phrases. By targeting words within sentences, rather than whole sentences, and leveraging contemporary machine-learning techniques, this RNN-based approach improved accuracy and the overall vocabulary size and number of sentences that could be decoded. Despite their successes, these approaches are all limited by restrictive, predefined vocabulary sizes. To address this, several studies have drawn inspiration from the field of automatic speech recognition (ASR) to facilitate generalizability to larger vocabularies.

One such approach to generalize decoding to larger vocabularies is to decode subword linguistic units, such as phonemes or characters, rather than individual words or sentences. This is a common ASR approach in which language models — trained to capture the statistical patterns of subword units and words — are used to convert decoded phoneme or character sequences into sentences. Multiple studies have demonstrated that isolated phonemes can be decoded from neural activity (Herff & Schultz, 2016; Mugler et al., 2018). Herff et al. 2016 further demonstrated that, during continuous speech, phoneme sequences could be decoded provided that synchronized produced acoustics were available to aid model training (Herff & Schultz, 2016). A language model applied to the decoded phoneme sequences could then generate sentences. Subsequent work built on this finding, applying additional ASR

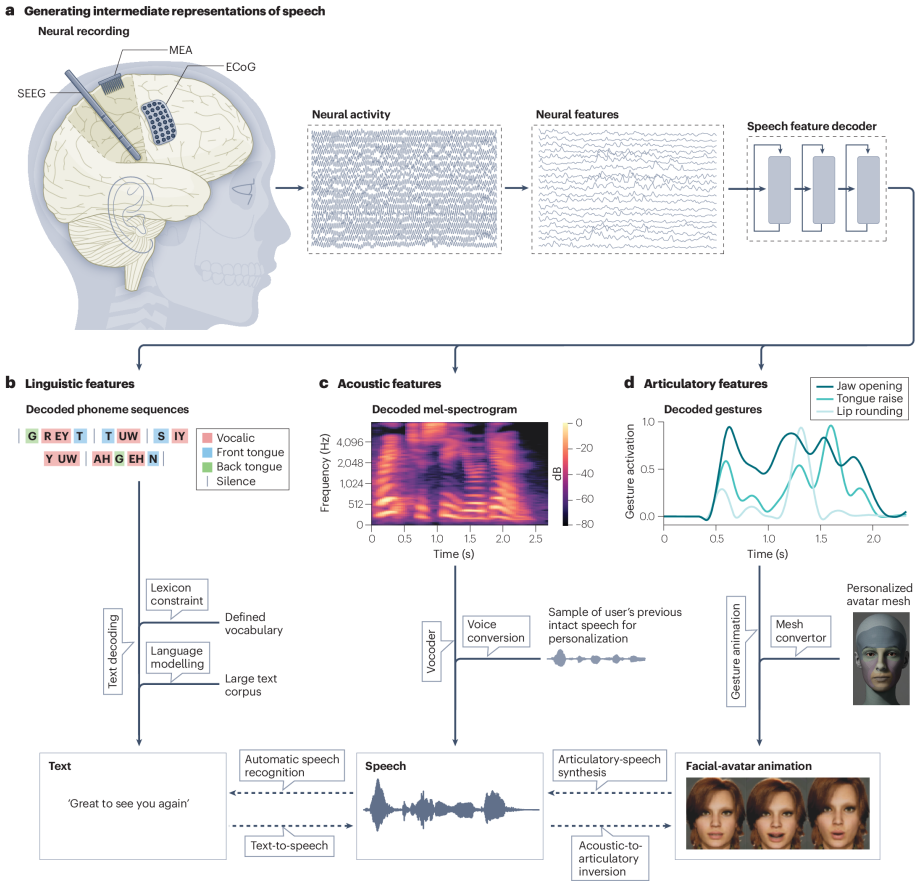

Figure 1.2: **Decoding Speech from Neural Activity a**, Recording of neural activity can be achieved using different neural interfaces such as electrocorticography (ECoG), microelectrode array (MEA) and stereoelectroencephalography (SEEG). Recorded neural activity is processed into neural features, which are then passed to a speech feature decoder, which might have been trained to output linguistic, acoustic or articulatory features as intermediate speech representations. **b**, For text decoding, models can be trained to decode neural features into sequences of linguistic features, such as phonemes, and then a defined vocabulary and natural language modelling can be used to transform phoneme sequences into text sequences of plausible words and sentences. **c**, For speech synthesis, models can be trained to decode neural features into sequences of acoustic features, such as the mel-spectrogram, which can then be vocoded into an audible speech waveform, often using pretrained models from the field of speech processing. Importantly, the vocoder can be personalized in a way that captures the previous intact voice of the individual. **d**, Models may also be trained to decode neural features into sequences of articulatory features, such as the relative displacement of different locations in the vocal tract over time. A gesture-animation system may be applied to the gesture activation sequences to animate a digital avatar. Similar to speech synthesis, the avatar may be personalized to reflect the likeness of the users, using digital-face capture software. Optional conversion between text, speech and facial-avatar animation outputs is feasible using pretrained speech-processing models (dashed arrows).

techniques to mitigate the need to synchronize linguistic subunits directly with the neural data. By using the connectionist temporal classification (CTC) loss (Graves et al., 2006), which scores the loss between predictions from an input sequence (such as neural activity) and an output sequence (such as linguistic subunits) without requiring their alignment, Sun et al. (Sun et al., 2020) trained an RNN to decode character sequences from brain activity during sentence production. Language models similarly converted the decoded character sequences into sentences. Even though precise alignment between the input neural activity and output character sequences was not required, aligning acoustic features to the brain activity during model training increased performance. Here, as in many text-decoding studies and ASR approaches, performance evaluation used word, character and/or phoneme error rates (WER, CER and PER, respectively). Error rates are defined as the edit distance between the ground-truth and decoded sequences.

In able speakers, neural activity could be aligned to their produced speech acoustics to improve model performance. However, for individuals who cannot speak, there may be no way to reliably align neural activity with ground-truth produced acoustics. It may be possible to align neural activity to the onsets or offsets of attempted speech, through audio (if the individual is able to vocalize) or through video of the face (if the individual retains some residual motor function), but these methods require separate annotation (and may not be possible if the individual is fully locked in). As a first approach to counter a lack of alignment between neural activity and ground-truth targets, Moses et al. 2021 instructed a participant with anarthria to attempt to speak sentences word-by-word (with pauses between words) (Moses et al., 2021). A speech-detection model used cortical activity, primarily from the SMC, to predict when the individual was attempting to speak a word and passed the corresponding neural features to a word-classification model trained to predict probabilities across 50 words in a predefined vocabulary. A language model combined the neural-based probabilities of each word with linguistic likelihoods of word sequences, improving WERs of the decoded sentences. Although the speech-detection model used provided word-level alignment, direct classification from a 50-word vocabulary limited this approach. The development of a spelling system in which the same participant with anarthria attempted to silently speak 26 NATO code words — phonetically discriminable words which represent each letter in the alphabet — rather than the 50 common words previously used facilitated access to a larger vocabulary (Metzger et al., 2022). The individual then had the ability to spell character-by-character and language models could transform decoded-character sequences into sentences comprised from a vocabulary of more than 1,000 words. Although this approach facilitated access to a large vocabulary, while still using intended speech, decoding speeds were slower and communication was less natural for the individual.

Recent works have achieved large-vocabulary decoding while maintaining decoding speeds closer to natural speech, by using CTC loss. Three studies (each in a different individual with vocal-tract paralysis: two with MEAs (Card et al., 2024; Willett et al., 2023b) and one with ECoG (Metzger et al., 2023)) leveraged RNN models trained to map an input

sequence of neural activity to an output sequence of phonemes without the need for alignment between the neural activity and ground-truth sequence of phonemes. Language models then mapped decoded phoneme sequences into words and sentences. A key advantage of using CTC loss is that the individual can more naturally produce a phrase in this approach, without having to associate isolated speech attempts with individual words. Furthermore, decoders can generalize to larger vocabularies by targeting low-dimensional subword units (such as phonemes) that can be built into a much higher-dimensional space of words and sentences. Overall, this has led to state-of-the-art decoding performance for individuals with paralysis.

1.3 Speech Synthesis

An alternative to decoding text is to synthesize audible speech from brain activity. As spoken language is a more fundamental form of human communication than written language (text), this approach can enable more fine-grained and natural control over decoded outputs. Furthermore, self-perception of speech is an important component to speech-motor control. It is possible that low-latency restoration of audible speech could have analogous benefits to rapid closed-loop feedback for neuroprosthetic control in other motor domains. Finally, an individual may feel greater embodiment when using a speech neuroprosthesis with a personalized voice that reflects their likeness (Littlejohn et al., 2024; Metzger et al., 2023). However, these potential benefits come with increased difficulty; speech synthesis has generally proven more challenging than text decoding, as it does not leverage language models and predefined vocabularies.

Initial work to synthesize speech from neural activity in able speakers has leveraged a few approaches (Wairagkar et al., 2023; Herff et al., 2019; Anumanchipalli et al., 2019; Angrick et al., 2019). Herff et al. 2019 developed a concatenative synthesizer to transform neural activity recorded from SEEG into audible speech (Herff et al., 2019). They first built a brain-to-speech lookup library during training by associating 150 ms neural activity segments with the synchronized 150 ms of audible speech. During evaluation, consecutive 150 ms windows of neural activity were correlated with each neural entry in the lookup library and the speech segments corresponding to the highest correlated neural entry were concatenated together to form a decoded speech waveform. An advantage of this concatenative-synthesis approach is its feasibility with smaller dataset sizes; however, it does not fully leverage advances in machine learning, relying on time window correlations.

Another approach, pursued by Angrick et al. 2019 and Anumanchipalli et al. 2019, involved a two-stage decoding process. In the first stage, a deep-learning model is trained to regress neural activity (recorded from SEEG (Angrick et al., 2019) or ECoG (Anumanchipalli et al., 2019)) to a time series of acoustic features, such as the mel-spectrogram. In the second stage, a speech synthesizer — a vocoder — is used to convert the acoustic representation into an

audible speech waveform. This acoustic regression approach achieved higher-performance speech synthesis, probably owing to the ability of a deep-neural network to learn complex nonlinear mappings between inputs and outputs. Anumanchipalli et al. 2019 added a first step of decoding an articulatory representation from neural activity, which was then mapped to the intermediate acoustic representation and vocoded into speech (Anumanchipalli et al., 2019). Their approach further improved speech-synthesis performance by leveraging the underlying articulatory organization of the SMC. In both implementations, computing correlation or distortion between the mel-spectrogram of decoded and ground-truth speech assessed performance. Anumanchipalli et al. 2019 also computed a human-transcribed WER by asking volunteers to transcribe decoded speech to text, which can then be compared with the ground truth (Anumanchipalli et al., 2019).

The previously described speech-synthesis approaches required precise alignment between neural activity and ground-truth acoustics. However, as described earlier, a major challenge of creating speech neuroprostheses for individuals with vocal-tract paralysis is the lack of intelligible ground-truth acoustic signals to align with the neural data during training. Individuals with dysarthria may retain some intelligible vocalization over single, isolated words, allowing the acoustic regression approach to prove successful in synthesizing single words (Angrick et al., 2023). However, individuals with severe cases of dysarthria or anarthria may lack the ability to make any intelligible vocalizations, especially for longer sentences. Similar to text decoding, a way to circumvent the need for alignment is to use CTC loss to train models that map input sequences of neural activity to output sequences of acoustic signals that may then be vocoded into speech. Our work in this thesis used this approach to decode neural activity during silent speech attempts from a participant with anarthria into synthesized speech; however, rather than regressing the mel-spectrogram, they decoded input sequences of neural activity into output sequences of discrete acoustic-speech units (Metzger et al., 2023). During training, a large self-supervised audio model converted target waveforms (generated from a text-to-speech (TTS) model) into sequences of discrete acoustic-speech units. During online inference, the decoded discrete acoustic-speech unit sequences were vocoded into sentence-level speech that was intelligible to untrained human listeners. Although this approach facilitated alignment-free speech synthesis, suited to scale to fully locked in individuals, it is not yet ideal for low-latency streaming. In the last chapter of this thesis, we used another sequence-to-sequence loss, known as the RNN-T loss (Graves, 2012), to achieve large-vocabulary streaming brain-to-voice synthesis (Littlejohn et al., 2024).

Because speech is variable across people and expressive, neuroprostheses that synthesize speech are well suited for personalization. A voice-conversion model can be applied to convert decoded speech waveforms into personalized waveforms that resemble the likeness of the user. Voice-conversion models require as little as 3 s of recorded speech (Casanova et al., 2022); however, larger personalized training datasets can be leveraged if the individual has additional pre-injury audio recordings. The work presented in this thesis first used this voice-conversion approach to personalize synthesized speech for an individual with vocal-tract paralysis using

speech samples recorded before the injury (Metzger et al., 2023). Card et al. 2024 trained a personalized TTS model on a small corpus of pre-injury speech, which is an alternative to using a voice-conversion model (Card et al., 2024). In certain cases, individuals could create voice banks of themselves speaking defined lists of natural sentences, which might be particularly relevant for neurodegenerative diseases, such as ALS, in which people anticipate losing the ability to speak in the near future (Yamagishi et al., 2012).

Given the strong link between speech acoustics and underlying vocal-tract movements, articulation constitutes another relevant feature space for speech decoding and speech synthesis (Chartier et al., 2018; Bouchard et al., 2013). Previous work has demonstrated that first decoding articulatory features and transforming them to acoustic features improved the quality of synthesized speech (compared with decoding acoustics directly from neural data) (Anumanchipalli et al., 2019). In addition, articulatory features may be useful as a standalone output space for both speech and non-speech orofacial gestures (Salari et al., 2020). Non-verbal facial expressions that accompany virtual or face-to-face communication provide numerous benefits compared with audio-only communication, including improved conveyance of emotion and attitude, that provides increased clarity (Sumby & Pollack, 1954). Our work covered in this thesis demonstrated that it is possible to use these decoded articulatory features to animate a digital avatar face, which was personalized to resemble the likeness of the individual (Metzger et al., 2023).

The central thesis of this dissertation is that neural activity from the speech motor cortex can be decoded in real time to restore expressive, high-rate communication for individuals with severe paralysis, through systems that produce text, audible speech, and articulatory-driven facial animation. By leveraging large-vocabulary decoding methods, alignment-free training strategies, and personalization techniques, speech neuroprostheses can approach the speed and naturalness of able-bodied speech (within a research setting).

2

High Performance Speech Decoding and Avatar Control

2.1 Summary

Speech neuroprostheses have the potential to restore communication to people living with paralysis, but naturalistic speed and expressivity are elusive (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 and personalization to the participant's pre-injury voice. For facial-avatar animation, we demonstrate the control of virtual orofacial movements for speech and non-speech communicative gestures. The decoders reached high performance with less than two weeks of training. Our findings introduce a multimodal speech-neuroprosthetic approach that has substantial promise to restore full, embodied communication to people living with severe paralysis.

2.2 Main

Speech is the ability to express thoughts and ideas through spoken words. Speech loss after neurological injury is devastating because it substantially 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 (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. Although 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 (ECoG) to decode text and audio-visual speech outputs from articulatory vocal-tract representations distributed throughout the sensorimotor cortex (SMC). Owing to severe paralysis caused by a basilar-artery brainstem stroke that occurred more than 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 has used commercial head-tracking assistive technology to communicate slowly to select letters at up to 14 words per minute (WPM). 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 SMC that have persisted after years of paralysis.

2.3 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 ECoG array implanted over speech cortical areas of the SMC and superior temporal gyrus (Figure 2.1). 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.

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. This differs from imagined or inner speech because she was trying to engage her articulators to the best of her ability, although substantial orofacial weakness prevents her from naturally mouthing words. Meanwhile, we processed neural signals recorded from all 253 ECoG electrodes to extract high-gamma activity (HGA; between 70 and 150 Hz) and low-frequency signals (between 0.3 and 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.

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., 1998). 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 from this set we sub-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 film 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 before real-time testing, we recorded ECoG data as the participant silently attempted to speak individual sentences. A major difficulty in learning statistical mappings between the ECoG features and the sequences of phones and speech-sound features in the sentences was caused by the absence of clear timing information of words and phonemes in the silently attempted speech. To overcome this, 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, speech and articulatory decoding models to enable prediction of phone probabilities, discrete speech-sound units and discrete articulator movements, respectively, from the ECoG signals.

2.4 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 (Moses et al., 2021; Metzger et al., 2022). Here we address these limitations by implementing a flexible approach using phone decoding, enabling decoding of arbitrary phrases from large vocabularies while approaching naturalistic speaking rates.

To evaluate real-time performance (Figure 2.2), we decoded text as the participant attempted to silently say 249 randomly selected sentences from the 1024-word-General set that were

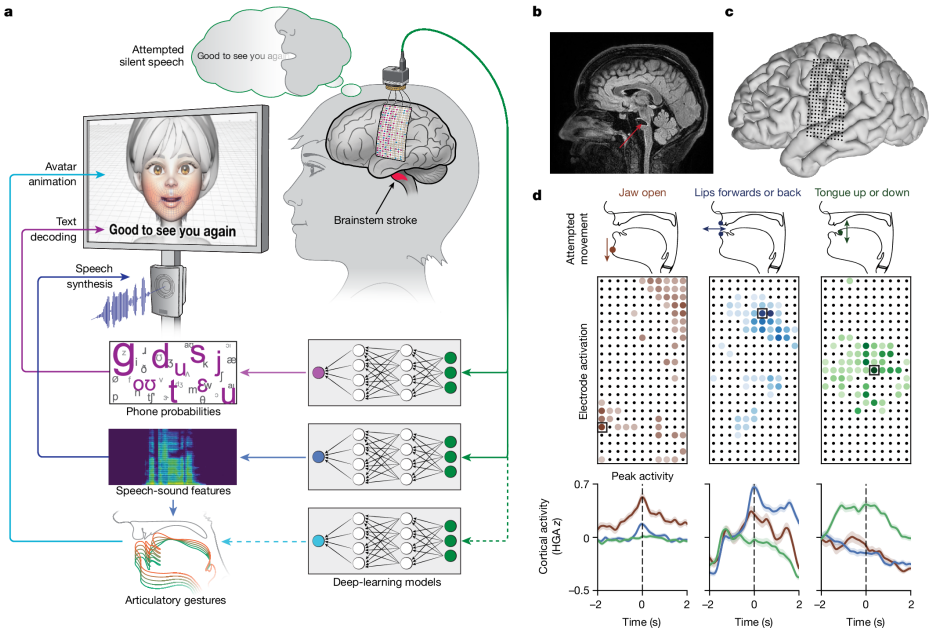

Figure 2.1: **Multimodal speech decoding in a participant with vocal-tract paralysis.** **a**, Overview of the speech-decoding pipeline. A brainstem-stroke survivor with anarthria was implanted with a 253-channel high-density ECoG array 18 years after injury. Neural activity was processed and used to 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 magnetic resonance imaging scan showing brainstem atrophy (in the bilateral pons; red arrow) resulting from stroke. **c**, Magnetic resonance imaging reconstruction of the participant's brain overlaid with the locations of implanted electrodes. The ECoG array was implanted over the participant's lateral cortex, centred 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. Colour indicates the magnitude of the average evoked HGA response with each type of movement. Bottom: z-scored trial-averaged evoked HGA responses with each movement type for each of the outlined 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 (n = 130 trials for jaw open, n = 260 trials each for lips forwards or back and tongue up or down).

not used during model training. To decode text, we streamed features extracted from ECoG signals starting 500 ms before the go cue into a bidirectional recurrent neural network (RNN). Before 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 WPM. WER, PER and CER measure the percentage of decoded words, phones and characters, respectively, that were incorrect (see Table 2.1 for example decodes).

We computed error rates across sequential pseudo-blocks of ten-sentence segments (and one pseudo-block of nine sentences) using text decoded during real-time evaluation. We achieved a median PER of 18.5% (99% confidence interval (CI) [14.1, 28.5]), a median WER of 25.5% (99% CI [19.3, 34.5]) and a median CER of 19.9% (99% CI [15.0, 30.1]). 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 five-way Holm–Bonferroni correction). The average WER passes the 30% threshold below which speech-recognition applications generally become useful (Watanabe et al., 2017) 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 by augmenting the 1024-word-General vocabulary with the n - 1,024 most frequently occurring words outside this set in large-scale corpora, in which 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 carried out decoding with the real-time evaluation trials. We observed robust decoding performance as vocabulary size grew. 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]. This

decoding rate exceeds our participant's typical communication rate using her assistive device (14.2 WPM) and is closer to naturalistic speaking rates than has been previously reported with communication neuroprostheses (Moses et al., 2021; Metzger et al., 2022; Vansteensel et al., 2016; Pandarinath et al., 2017; Willett et al., 2021).

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]), which is only 10.9 percentage points higher than that of 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 five-way Holm–Bonferroni correction).

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 five models with different random initializations on all of the data collected on or before that date, and then simulated performance on the real-time blocks. We observed steadily declining error rates over the course of 13 days of training-data collection, during which we collected 9,506 sentence trials corresponding to about 1.6 h 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 that of our previous work (Moses et al., 2021; Metzger et al., 2022) and is likely to continue to improve with more data.

To assess signal stability, we measured real-time classification performance during a separate word and motor task that we collected data for during each research session with our participant. In each trial of this task, we prompted the participant to either attempt to silently say one of the 26 code words from the NATO (North Atlantic Treaty Organization) phonetic alphabet (alpha, bravo, charlie and so forth) or attempt one of four hand movements (described and analysed in a later section). We trained a neural-network classifier to predict the most likely NATO code word from a 4-s window of ECoG features (aligned to the task go cue) and evaluated real-time performance with the classifier during the NATO-motor task. We continued to retrain the model using data available 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 eight of these sessions. Accuracy remained high after a 61-day hiatus 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 predefined 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, and 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]). These results illustrate extremely rapid and accurate decoding for finite, predefined sentences that could be used frequently by users.

Target Sentence	**Decoded Sentence**	**WER (%)**	**Percentile (%)**
You should have let me do the talking	You should have let me do the talking	0	44.6
I think I need a little air	I think I need a little air	0	44.6
Do you want to get some coffee	Do you want to get some coffee	0	44.6
What do you get if you finish	Why do you get if you finish	14	47.0
Did you know him very well	Did you know him well	17	49.4
You got your wish	You get your wish	25	61.8
No tell me why	So tell me why	25	61.8
You have no right to keep us here	You have no right to be out here	25	61.8
Why would they come to me	Why would they have to be	33	65.1
Come here I want to show you something	Have here I want to do something	38	65.5
All I told them was the truth	Can I do that was the truth	43	70.3
You got it all in your head	You got here all your right	43	70.3
Is she a friend of yours	I see afraid of yours	67	85.1
How is your cold	Your old	75	89.2

Table 2.1: **Illustrative text-decoding examples for the 1024-word-General set**. Examples are shown for various levels of WER during real-time decoding with the 1024-word-General set. Each percentile value indicates the percentage of decoded sentences that had a WER less than or equal to the WER of the provided example sentence.

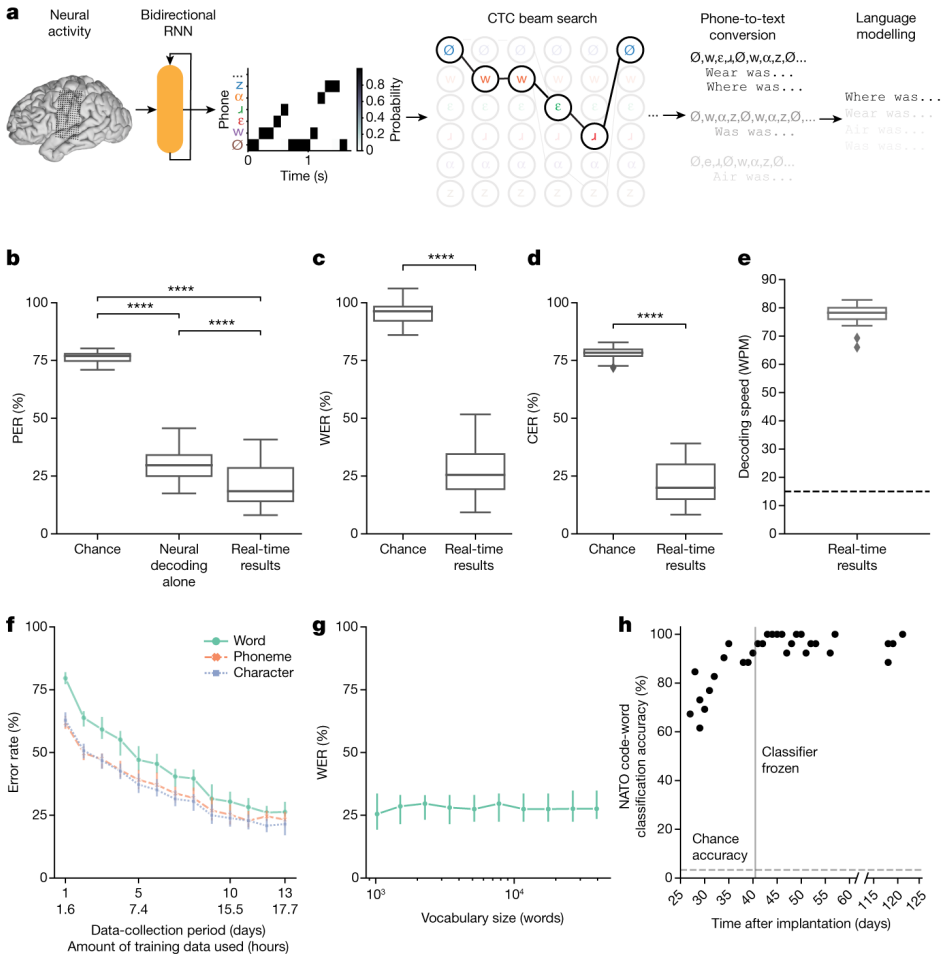

Figure 2.2: **High-performance text decoding from neural activity.** **a**, During attempts by the participant to silently speak, a bidirectional RNN decodes neural features into a time series of phone and silence (denoted as Ø) probabilities. From these probabilities, a CTC beam search computes the 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 PERs, calculated using shuffled neural data (Chance), neural decoding without applying vocabulary constraints or language modelling (Neural decoding alone) and the full real-time system (Real-time results). **c, d,** Word (**c**) and character (**d**) error rates for chance and real-time results. In **b-d**, ****P $<$ 0.0001, two-sided Wilcoxon signed-rank test with five-way Holm–Bonferroni correction for multiple comparisons. **e**, Decoded WPM. Dashed line denotes previous state-of-the-art speech BCI decoding rate in a person with paralysis (Moses et al., 2021). **f**, Offline evaluation of error rates as a function of training-data quantity. **g**, Offline evaluation of WER as a function of the number of words used to apply vocabulary constraints and train the language model. **h**, Decoder stability as assessed using real-time classification accuracy during attempts to silently say 26 NATO code words across days and weeks. The vertical line represents when the classifier was no longer retrained before each session.

2.5 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 paralysed.

We carried out 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 2.3). To synthesize speech, we passed time windows of neural activity around the go cue into a bidirectional RNN. Before 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., 2021) that encodes a continuous speech waveform into a temporal sequence of discrete speech units that captures latent phonetic and articulatory representations (Cho et al., 2023). Because our participant cannot speak, we acquired reference speech waveforms from a recruited speaker for the AAC sentence sets or using 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 pretrained unit-to-speech model that first generated a mel spectrogram and then 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 timescale information with corresponding reference spectrograms. 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. 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 (Yamagishi et al., 2010) and better than those in previous neural-decoding work with participants that were able to speak naturally (Anumanchipalli et al., 2019).

Human-transcription assessments are a standard method to quantify the perceptual accuracy of synthesized speech (Wolters et al., 2010). 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 WERs 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]) and 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. As for 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.

2.6 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 substantial 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 & Moskowitz, 2018) and intelligibility (Sumby & Pollack, 1954). Therefore, animation of a facial avatar to accompany synthesized speech and further embody the user is a promising means towards 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 brain–computer interface (BCI) to decode neural activity into articulatory speech gestures and render a dynamically moving virtual face during the audio-visual task condition (Figures 2.4 and 2.5).

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). This technology uses speech-to-gesture methods that predict articulatory gestures from sound waveforms and then synthesizes the avatar animation from these gestures (Berger et al., 2011). We designed a three-dimensional (3D) virtual environment

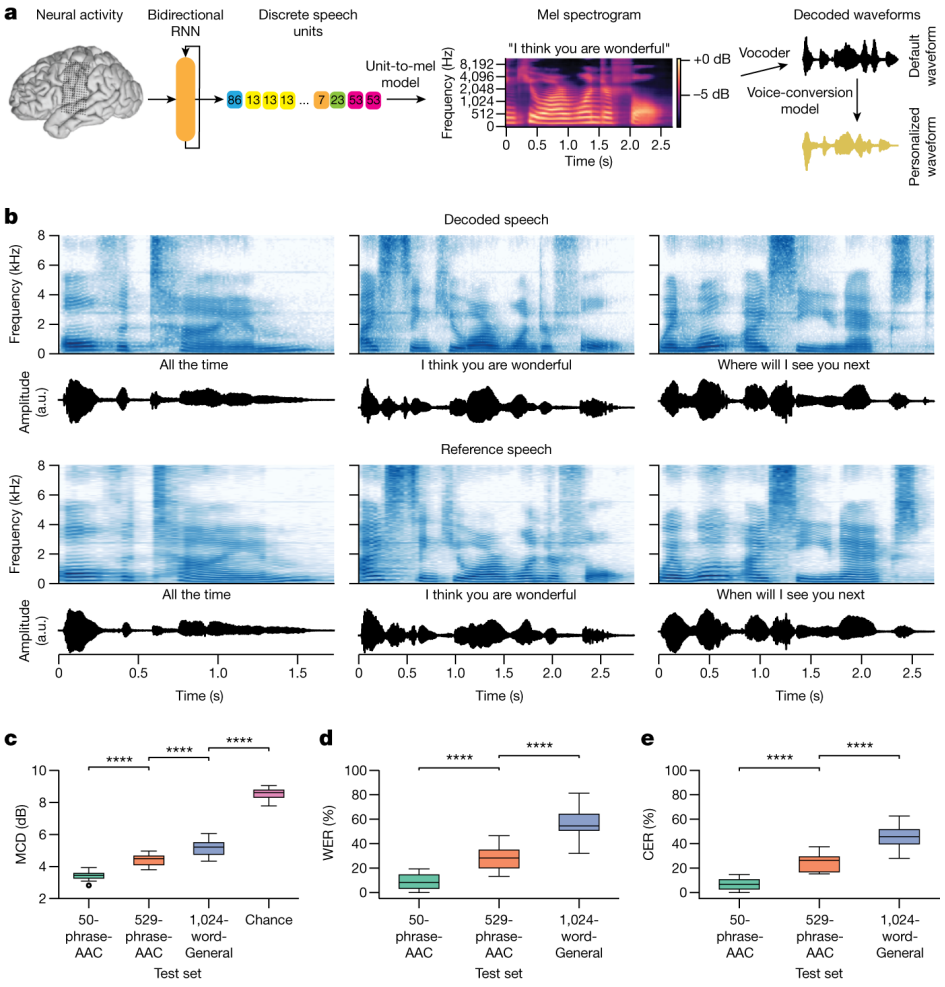

Figure 2.3: **Intelligible speech synthesis from neural activity.** **a**, Schematic diagram of the speech-synthesis decoding algorithm. During attempts by the participant to silently speak, a bidirectional RNN decodes neural features into a time series of discrete speech units. The RNN was trained using 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 two rows: three example decoded spectrograms and accompanying perceptual transcriptions (top) and waveforms (bottom) from the 529-phrase-AAC sentence set. Bottom two rows: the corresponding reference spectrograms, transcriptions and waveforms representing the decoding targets. **c**, MCDs for the decoded waveforms during real-time evaluation with the three sentence sets and from chance waveforms computed offline. **d**, Perceptual WERs from untrained human evaluators during a transcription task. **e**, Perceptual CERs from the same human-evaluation results as **d**. In **b**–**e**, all decoded waveforms, spectrograms and quantitative results use the non-personalized voice. A.u., arbitrary units.

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 whether 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 acoustic waveforms through the animation system's speech-to-gesture model. We then discretized the articulatory gestures using a vector-quantized variational autoencoder (VQ-VAE) (van den 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 (Figure 2.6).

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 and 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.

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) (King, 2009) to avatar and healthy-speaker videos to extract trajectories important for speech: jaw opening, lip aperture and mouth width. For each pseudo-block of ten test sentences, we computed the mean correlations across sentences between the

trajectory values for each possible pair of corresponding videos (36 total combinations with 1 avatar and 8 healthy-speaker videos). Before calculating correlations between two trajectories for the same sentence, we applied dynamic time warping 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. Although correlations among pairs of healthy speakers were higher than between the avatar and healthy speakers (all $P < 0.0001$, two-sided Mann–Whitney U-test with nine-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 dynamic time warping (all $P < 0.0001$, two-sided Mann–Whitney U-test with nine-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 carried out two additional tasks: an articulatory-movement task and an emotional-expression task. In the articulatory-movement task, the participant attempted to produce six orofacial movements. In the emotional-expression task, the participant attempted to produce three types of expression—happy, sad and surprised—with either low, medium or high intensity, resulting in nine unique expressions. Offline, for the articulatory-movement task, we trained a small feedforward neural-network model to learn the mapping between the ECoG features and each of the targets. We observed a median classification accuracy of 87.8% (99% CI [85.1, 90.5]; across $n = 10$ cross-validation folds) when classifying between the six articulatory 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) when classifying between the nine possible expressions and a median classification accuracy of 96.9% (99% CI [93.8,100]) when considering the classifier's outputs for only the strong-intensity versions of the three 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.

2.7 Articulatory Representations Drive Decoding

In healthy speakers, neural representations in the 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 centred over the SMC of our participant, we reasoned that articulatory representations persisting after paralysis drove speech-decoding performance. To assess this, we fitted 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, yielding a phonetic-tuning space in which each electrode had an associated vector of phone-encoding weights. Within this space, we determined whether phone clustering was organized by the primary orofacial articulator of each phone (place of articulation (POA), Figure 2.7), which has been shown in previous studies with healthy speakers (Chartier et al., 2018; Bouchard et al., 2013). We parcelled 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). 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). We visualized the phonetic tunings in a 2D space, revealing separability between labial and non-labial consonants and between lip-rounded and non-lip-rounded vowels.

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). As 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). 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 probably 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).

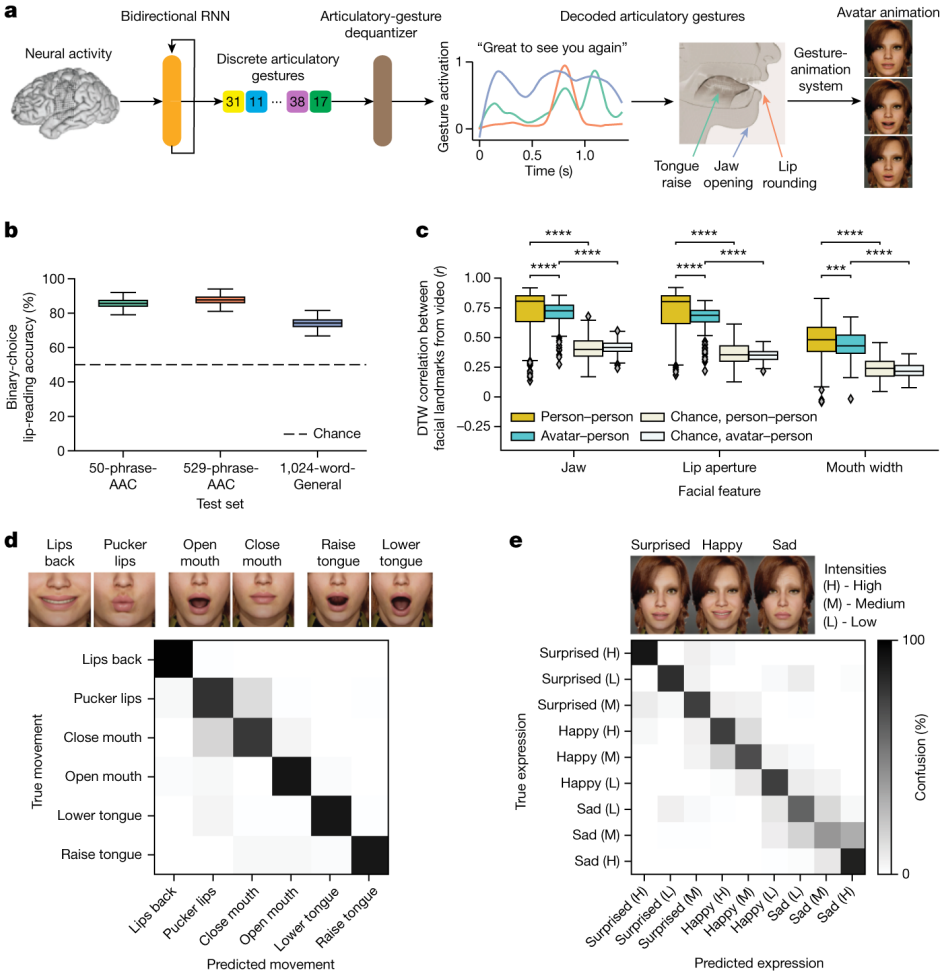

Figure 2.4: **Direct decoding of orofacial articulatory gestures from neural activity to drive an avatar.** **a**, Schematic diagram of the avatar-decoding algorithm. Offline, a bidirectional RNN decodes neural activity recorded during attempts to silently speak into discretized articulatory gestures. A dequantizer is then applied to generate the final predicted gestures, which are then passed through a pretrained 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. **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. **d**, Top: snapshots of avatar animations of six non-speech articulatory movements in the articulatory-movement task. Bottom: confusion matrix depicting classification accuracy across the movements. **e**, Top: snapshots of avatar animations of three non-speech emotional expressions in the emotional-expression task. Bottom: confusion matrix depicting classification accuracy across three intensity levels (high, medium and low) of the three expressions, ordered using a hierarchical agglomerative clustering on the confusion values.

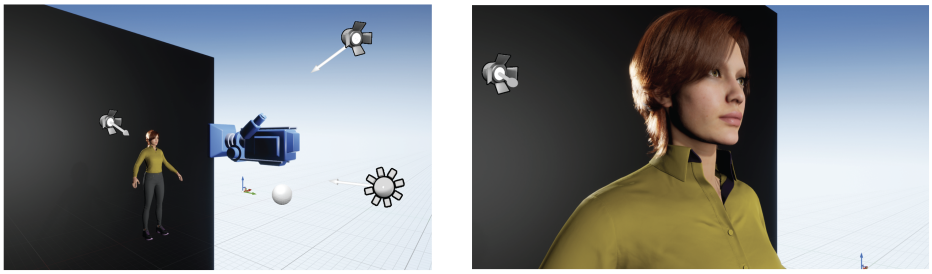

Figure 2.5: **Virtual environment for avatar decoding.** Virtual environment (designed in Unreal Engine 4.26) containing the camera and setup (left) for the "Vivian" MetaHuman's character (right).

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). 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). Although most electrodes were selective to either lip or tongue movements, others were activated by both. 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 2.4). 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 (Moses et al., 2021; Metzger et al., 2022; Simonyan et al., 2014) (Figure 2.8). 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 reasoned, many of the highest-contributing electrodes also exhibited substantial articulatory-feature encoding defined in 2.7 and were similarly important for all three modalities. 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.

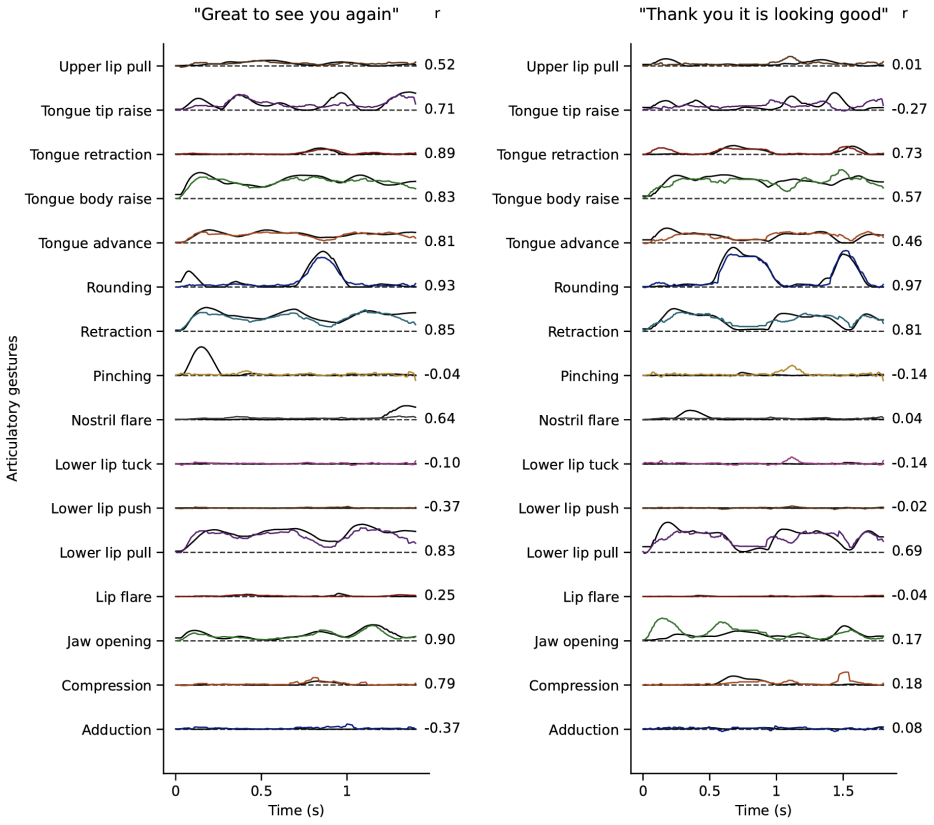

Figure 2.6: **Examples of directly decoded avatar articulatory gestures.** Examples of directly decoded articulatory gestures (colored) compared with reference articulatory gestures (black). Examples were taken from the 50-phrase-AAC sentence set. Dynamic time warping (Berndt & Clifford) was applied to align traces prior to plotting and computation of Pearson's r, which is displayed to the right of each gesture. Reference articulatory gestures were computed using the speech-to-gesture acoustic-to-articulatory inversion model from Speech Graphics' SG Com.

These results are in line with growing evidence for motor-movement encoding in the postcentral gyrus (Umeda et al., 2019; Murray & 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 & Coulter, 1981; Arce et al., 2013). We found that some temporal-lobe electrodes were not only active during passive listening but also contributed to silently attempted speech decoding ($r \geq 0.55$, $P < 0.0001$, Pearson correlation permutation test), suggesting that they may record cortical activity from the subcentral gyrus (Eichert et al., 2021) or sites with production activity within the temporal lobe (Binder, 2015).

2.8 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 system that decodes articulatory cortical activity into multiple output modalities in real time, 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 BCIs by a factor of 4 or more (Moses et al., 2021; Metzger et al., 2022; Willett et al., 2021) and expanding the vocabulary size of our previous direct-speech BCI by a factor of 20 (Moses et al., 2021). We also showed that intelligible speech can be synthesized from the brain activity of a person with paralysis. Finally, we introduced a 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. We believe that, together, 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: advances in the neural interface, providing denser and broader sampling of the distributed orofacial and vocal-tract representations across the lateral SMC; highly stable recordings from non-penetrating cortical-surface electrodes, enabling training and testing across days and weeks without requiring retraining on the day of testing; custom sequence-learning neural-decoding models, facilitating training without alignment of neural activity and output features; self-supervised learning-derived discrete speech units, serving as effective intermediate representations for intelligible speech synthesis; control of a virtual face from brain activity to accompany synthesized speech and convey facial expressions; and persistent articulatory encoding in

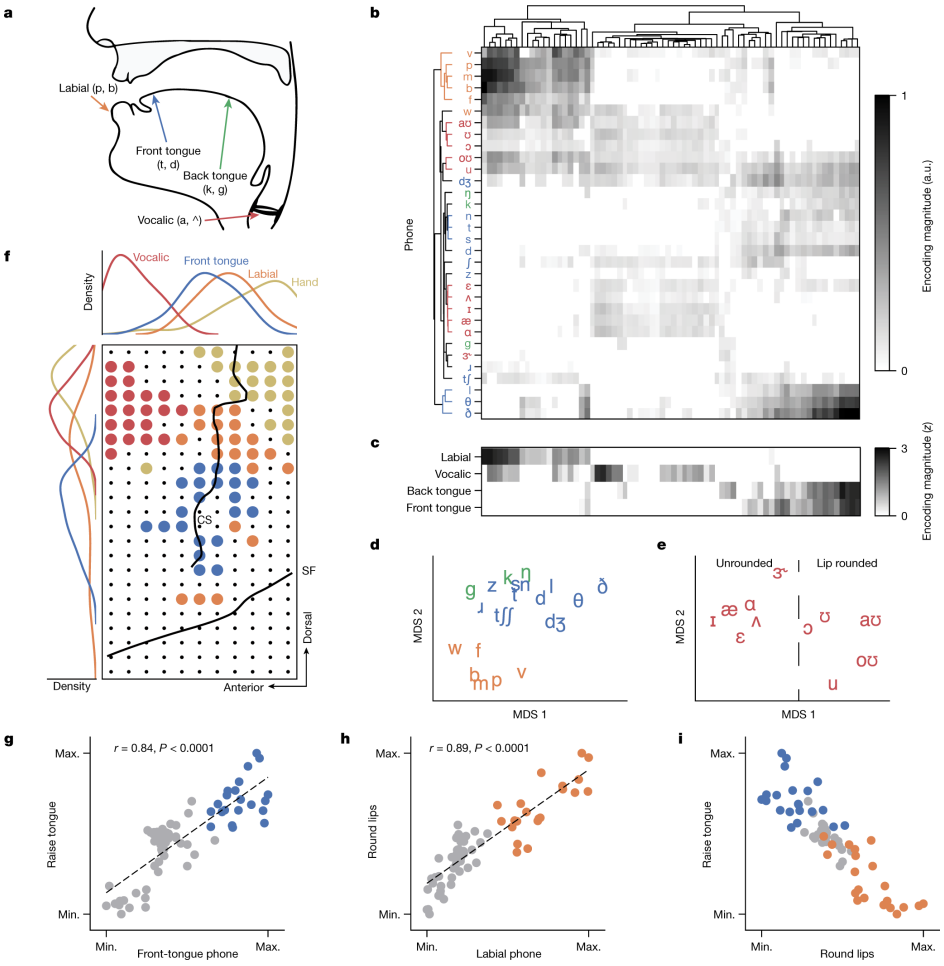

Figure 2.7: **Articulatory encodings driving speech decoding.** **a** Mid-sagittal schematic of the vocal tract with phone POA features labelled. **b**, Phone-encoding vectors for each electrode computed by a temporal receptive-field model on neural activity recorded during attempts to silently say sentences from the 1024-word-General set, organized by unsupervised hierarchical clustering. a.u., arbitrary units. **c**, z-scored POA encodings for each electrode, computed by averaging across positive phone encodings within each POA category. **d, e**, Projection of consonant (**d**) and vowel (**e**) phone encodings into a 2D space using 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 ECoG 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 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, labial phone encoding and lip-puckering attempts, and tongue-raising and lip-rounding attempts, respectively.

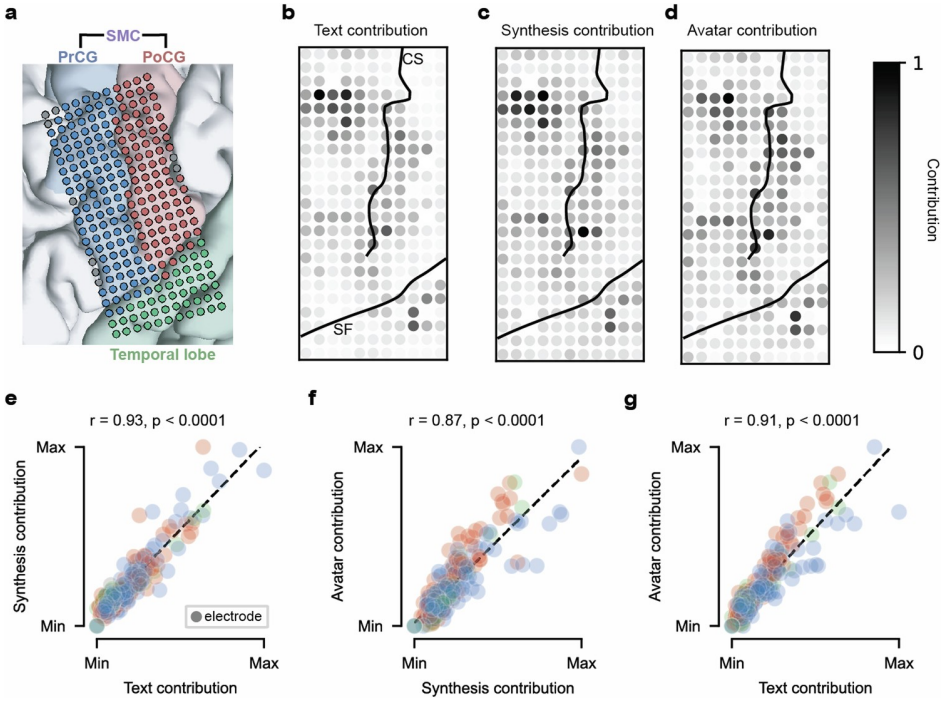

Figure 2.8: **Electrode contributions to decoding performance.** **a**, MRI reconstruction of the participant's brain overlaid with the locations of implanted electrodes. Cortical regions and electrodes are colored according to anatomical region (PoCG: postcentral gyrus, PrCG: precentral gyrus, SMC: 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.

the SMC of our participant that is consistent with previous intact-speech characterizations despite more than 18 years of anarthria, including hand and orofacial-motor somatotopy organized along a dorsal–ventral axis and phonetic tunings clustered by POA.

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). Additionally, providing instantaneous closed-loop feedback during decoding has the potential to improve user engagement, model performance and neural entrainment (Brumberg et al., 2018; Sadtler et al., 2014). Also, further advances in electrode interfaces (Chiang et al., 2020) 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.

2.9 Methods

Clinical-Trial Overview

This study was completed within the BCI Restoration of Arm and Voice clinical trial (ClinicalTrials.gov; NCT03698149). The primary endpoint of this trial is to assess the long-term safety and tolerability of an ECoG-based interface. All data presented here are part of the ongoing exploratory clinical trial and do not contribute towards 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 in which the details of study enrolment, including risks related to the study device, were thoroughly explained to her.

Participant

The participant, who was 47 years old at time of enrolment into the study, was diagnosed with quadriplegia and anarthria by neurologists and a speech–language pathologist. She experienced a pontine infarct in 2005, when she 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 enrolment evaluation, she scored 29/30 on the Mini Mental State Exam and was unable to achieve the final point only 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 model (Shi et al., 2022) 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 disc-shaped electrodes arranged in a lattice formation with 3-mm centre-to-centre 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. 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 HGA (Crone et al., 1998) and low-frequency signals (LFSs) 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 1 kHz after applying an anti-aliasing

filter with a cutoff of 500 Hz. Then we compute HGA as the analytic amplitude of these signals after band-passing them in the high-gamma frequency range (70–150 Hz), and then downsample them to 200 Hz. For LFSs, we apply only a low-pass anti-aliasing filter with a cutoff frequency of 100 Hz, and then downsample signals to 200 Hz. For data normalization, we applied a 30-s sliding-window z score in real time to the HGA and LFS features from each ECoG channel.

We carried out 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 previous work but have continued to augment and maintain over time (Moses et al., 2021; Metzger et al., 2022; 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 laboratory's server infrastructure, where we analysed 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 carry out 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, as we did 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 carry out 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 carry out 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 carry out 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 AAC (Beukelman et al., 1998). 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 film 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 & Loper, 2004) and the Cornell film corpus (Danescu-Niculescu-Mizil & Lee, 2011). We drew 18,284 sentences from these corpora that were composed entirely from the 1,152-word vocabulary from our previous work (Metzger et al., 2022), 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 and 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 collect 9,406 unique sentences (100 sentences were collected twice, for a total of 9,506 trials) 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 before real-time testing. As the synthesis model required several days to train to convergence, this model used only 6,449 trials for training data as 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 before 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 the 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 previous 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 (Moses et al., 2021; Metzger et al., 2022). We then normalized the HGA and LFSs separately to have an L2 norm of 1 across all time steps for each channel. We used all available electrodes during decoding.

We trained an RNN to model the probability of each phone at each time step, given these neural features. We trained the RNN using the CTC loss (Graves et al., 2006) 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 p z/ (the phonetic transcription of 'was') could be a result of the RNN predicting the following valid time series of phones: /w p z z/, /w w p p z/ z/, /w w p z/ and so forth.

We determined reference sequences using g2p-en (Park & 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 (Graves et al., 2013; Hannun et al., 2014). The convolutional portion of our RNN was composed of a 1D 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 41D 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-s 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-s window from a 9-s window of neural activity spanning from 1 s before to 8 s after the go cue, as in previous work (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 s after the go cue. To improve communication rates and decoding of variable-length sentences, we terminated trials before a full 8-s window if the decoder determined that 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 carried out the following steps: 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; if the RNN predicted the silence token for the most recent 8 time steps (960 ms) with higher than 88.8% average probability (or, in 2 out of the 249 real-time test trials, if the 7.5-s 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 in which the current decoded text was presented to the participant every 800 ms; however, the participant generally preferred seeing only 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 (Jurafsky & Martin, 2009) with Kneser-ney smoothing (Kneser & 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 before any pruning. The 5-gram language model is trained to predict the probability of each word in the vocabulary given the preceding words (up to 4). We chose this approach because the linguistic structure and content of conversational tweets and film lines are more relevant for everyday usage than formal written language commonly used in many standard speech-recognition databases (Panayotov et al., 2015; Ito & 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:

$$\text{rate} = \frac{n}{T}$$

where n is the number of words in the decoded output, and T is the time (in minutes) during which the 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 that immediately preceded the samples that triggered early stopping, giving the resulting formula:

$$\text{rate} = \frac{n}{t_{\text{silence detected}} - t_{\text{go cue}}}$$

Here, n remains the number of words in the decoded output, $t_{\text{silence detected}}$ is the time of the data sample that immediately preceded the samples that triggered early stopping, and $t_{\text{go cue}}$ is the time when the go cue appeared.

Error-rate Calculation

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. PER and CER are defined analogously for phones and characters, respectively. When measuring PERs, we ignored the silence token at the start of each sentence, as this token is always present at the start of both the reference phone sequence and the phone decoder's output.

For BCIs, error-rate distributions are typically assessed across sets of 5 or more sentences rather than single trials, as single-trial error rates can be noisy and are highly dependent on sentence length (Metzger et al., 2022; Moses et al., 2021; Willett et al., 2021). Hence, we sequentially parcelled 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; Moses et al., 2021), 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 in which 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 changed only the underlying 5-gram language model to be trained on all sentences 4 to 8 words in length in the Twitter and Cornell film 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 (1,024 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 that were present in both the Twitter and Cornell films corpora and the pronunciation dictionary. The most frequent words that 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 tested the text decoder in real time only 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 s 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 fitted 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 (Metzger et al., 2022) to output the probability of each of the 26 NATO code words and the 4 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. We computed NATO code-word classification accuracy using a model that was also capable of predicting the motor targets; here we measured performance only 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., 2021) from neural data. HuBERT is a speech-representation learning model that is trained to predict acoustic k-means-cluster identities corresponding to masked time points from unlabelled 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.

As our participant cannot speak, we generated reference sequences of speech units by applying HuBERT to a speech waveform that 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 before our participant's enrolment in the trial) who was instructed to read each sentence aloud in a consistent manner. Owing to the large number of sentences in the 1024-word-General set, we used the Wavenet text-to-speech model (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 h 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 (HGA and LFSs), consisted of the following layers (in order): a 1D convolutional layer, with 260 kernels with width and stride of 6; three layers of bidirectional gated recurrent units, each with a hidden dimension size of 260; and a 1D 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 (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 retained only the most likely predicted unit at each time step. We ignored time steps in which 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 pretrained unit-to-speech vocoder (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 s before to 4.62 s after the go cue for the 50-phrase-AAC and 529-phrase-AAC sentence sets and from 0 s before to 7.5 s 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 using 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. To reduce the latency of the personalized speech synthesizer during real-time inference for a qualitative demonstration, we trained a HiFi-CAR convolutional neural network (Wu et al., 2022) to vocode HuBERT units into personalized speech. This model used voice-converted LJSpeech (by means of YourTTS) as training data.

Evaluation

To evaluate the quality of the decoded speech, we computed the MCD between the decoded and reference waveforms ($\hat{y}$ and y, respectively) (Kubichek). This is defined as the squared error between dynamically time-warped sequences of mel cepstra (mc_d, in which 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:

$$\mathrm{MCD}(\hat{y}, y) = \frac{10}{\log(10)} \sqrt{\sum_{d=1}^{24} \left(\mathrm{mc}_d^y - \mathrm{mc}_d^{\hat{y}}\right)^2}$$

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 ten 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 carry out direct decoding of articulatory gestures from neural activity (the direct approach for avatar animation), we first trained a VQ-VAE to encode continuous Speech Graphics' gestures into discrete articulatory-gesture units (van den 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 1D convolutional units with 40 filters, a kernel size of 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 1D convolution, with 1 filter and a kernel size and stride of 1, to generate the predicted codebook embedding. We then used nearest-neighbour lookup to predict the discrete articulatory-gesture units. We used a codebook with 40 different 1D vectors, in which 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 1D codebook vector. Next we applied a 1D convolution layer, with 40 filters and a kernel size and stride of 1, to increase the dimensionality. Then, we applied 3 layers of 1D 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 1D 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 loss for the most important gestures more highly. We upweighted the jaw opening's mean-squared error loss by a factor of 20, and the gestures associated with important tongue movements (tongue-body raise, tongue advance, tongue retraction and 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 s after the go cue for the 1024-word-General set and from 0.5 s before to 5.5 s after for the 50-phrase-AAC and 529-phrase-AAC sets. The RNN then processed these neural features using the following components: a 1D convolution layer, with 256 filters with kernel size and stride of 2; three layers of gated recurrent units, each with a hidden dimension size of 512; and a dense layer, which produced a 41D 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.

During inference, the RNN yielded a predicted probability of each discretized articulatory-gesture unit every 60 ms. To transform these output probabilities into a sequence of discretized units, we retained only 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 carries out 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 carries out a forward mapping from articulatory-gesture activations to animation parameters on a 3D MetaHuman character created by Epic Games. The output animation was rendered using Unreal Engine 4.26 (noa, 2020).

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 (for which 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. As 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 (Ekman & Friesen, 2019), a well-known method for taxonomizing human facial movements. However, these articulatory gestures from Speech Graphics are more oriented towards 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 crowd-sourcing platform (Amazon Mechanical Turk) to design and conduct a perceptual assessment of the animations. Each decoded animation was assessed by six 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 (King) 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 eight 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 among 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 as 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. Before data collection, the participant viewed a video of an avatar carrying out the following six movements: open mouth, pucker lips, lips back (smiling or lip retraction), raise tongue, lower tongue and close mouth (rest or idle). Then, the participant carried out 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 s starting at the go cue.

To train and test the avatar-movement classifier, we used a window of neural activity spanning from 1 s before to 3 s after the go cue for each trial. We first downsampled the ECoG features (HGA and LFSs) 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 LFS and HGA features. 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 4,048D neural-feature vector (the product of 256 electrodes, 2 feature sets, 4 statistics and 2 data halves) for each trial. We then trained a multilayer 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 6D output vector. We then applied a softmax activation to get a probability for each of the six different gestures. We evaluated the network using tenfold 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 nine 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 s 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 (Metzger et al., 2022). We initialized the expression classifier with a pretrained NATO-and-hand-motor classifier (trained on 1,222 trials of NATO-motor task data collected before 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 fitted ten unique models to ensemble predictions on the held-out test set.

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 fitted 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 tenfold cross-validation using data recorded with the 1024-word General sentence set.

We first decimated the HGA by a factor of 24, reducing its sampling rate from 200 Hz to 8.33 Hz to match that of the phone-emission probabilities. Then, we fitted 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 one 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. An independent model was fitted for each electrode.

The true HGA at time t, denoted as $\mathrm{HGA}(t)$, was modeled as a weighted linear combination of phone-emission probabilities, indexed by p, from 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 $d_1 \ldots D$, with $d_1 = -4$ and $D = 4$.

During training, the squared error between the predicted HGA, $\mathrm{HGA}^*(t)$, and the true HGA, $\mathrm{HGA}(t)$, was minimized, using the following formulae:

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

$$\min_{w} \sum_{t} \left[\mathrm{HGA}^*(t) - \mathrm{HGA}(t)\right]^2$$

We implemented the model with the MNE toolbox's receptive-field ridge regression in Python (Gramfort et al., 2013). We used tenfold cross-validation to select the optimal alpha ridge-regression parameter by sweeping over the values $[1 \times 101, 1 \times 100, 1 \times 101, \ldots\ 1 \times 105]$, using 10% of our total data as a held-out tuning set. We then conducted another round of tenfold cross-validation on the remaining 90% of our total data to evaluate performance with the optimized alpha parameter. We averaged the coefficients for the model across the ten 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 (Müllner, 2011). We carried out clustering along both the electrode and phone dimensions.

To assess any relationships between phone encodings and articulatory features, we assigned each phone to a POA feature category, similar to what was done in previous work (Chartier et al., 2018; Bouchard et al., 2013). Specifically, each phone was primarily articulated at the lips (labial), the front tongue, the back tongue or the 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, with F=4 being 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 four 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 over the electrodes to visualize each phone in a 2D space. We implemented this using the scikit-learn Python package (Pedregosa et al., 2018).

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. 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 & 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 (event-related potentials (ERPs)) 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, 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 what was done in previous work (Cheung et al., 2016).

Statistical Analyses

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 were 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 in which we randomly sampled the distribution (for example, trials or pseudo-blocks) of interest with replacement 2,000 or 1,000 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 in which data were randomly shuffled 1,000 times. To compare success rates of decoding during our freeform demonstration with the main real-time evaluation, we used a t-test.

3

A Fast Streaming Brain-to-Voice Neuroprosthesis

3.1 Summary

Natural spoken communication happens instantaneously. Speech delays longer than a few seconds can disrupt the natural flow of conversation. This makes it difficult for individuals with paralysis to participate in meaningful dialogue, potentially leading to feelings of isolation and frustration. Here we used high-density surface recordings of the speech sensorimotor cortex in a clinical trial participant with severe paralysis and anarthria to drive a continuously streaming naturalistic speech synthesizer. Streaming speech synthesis refers to a speech synthesizer that does not wait until the end of attempted speech to produce audio but rather immediately begins producing audio in tandem with the user's attempt to speak. We designed and used deep learning recurrent neural network transducer models to achieve online large-vocabulary intelligible fluent speech synthesis personalized to the participant's preinjury voice with neural decoding in 80-ms increments. The models demonstrated implicit speech detection capabilities and could continuously decode speech indefinitely, enabling uninterrupted use of the decoder and further increasing speed. Our framework also successfully generalized to other silent-speech interfaces, including single-unit recordings and electromyography. Our findings introduce a speech-neuroprosthetic paradigm to restore naturalistic spoken communication to people with paralysis.

3.2 Main

Loss of spoken communication after neurological injury can be debilitating (Felgoise et al., 2016; Huggins et al., 2011; Branco et al., 2021); however, a neuroprosthesis that restores naturalistic and embodied communication from the individual's intent to speak would be transformative and substantially improve their quality of life (Peters et al., 2015; 2024; Silva et al., 2024a). The speech sensorimotor cortex has been shown to encode a rich array of articulatory and speech production-related information (Bouchard et al., 2013; Carey et al., 2017; Lotte et al., 2015; Dichter et al., 2018) that can be used to drive speech decoding in healthy speakers in the form of text (Herff & Schultz, 2016; Mugler et al., 2018; Makin et al., 2020; Sun et al., 2020) or synthesized speech (Herff et al., 2019; Anumanchipalli et al., 2019; Angrick et al., 2021; Wairagkar et al., 2023). Translating these findings into speech neuroprostheses that can restore fluent communication in people with severe vocal tract paralysis has been elusive (Silva et al., 2024a). This is because synthesizing intelligible speech without vocalization or coordinated articulation is a challenging goal. Individuals who have lost the ability to speak cannot produce an intelligible acoustic signal that can be used as the target during supervised model training, a leading paradigm used in current decoders. Previous demonstrations (Metzger et al., 2023; Moses et al., 2021; Willett et al., 2023b; Angrick et al., 2023; Luo et al., 2023; Metzger et al., 2022; Silva et al., 2024b) have shown that it is possible to decode intended speech from neural activity alone, with three recent publications expanding to large-vocabulary text decoding either complimented with delayed text-to-speech (TTS) synthesis (Willett et al., 2023b; Card et al., 2024) or delayed speech synthesis and accompanying orofacial movements (Metzger et al., 2023). However, the approaches to output audible speech process an entire window of neural activity corresponding to a speech attempt and wait until the end of the attempted speech production to synthesize speech directly (Metzger et al., 2023; Angrick et al., 2023) or apply TTS (Willett et al., 2023b; Card et al., 2024), causing extended periods of delays that scale linearly with the duration of the speech attempt. From this perspective, these approaches can be considered as running the inference offline and then subsequently playing the offline-generated speech waveform back to the participant. A naturalistic speech neuroprosthesis should operate in the same manner as natural spoken communication: as the participant tries to speak, speech is immediately synthesized from their neural activity. Furthermore, prior speech neuroprosthesis studies, including those decoding text followed by TTS, have depended on the participant's attempted overt vocalizations. The neural data generated from these attempts were used to train the decoding model (Willett et al., 2023b; Moses et al., 2021; Luo et al., 2023; Angrick et al., 2023; Silva et al., 2024b; Card et al., 2024), which may prove challenging to adapt to someone who has difficulty with or has lost this ability. Although our prior work circumvented this challenge (Metzger et al., 2023), it did not synchronously stream audible speech; instead, it relied on extracting the entire window of neural data before sending the data buffer to the decoder and outputting delayed synthesized speech.

Speech delays can hinder the flow of conversation, leading to misunderstandings or frustration for both the speaker and the listener and can even degrade the perceived quality of the speech (Schoenenberg et al., 2014; Krauss & Bricker, 1967). Therefore, having a long wait from intent to speak to the produced sound severely limits the natural engagement in communication and decreases the overall communication rate (Metzger et al., 2023; Willett et al., 2023b; Card et al., 2024). Low-latency spoken communication is crucial for maintaining the perceived quality of the conversation (Schoenenberg et al., 2014) and minimizing the frequency of confusion, for example, via interruptions, repetition of one's own speech or unnecessarily long silent periods (Brady, 1971). Hence, a practical speech neuroprosthesis must continuously synthesize speech from neural data in tandem with the user's attempt to speak. As persons with severe vocal tract paralysis may not be able to vocalize, speech brain–computer interfaces (BCIs) should not rely on audible vocalization at any point during training or inference. Additionally, speech BCI users should be able to speak using their decoding system continuously and volitionally after attempting to speak (Silva et al., 2024a). This capability is crucial as it mirrors the healthy speaker's ability to produce spontaneous speech, enabling speech BCI users to initiate spoken conversation or respond in real time without external prompts or cues. Last, provided that the device's input features encode articulation at a sufficient spatiotemporal resolution for speech decoding, the speech synthesis model architecture ideally should synthesize intelligible speech when trained on any such device (Mermelstein, 1973).

Toward this goal, we developed a 'streaming' speech neuroprosthesis that seamlessly converts short windows of neural activity to audible sound without waiting for an entire sentence to be attempted. We worked with a participant who could not speak or vocalize intelligible speech due to severe paralysis caused by a brainstem stroke (diagnosed with anarthria). Using a 253-channel high-density electrocorticography (ECoG) array, implanted largely over the surface of her speech sensorimotor cortex, we recorded neural activity while the participant silently attempted to say complete sentences constructed from a vocabulary of 1,024 words. In other words, the participant attempted to 'mime' or 'mouth' the target sentence without making any vocal sounds. We then designed and evaluated a bimodal speech decoding system that enabled large-vocabulary streaming speech synthesis and text decoding from neural activity during these silent-speech attempts. The decoders had significantly lower latency and faster communication rates than previous speech synthesis models, which waited until silent-speech attempts were completed before generating speech (Metzger et al., 2023). The speech synthesizer could generalize above chance to novel words not seen during training. Offline, our decoders demonstrated implicit speech detection capabilities and could operate continuously over several minutes, facilitating future speech BCI use outside a dedicated trial-level task structure. Although the speech was synthesized with a minimal delay as the participant silently attempted to speak, we show that the decoders were driven by the articulatory control in the motor cortex rather than model-generated auditory feedback, and performance remained stable. Finally, we demonstrate offline that our sequence modeling architectures generalize to multiple silent-speech interfaces, namely single-unit recordings from

another participant with paralysis (Willett et al., 2023a) and vocal tract surface electrodes measuring movement from a healthy speaker's silently mimed speech (Gaddy, 2020).

3.3 A Naturalistic Streaming Silent-Speech Decoding System

We designed a speech synthesis neuroprosthesis that enabled a clinical trial participant (ClinicalTrials.gov, NCT03698149) to naturalistically speak by synthesizing intended speech from neural signals acquired from a 253-channel ECoG array implanted over the surface of her speech sensorimotor cortex and a small portion of the temporal lobe (Figure 3.1). To train the system, we recorded neural data as the participant silently attempted to speak individual sentences. The participant was presented with a text prompt on a monitor and was asked to begin silently attempting to speak once a visual 'GO' cue turned green. During the online evaluation, the task design was the same. In addition, the synthesized speech was streamed through a nearby analog speaker, and decoded text was displayed on the monitor. The neural decoders in our system were bimodal in that they were jointly trained not only to synthesize speech but also to decode text simultaneously.

We streamed high gamma activity (HGA; 70–150 Hz) and low-frequency signals (LFS; 0.3–17 Hz; see Methods)(Metzger et al., 2022; Crone et al., 1998) to a custom bimodal decoding model, which processed the neural features in 80-ms increments starting 500 ms before the GO cue in each trial to decode audible speech and text. Inspired by streaming automatic speech recognition (ASR) approaches (Zhang et al., 2020; Shi et al., 2020; He et al., 2018), we used a recurrent neural network (RNN) transducer (RNN-T) framework (Graves, 2012), a flexible, general purpose neural network architecture that does not need future input context. We adapted this framework to facilitate streaming speech synthesis and text decoding from neural features. Specifically, a unidirectional RNN processes neural features in real time, producing an encoding vector corresponding to the speech content. For speech synthesis, these encodings were combined autoregressively with a streaming acoustic-speech unit language model to produce a probability distribution over the next acoustic-speech unit from 100 candidates. Similarly, for text, the same encodings were combined autoregressively with a streaming subword text-encoding language model to produce a probability distribution over the next subword text encoding from 4,096 candidates. For acoustic-speech units and text encodings, RNN-T beam search was used to determine the most likely token during inference (Graves, 2012). The predicted acoustic-speech unit was passed as input into a personalized speech synthesizer to generate a waveform chunk played synchronously as the participant attempted to speak.

To overcome limitations in aligning neural data to speech behavior due to the lack of intelligible speech from the participant, we used the RNN-T loss function during training. Notably,

the RNN-T loss models not only the probability of the output acoustic-speech units or text encodings but also their interdependence. This allows for the learning of a streaming language model over the acoustic-speech units and text encodings without requiring an external language model. We trained the streaming language model portions of our architecture offline on speech recognition tasks and froze them before training the rest of the pipeline. The target acoustic-speech units were extracted using HuBERT, a self-supervised speech representation learning model that encodes a speech waveform into a temporal sequence of units that capture the speech waveform's underlying phonetic and articulatory characteristics (Hsu et al., 2021; Cho et al., 2023). Because our participant cannot speak, we generated the initial reference waveforms using a TTS model (Casanova et al., 2022). No audible vocalization was required to train the speech decoders. Finally, for the speech synthesizer, we trained a personalized autoregressive acoustic-speech unit synthesizer, which models the duration of the acoustic-speech units to better match the participant's speaking rate. The synthesized speech was conditioned on a short voice clip of the participant recorded before she lost her speech ability. Figure 3.2 provides a more detailed schematic of the streaming speech neuroprosthesis's inference and training paradigm.

We evaluated the system using a small-vocabulary sentence set '50-phrase-AAC' and an extensive-vocabulary sentence set, '1,024-word-General'. The 50-phrase-AAC set was designed as a predefined phrase set to express primary caregiving needs. By contrast, the 1,024-word-General set was designed as a large-vocabulary, large-sentence set containing 12,379 unique sentences composed of 1,024 unique words sampled from Twitter and movie transcriptions (Metzger et al., 2023). The participant completed nearly two full passes through the corpus during training, resulting in 23,378 total silent-speech attempts. Each sentence was seen at least twice during training, and a subset of the sentences was also collected multiple times, resulting in the model seeing each test sentence an average of 6.94 times during training. Additionally, to test the generalizability of the neural decoder, we evaluated performance on novel sentences composed of words within the vocabulary but never seen by the participant as well as evaluated performance on novel words outside of the 1,024 word vocabulary.

As an alternative to speech synthesis via acoustic-speech units, we also implemented a version of the pipeline that performs text decoding followed by word-by-word TTS for the 1,024-word-General sentence set. Here, we used the text-decoding portion of the same model to predict the next text segment, and this was then used to condition a TTS model that synthesized speech for that segment. This offers higher intelligibility at the cost of naturalism and could be adapted to any text-decoding algorithm (Metzger et al., 2023; Willett et al., 2023b; Card et al., 2024), provided that the language model and decoder are streaming and causal.

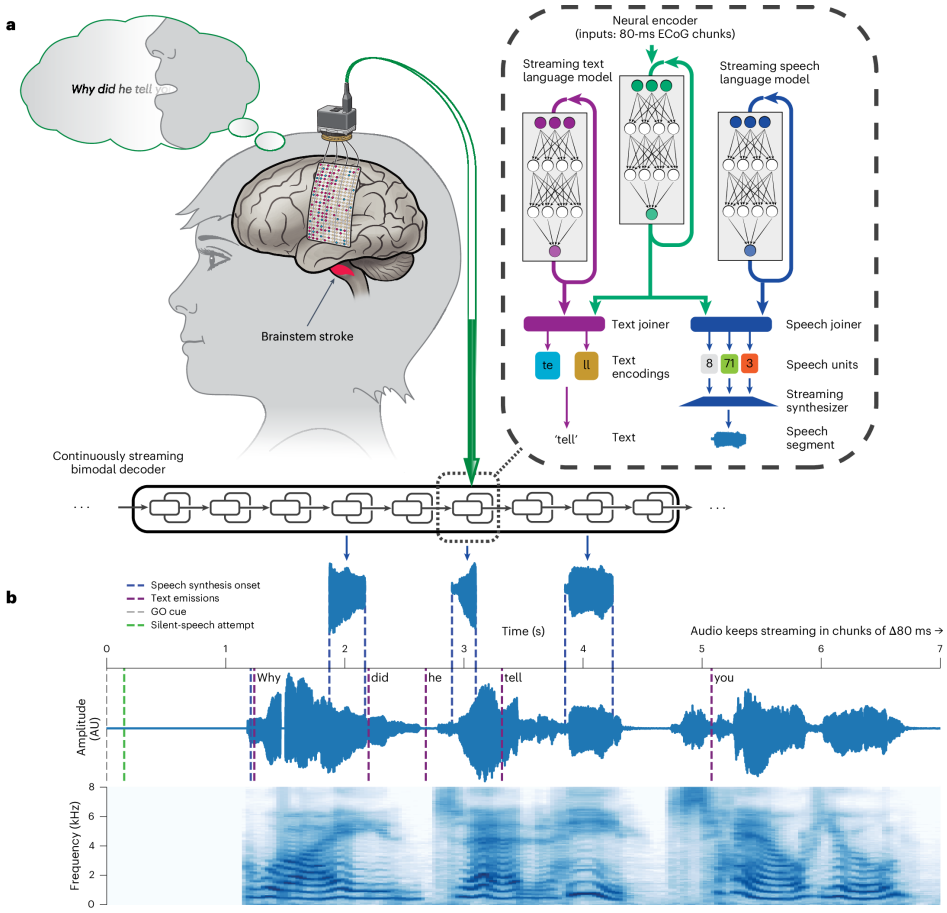

Figure 3.1: **Overview of a naturalistic streaming silent-speech neuroprosthesis. a** Overview of the streaming speech synthesis and text-decoding pipeline. A person with severe paralysis due to a brainstem stroke was implanted with a 253-channel ECoG array 18 years after injury. Deep learning models were trained to map neural activity during silently attempted speech to personalized speech and text in increments of 80 ms. For speech synthesis, acoustic-speech units are decoded and then synthesized into speech. For text, subword text encodings are predicted and then dequantized into words. For both outputs, a streaming language model takes in the previous prediction in parallel with the neural encoder inference to allow for language modeling during streaming decoding. **b**, An exemplar online waveform (top) and spectrogram (bottom) from the 1,024-word-General set. The detected GO cue and speech attempt are demarcated in black and green, respectively. Timings for text emissions are demarcated in purple. All elements are temporally to scale; AU, arbitrary units.

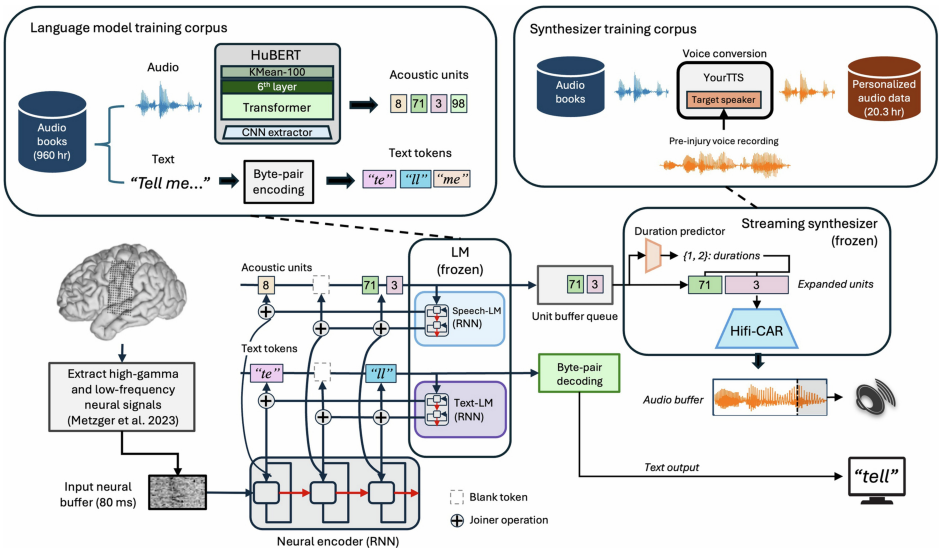

Figure 3.2: **Schematic of streaming silent-speech neuroprosthesis architecture and training**, (Top) For training, A large speech corpus is used to generate reference text and speech audio. A large acoustic language model (HuBERT) encodes the audio waveforms into acoustic-speech units to be used as targets during training. Similarly, a byte-pair encoding model encodes the sentences into word and subword units. For the speech-synthesizer component, a voice conversion module (YourTTS) is used to convert the audio into personalized audio conditioned on a short clip from the participant recorded in passing before her loss of speech. (Bottom) High-gamma and low-frequency neural signals are extracted during inference, and 80-ms chunks are fed into the multimodal neural encoder. The language model predictions are joined with the neural encoder outputs. For acoustic-speech units, a duration predictor predicts the acoustic-speech unit duration and duplicates the unit accordingly. The speech synthesizer then vocodes the most recent predicted acoustic-speech unit into an 80 ms audio chunk.

3.4 Fast Streaming Intelligible Speech Synthesis

To evaluate online performance (Figure 3.3), we synthesized speech and displayed decoded text as the participant silently attempted to say 100 different sentences from the 1,024-word-General set and 150 trials (50 phrases, three times each) from the 50-phrase-AAC set.

The system's decoding speed was a median of 47.5 words per minute (WPM; 99% confidence interval (CI) 45.2, 49.3) and 90.9 WPM (99% CI 88.4, 95.4) for the 1,024-word-General and 50-phrase-AAC sets, respectively, and was significantly faster than the previous decoding approach ($P < 0.0001$ for both comparisons, two-sided Wilcoxon rank-sum tests; the previous

approach was 28.3 WPM (99% CI 27.5, 33.4) and 52.9 WPM (99% CI 50.0, 54.2), respectively) (Metzger et al., 2023). The 50-phrase-AAC sentence set had a much higher median WPM because the participant's rate of silently attempted speech was, by design, faster for those sentences. We observed minimal delay speech synthesis and text decoding as measured from the time between the speech attempt and the onset of decoding output. We used a speech detection algorithm to predict the start of the attempted speech from the participant's neural activity (see Methods) (Metzger et al., 2022; 2023; Moses et al., 2021). For the 1,024-word-General set, we achieved median latencies of 1.12 s (99% CI 1.03, 1.26) for speech synthesis and 1.01 s (99% CI 0.90, 1.13) for text decoding, respectively. For the 50-phrase-AAC set, we achieved median latencies of 2.14 s (99% CI 2.05, 2.22) for speech synthesis and 2.35 s (99% CI 2.28, 2.43) for text decoding, respectively. When measuring the latency between the GO cue and the decoding onsets, for the 1,024-word-General set, we observed median latencies of 1.67 s (99% CI 1.57, 1.81) and 1.56 s (99% CI 1.4, 1.56) for speech synthesis and text decoding, respectively, increased due to the participant's reaction time and the time between the beginning of articulation and intended acoustics. Similarly, for cue-based latency measurements, for the 50-phrase-AAC set, we achieved median latencies of 2.61 s (99% CI 2.53, 2.70) and 2.90 s (99% CI 2.70, 2.90) for speech synthesis and text decoding, respectively. These latencies are faster than our participant's typical communication latency using her previous assistive communication device (23.2 s) and are lower latency than the previous state-of-the-art speech synthesizer (delayed synthesis) (Metzger et al., 2023). We observed that no trials emitted speech or text outputs before the GO cue during the preparation phase of speech production, indicating that the model does not prematurely emit decodings during real-time inference. Offline, we characterized the per-80-ms-chunk inference latency of the 1,024-word-General sentence set model and observed it to be a median of 11.83 ms (99% CI 11.82, 11.84), with the language model subcomponents contributing the most delay. The decoder had over a 99% success rate in achieving sub-80-ms inference latencies. This metric represents the latency achievable for each step of speech synthesis, which constrains how natural and fluent sounding the streaming synthesis can be. However, latencies measured from the detected speech attempt and generated audio are much higher, indicating the need for further metrics to distinguish the computational latency from the online decoding latency. The speech synthesis latency determines the functional speed (that is, the system's WPM relative to the participant's speaking rate). By contrast, the short inference latency enables naturalistic continuous speech synthesis without phrase-level or word-level pauses.

For evaluation of the content of the decoded outputs, we used phoneme error rate (PER), word error rate (WER) and character error rate (CER), which are standard in ASR and measure the percentage of incorrect phonemes, words and characters, respectively. We computed error rates across sequential pseudoblocks of ten-sentence segments. For speech synthesis, transcripts were obtained from perceptual evaluations performed by third-party evaluators. We achieved median speech PERs of 10.8% (99% CI 4.21, 15.9) and 45.3% (99% CI 40.7, 64.4) for the 50-phrase-AAC and 1,024-word-General sets, respectively. For text, we achieved PERs of 7.58% (99% CI 0.0, 9.27) and 23.9% (99% CI 18.4, 31.6) for the 50-phrase-AAC

and 1,024-word-General sets, respectively. Additionally, we achieved median speech WERs of 12.3% (99% CI 5.26, 23.1) and 58.8% (99% CI 50.5, 76.0) for the 50-phrase-AAC and 1,024-word-General sets, respectively. For text, we achieved WERs of 10.3% (99% CI 0.0, 13.8) and 31.9% (99% CI 27.5, 42.8) for the 50-phrase-AAC and 1,024-word-General sets, respectively. Similarly, we achieved median speech CERs of 11.2% (99% CI 4.67, 15.0) and 44.7% (99% CI 39.2, 63.3) for the 50-phrase-AAC and 1,024-word-General sets, respectively. For text, we achieved CERs of 7.23% (99% CI 0.0, 9.72) and 22.8% (99% CI 18.4, 31.0) for the 50-phrase-AAC and 1,024-word-General sets, respectively.

To characterize the streaming model's ability to synthesize words outside of the 1,024-word training vocabulary, we applied our 1,024-word-General model to silent-speech attempts of 24 isolated unseen words (for example, 'Zulu', 'Romeo', 'Quebec' and so on). We applied the decoder to individual word-level attempts to avoid any bias in the synthesized words that could arise from the language model's prior acoustic-speech context from words in the 1,024-word vocabulary. We were able to achieve a median of 46.0% classification accuracy (99% CI 37.0, 55.4, and above chance (3.85%), computed via re-evaluating performance after shuffling the channels of the neural data and using this as the input to the model; $P < 0.001$, two-sided Wilcoxon rank-sum test) over the spectral features of the decoded waveforms, with clustering of the acoustics generally being by place of articulation, and many waveforms resembled their respective unseen word class (Figure 3.4.

To visualize the acoustic unit decoding process offline, we plotted the probability of a given acoustic unit sequence across time as the decoder was applied to the input neural data from a 1,024-word-General set trial from the real-time evaluation set. After a brief period of silence, we see that the synthesis decoding model emits speech units relatively continuously. By contrast, the text byte pair encodings were decoded stepwise at the word level. For 50-phrase-AAC, speech unit decoding was more similar to stepwise text decoding, and latencies were generally higher, which may be explained by the language model waiting to accumulate more neural context because of the strong bias during training to produce 1 of 50 phrases. To better characterize the synchronization between the text and speech onsets, we observed a median absolute timing difference between text onset and speech onset of 185 ms (99% CI 170, 210) and 170 ms (99% CI 130, 210) for the 1,024-word-General and 50-phrase-AAC sets, respectively. This is important because text and speech synthesis are jointly modeled, whereas all previous approaches model the outputs independently or are not multimodal at all and, hence, do not provide actual conditioning.

For all error rate metrics, performance was better than chance, which we computed by re-evaluating performance after shuffling the channels of the neural data and using this as the input to our decoding models ($P < 0.01$ for all 12 comparisons, two-sided Wilcoxon rank-sum tests with six-way Holm–Bonferroni correction). To evaluate the contribution of the language model, we re-evaluated performance when providing the language models with unit contexts generated by uniformly sampling tokens from the training corpus distribution. For all real-time error rate metrics, performance did not exceed chance ($P > 0.05$ for all 12 comparisons, two-

sided Wilcoxon rank-sum tests with six-way Holm–Bonferroni correction). In contrast to prior works (Metzger et al., 2023; Willett et al., 2023b; Moses et al., 2021; Metzger et al., 2022; Silva et al., 2024b; Card et al., 2024), the training procedure models both the input neural data and the target sequences (Graves, 2012), making both components essential for accurate streaming transducer-based decoding. Together, these results demonstrate naturalistic streaming speech synthesis from neural activity that produces low-latency intelligible speech and greatly improves communication speed compared to synthesizing speech after the intended behavior. The audio and text were tightly synchronized via a shared streamable bimodal neural encoder, and the streaming synthesis was enabled via an RNN-T architecture for streaming speech synthesis.

3.5 Long-Form Continuous Speech Decoding and Implicit Detection

Ideal speech neuroprostheses should be able to operate continuously using a unified model that is aware of when the participant is speaking or not. They should not be confined to isolated trial-level attempts, enabling future speech BCI users to use the device outside a dedicated task structure. Although we demonstrate online continuous speech synthesis at the trial level, the speech synthesizer should ideally generalize to longer-form decoding (that is, over several minutes or hours rather than several seconds). Furthermore, previous approaches to speech detection operated using a separate model, which resulted in waiting until the end of the speech attempt to decode speech (Metzger et al., 2023; Moses et al., 2021; Metzger et al., 2022; Luo et al., 2023; Angrick et al., 2023), hence not demonstrating continuous low-latency speech decoding. To address this, offline, we applied our 1,024-word-General sentence set model to four entire blocks of neural activity consisting of 25 trials each. Specifically, instead of passing in neural data 500 ms before the trial and waiting for a fixed period to allow for trial-level decoding, we autoregressively passed the entire block of neural activity to the model in nonoverlapping chunks of 80 ms (Figure 3.5).

We observed that the RNN-T model offers an implicit speech detection mechanism that allows it to operate continuously over periods beyond single trials. Once the RNN-T is confident enough that speech is occurring, it begins synthesizing speech. Likewise, it rarely produces speech when speech is not implicitly detected. We observed only 3 false-positive speech synthesis segments out of 100 in which the participant had not attempted speech. This demonstrates that the system correlates with volitional silent-speech attempts and rarely produces speech when not desired. To test whether the system is robust to unintended activations (for example, those that may occur during rest), we applied our model to 16 minutes of rest data, during which the participant rested and did nothing, aggregated across ten sessions and found that the system never falsely decoded speech. To characterize this

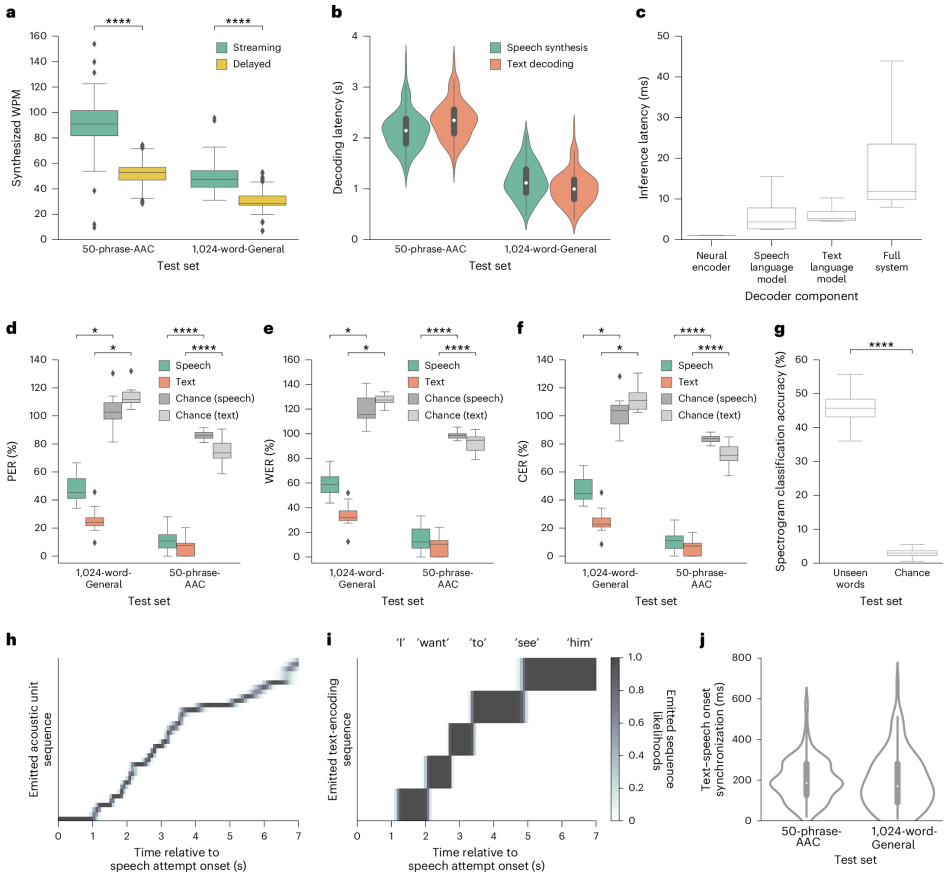

Figure 3.3: **Online continuously streaming synchronized speech synthesis and text decoding from neural activity.** **a** Synthesized WPM compared to delayed synthesis (Metzger et al., 2023). **b**, Latency for speech synthesis and text decoding. Latency is defined as the time from the detected start of attempted speech to decoding output onset. **c**, Inference latency for each 80-ms chunk of input neural data. **d**, PERs. **e**, WERs. **f**, CERs. Decoded speech synthesis transcripts in **d-f** were obtained from untrained human evaluators via a transcription task. Data were analyzed by two-sided Wilcoxon signed-rank test with a six-way Holm–Bonferroni correction for multiple comparisons. In **a** and **d-g**, ****P $<$ 0.0001 and *P $<$ 0.05. **g**, Offline synthesis spectrogram classification performance when applying the 1,024-word-General sentence set model to silent-speech attempts of 24 unseen words (n = 1,000 bootstrapped points). **h**, For an example trial, the likelihoods of emitting acoustic-speech unit sequences are shown time aligned with the detected speech onset, illustrating how the model selects the most probable speech unit sequences over time based on the input neural data. **i**, Same as **h** but for text encodings. The decoded word is at the top time aligned to when it is most probable. **j**, Synchronization time between the speech synthesis onset and text decoding onset. In **a**, **b**, **d-f**, and **j**, the results were obtained online. In **c**, **h**, and **i**, the data were simulated offline on real-time test blocks.

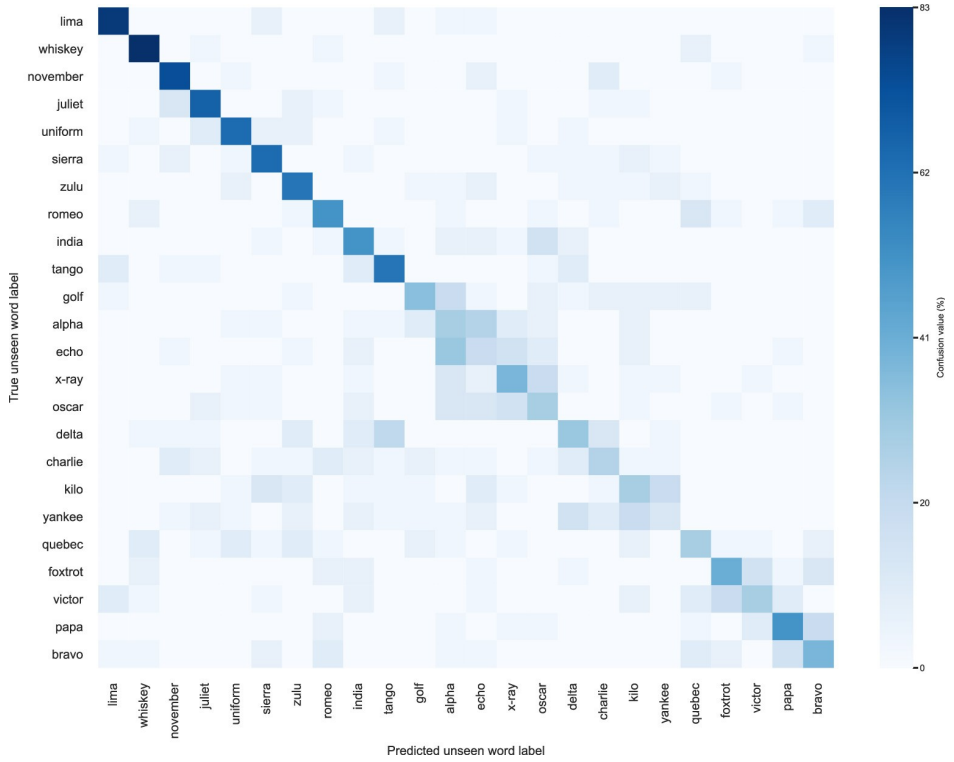

Figure 3.4: **Speech synthesis performance on unseen words.**, The confusion matrix presents the hierarchically clustered results of a 20-fold cross-validated classification using a random forest classifier on spectrograms extracted from decoded waveforms corresponding to 744 trials across 25 blocks of the participant miming one of 26 unseen words from the NATO phonetic alphabet (for example, 'Alpha,' 'Bravo,' 'Charlie,' etc.). The task aimed to test the generalizability of our 1024-word speech synthesis model to unseen words, preventing the language model from using context by isolating the words. The model achieved a median accuracy of 46.0% (99% CI [37.0, 55.4]; bootstrapped over 1,000 iterations), which was significantly above the 3.85% chance level ($P < 0.001$, two-sided Wilcoxon rank-sum test). Hierarchical clustering was applied, with darker cells along the diagonal indicating higher classification accuracy. All results were computed offline. The words 'Hotel' and 'Mike' were excluded to ensure a fair generalization analysis as they were part of the training set.

further, we computed the latency (absolute difference) between the speech synthesizer outputs and the detected onsets and offset for silently attempted speech (computed using a separate speech detection model). We observed median latencies of 1.19 s (99% CI 1.09, 1.29) for onsets and 1.89 s (99% CI 1.70, 2.04) for offsets. The RNN-T training process and beam search algorithm enable this long-form decoding functionality, which decides whether or not and how many acoustic-speech units and text encodings to emit, resulting in natural implicit speech detection.

To evaluate the model's ability to decode unprompted silent speech in freeform settings, offline, we applied our continuous long-form decoding system to four blocks of unstructured silent speech, where the participant attempted to speak any sentences she desired. In contrast to the delayed speech synthesis approach, we observed significant improvements in speech synthesis onset latency, offset latency and WPM ($P < 0.0001$, two-sided Wilcoxon signed-rank).

We also computed error rates between the predicted and target speech and text during long-form decoding. For speech, we achieved median PERs of 49.4% (99% CI 44.5, 56.7), median WERs of 65.0% (99% CI 60.3, 73.1) and median CERs of 49.3% (99% CI 43.9, 55.2). For text, we achieved median PERs of 34.7% (99% CI 26.5, 40.7), median WERs of 44.2% (99% CI 33.1, 51.5) and median CERs of 34.1% (99% CI 25.9, 39.2). All performances were above chance ($P < 0.0001$ for all six comparisons, two-sided Wilcoxon rank-sum tests with six-way Holm–Bonferroni correction computed over median-bootstrapped block metrics with 1,000 iterations sampled with replacement). These results demonstrate continuous speech synthesis beyond isolated trial-level attempts and offer promise for the future development of clinical user interfaces for continuous speech decoding.

3.6 Generalization across Silent-Speech Interfaces

Provided that the spatiotemporal resolution and coverage of the neural recording device is sufficient, decoding approaches for speech neuroprostheses should be able to generalize beyond a single recording modality or participant (Chiang et al., 2020). Furthermore, any such silent-speech recording interface with an articulatory basis should be able to be used as a synthesizer for intelligible speech (Mermelstein, 1973). To this end, we applied our speech decoding approach to three separate silent-speech datasets (Figure 3.6: a previous 1,024-word vocabulary ECoG corpus from our participant (Metzger et al., 2022), an open-vocabulary microelectrode array (MEA) dataset from a person with paralysis implanted with an intracortical BCI (Willett et al., 2023b) and an open-vocabulary articulatory electromyography (EMG) dataset from a single healthy speaker with surface electrodes placed along their vocal tract (Gaddy, 2020). Open vocabulary means there was no restriction on the vocabulary during training and testing. The EMG dataset was included as a baseline for silent-speech synthesis because healthy speakers can make the full range of articulatory

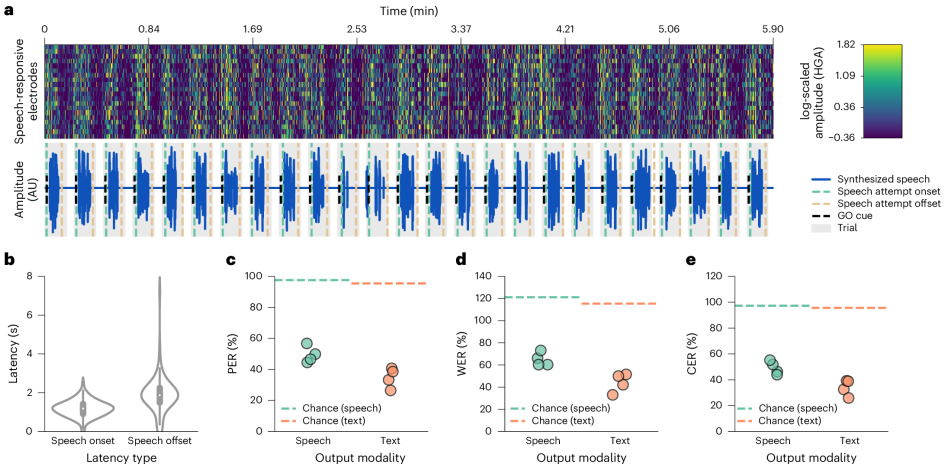

Figure 3.5: **Offline long-form continuous speech decoding with implicit speech detection.** **a**, Top: heat map of log-scaled HGA from the top 20 most speech-responsive electrodes during silent-speech attempts of an entire block (5.9 min) of 1,024-word-General sentences. Lighter indicates increased neural activity. Bottom: continuously synthesized speech waveform from the aforementioned neural activity. Neural data are passed into the model in 80-ms chunks and synthesized continuously in 80-ms chunks. The GO cue detected silent-speech attempt onsets, and detected silent-speech attempt offsets are marked in black, green and purple, respectively. Gray indicates single trials. Specifically, the decoder had access to the original time region of neural data collected during online inference. **b**, Latency between the detected onset of silently attempted speech to synthesized (n = 98 trials) speech onset and latency between the detected offset of silently attempted speech to synthesized speech offset. Each point in the distribution is a single trial. The white dot represents the median, the box spans the interquartile range (25th to 75th percentiles), whiskers extend to ±1.5 times the interquartile range, and the violin width illustrates data density at each point on the y axis. **c**, PERs. **d**, WERs. **e**, CERs. In **c-e**, predicted speech transcripts were obtained via ASR. Chance is computed by shuffling the electrodes and applying the decoder. Each dot indicates performance computed via continuous speech synthesis and text decoding over an entire block of attempted speech.

movements. Note that although each recording modality is a proxy for articulatory behavior during silent-speech attempts, the evaluation sentences, device implant, amount of training data, participant behavior, and participant etiologies differ. Hence, this analysis helps us to understand the generalization of the proposed sequence modeling framework but, for the stated reasons, we do not directly compare the performances of the three recording interfaces.

For each of the three datasets, we tested on data recorded during silent-speech attempts, used no vocalized audio from the participants during the training of the models and tested on unseen sentences. For MEA and EMG, we did not personalize the speech synthesizer voice. We used recording modality-specific feature extraction layers before applying the RNN in the encoder. Otherwise, the architecture was the same as demonstrated for online decoding. For ECoG, we observed a median PER of 49.0% (99% CI 44.2, 57.9), a median WER of 63.8% (99% CI 57.5, 73.9) and a median CER of 51.3% (99% CI 46.1, 58.9). For MEA, we observed a median PER of 54.5% (99% CI 48.3, 60.9), a median WER of 79.7% (99% CI 70.5, 84.0) and a median CER of 55.0% (99% CI 48.6, 61.7). For EMG, we observed a median PER of 44.8% (99% CI 43.3, 49.9), a median WER of 73.0% (99% CI 69.0, 80.9) and a median CER of 44.3% (99% CI 42.5, 49.9).

To assess the effects of using the entire window of neural activity on performance, we retrained our speech synthesis models but did not enforce a streaming constraint. We used no 80-ms streaming buffer for all three models, allowing the model to produce delayed synthesis at the expense of latency. For ECoG and MEA, we used bidirectional neural encoders. We used a unidirectional neural encoder for EMG. For ECoG, we observed a median PER of 38.7% (99% CI 34.4, 46.2), a median WER of 55.4% (99% CI 49.1, 60.8) and a median CER of 43.7% (99% CI 35.6, 49.0). For MEA, we observed a median PER of 39.1% (99% CI 31.9, 44.9), a median WER of 57.9% (99% CI 50.0, 66.7) and a median CER of 40.8% (99% CI 33.2, 44.2). For EMG, we observed a median PER of 37.9% (99% CI 31.4, 42.0), a median WER of 59.7% (99% CI 53.3, 64.2) and a median CER of 38.5% (99% CI 33.1, 42.0).

We also applied a connectionist temporal classification-based approach described in our previous work, which requires a full window of neural activity, to each modality (Metzger et al., 2023; Graves et al., 2006), achieving intelligible but delayed speech synthesis. For ECoG, we observed a median PER of 35.7% (99% CI 31.9, 40.8), a median WER of 48.4% (99% CI 41.2, 58.7) and a median CER of 36.5% (99% CI 31.6, 41.1). For MEA, we observed a median PER of 29.9% (99% CI 22.2, 32.4), a median WER of 40.9% (99% CI 38.3, 46.6) and a median CER of 30.1% (99% CI 24.0, 32.9). For EMG, we observed a median PER of 17.5% (99% CI 14.1, 20.4), a median WER of 32.0% (99% CI 26.7, 34.9) and a median CER of 16.8% (99% CI 13.9, 21.3).

We observed significantly above-chance performance for each modality and error rate metric ($P < 0.01$ for all 27 comparisons, two-sided Wilcoxon rank-sum tests with a nine-way Holm–Bonferroni correction computed over median-bootstrapped block metrics with 1,000 iterations sampled with replacement). Together, these results show that our streaming and

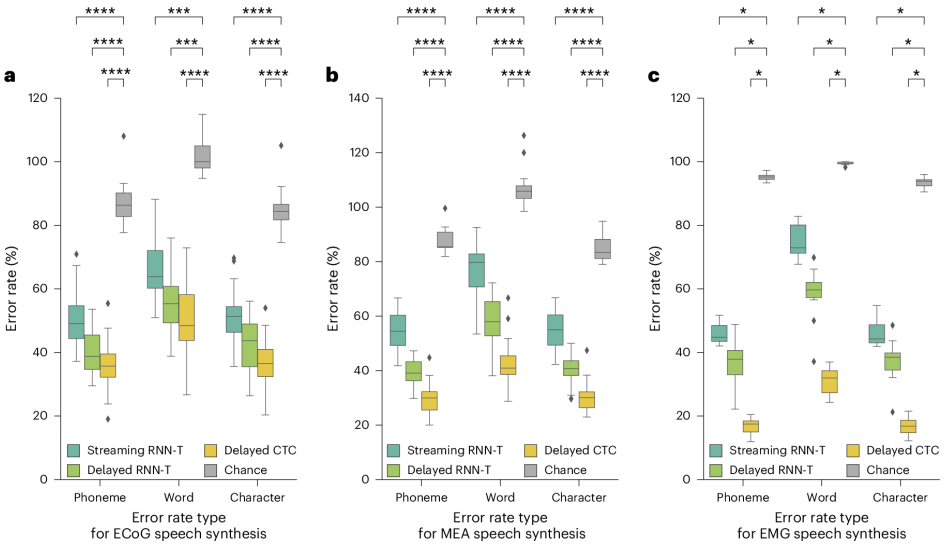

Figure 3.6: **Speech synthesis generalization across silent-speech interfaces. a**, ECoG speech synthesis error rates (1,024-word vocabulary) (Metzger et al., 2023) **b**, MEA speech synthesis error rates (open vocabulary) (Willett et al., 2023a) **c**, Surface EMG speech synthesis error rates (open vocabulary) (Gaddy, 2020) In **a-c**, to evaluate the generalizability of our approach, we retrained and tested a streaming RNN-T model, an RNN-T model without a streaming buffer constraint (delayed) and a connectionist temporal classification model (also delayed). Predicted speech transcripts were obtained via ASR. Chance distributions were generated by shuffling the electrode locations and applying the best-performing decoder ($n = 20$, $n = 12$ and $n = 10$ pseudoblocks for ECoG, MEA and EMG, respectively); *$P <$ 0.05, ***$P < 0.001$ and ****$P < 0.0001$ (two-sided Wilcoxon signed-rank test with nine-way Holm–Bonferroni correction for multiple comparisons); $P = 0.0000172$, $P = 0.000275$ and $P = 0.01758$ for all ECoG, MEA and EMG comparisons, respectively. All box plots depict the median (horizontal line inside box), 25th and 75th percentiles (box), 25th and 75th percentiles ±1.5 times the interquartile range (whiskers) and outliers (diamonds). All models were tested offline; CTC, connectionist temporal classification.

delayed model architectures generalize across silent-speech interfaces and are not limited to ECoG. Error rate metrics decrease with increased delay, at the expense of streaming, speed and latency. We expect the performance of these approaches to improve and further generalize as the density or number of electrodes increases when recording from the speech sensorimotor cortex (Card et al., 2024; Chiang et al., 2020; Duraivel et al., 2022).

3.7 Speech Decoding is Robust to Synthesized Auditory Feedback

As the decoder was trained without auditory feedback, one could hypothesize that, during inference where auditory feedback is present, neural activity patterns could differ, especially in areas responsive to listening (Chang et al., 2013; Ozker et al., 2023), which could disrupt the model's inference. To assess this, we compared the electrode contributions and decoding performance on trials with and without naturalistic auditory feedback using synthesized speech. We constrained the prompted texts to be consistent across conditions, leaving 50 trials for each condition. Both conditions used the 1,024-word-General sentence set (Figure 3.7).

We applied an ablation-based salience mapping technique to calculate the amount of the contribution of each electrode on the array to streaming speech decoding. The strongest contributions came from electrodes along the central sulcus and a middle section of the precentral gyrus. Most of the electrodes on the superior temporal gyrus (STG) had relatively low contributions, except for two smaller clusters on the anterior and dorsal–posterior portions of the array's STG coverage. The spatial patterns of electrode contributions were highly consistent regardless of the presence of feedback, with a significant and high correlation of 0.79 (Pearson r; $P < 0.0001$, Pearson correlation permutation test). In terms of decoding performance, there was no significant difference between the two conditions for both modalities. In both modalities, the median WERs were lower without auditory feedback; however, the difference was insignificant ($P = 0.95$ for speech and $P = 0.09$ for text using a two-sided Wilcoxon signed-rank test corrected by two-way Holm–Bonferroni correction). Speech synthesis requires the model to generate fine-grained acoustic units at a higher sampling rate than text decoding, which leads to higher variability in the synthesized waveforms and reduced sensitivity to any potential differences between feedback conditions (15.5% and 9.3% WER s.d. for real-time 1,024-word-General speech synthesis and text decoding, respectively). Furthermore, we performed region exclusion analysis for the auditory feedback condition for the postcentral gyrus, precentral gyrus and STG. We found similar performance trends to those reported in prior work with this participant without auditory feedback (Metzger et al., 2023). In summary, we show that the decoders were unaffected by model-generated auditory feedback, instead relying on electrodes that correspond to articulatory control.

3.8 Discussion

Naturalistic continuous speech synthesis from neural activity with minimal delay is a major goal for technologies that restore speech to persons with severe paralysis (Felgoise et al.,

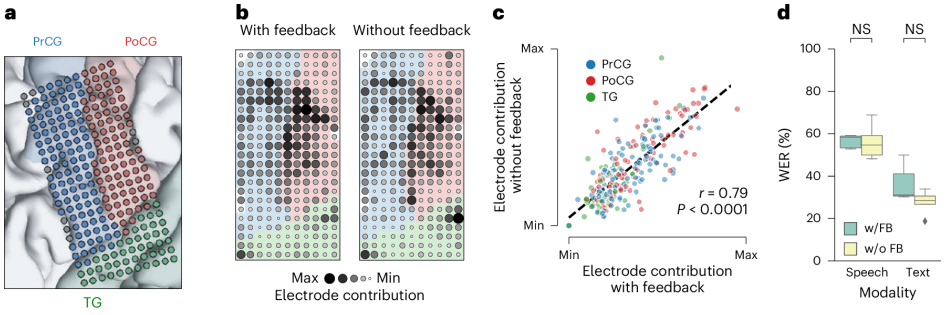

Figure 3.7: **Model-generated auditory feedback does not interfere with articulatory-driven speech decoding.** **a**, Placement of the electrodes on the speech sensorimotor cortex: precentral gyrus (PrCG; blue), postcentral gyrus (PoCG; red) and temporal gyrus (TG; green). **b**, Contribution maps calculated from two conditions: blocks with auditory feedback during online speech synthesis demonstrations (left) and blocks without decoder feedback (right). Both conditions use the 1,024-word-General sentence set. The values are normalized to be in [0,1], which is done separately for each condition. The contribution is calculated for each electrode by computing the difference in RNN-T loss induced by ablating the electrode. Both conditions show similar across-channel patterns of contributions. **c**, Contribution comparison for each channel colored by anatomical region: precentral gyrus (blue), postcentral gyrus (red) and temporal gyrus (green). The correlation between electrode contributions from the two conditions is 0.79 (Pearson's correlation r = 0.79 with P = 0.00 (by region, precentral gyrus: r = 0.74, P = 0.00; postcentral gyrus: r = 0.81, P = 0.00; temporal gyrus: r = 0.80, P = 0.00; one-sided Pearson correlation permutation test with 10,000 samples). **d**, For both speech and text, there is no significant difference in decoding performance between conditions; w/FB, with feedback; w/o FB, without feedback; for both modalities, P = 0.953 for speech and P = 0.172 for text. Data were analyzed by two-sided Wilcoxon signed-rank test with a two-way Holm–Bonferroni correction; n = 10 pseudoblocks for each modality and condition; NS (not significant), P > 0.05. Box plots show the median (horizontal line inside box), 25th and 75th percentiles (box), 25th and 75th percentiles ±1.5 times the interquartile range (whiskers) and outliers (diamonds).

2016; Huggins et al., 2011; Peters et al., 2024; 2015; Silva et al., 2024a; Guenther et al., 2009). We demonstrated that this can be achieved using high-density surface recordings of cortical activity in the speech sensorimotor cortex. Our deep learning models could stream synthesized speech in just 80-ms increments and with simultaneous text decoding, using the participant's neural activity as they silently attempted to 'mime' or 'mouth,' without any vocalization, complete sentences from a 1,024-word vocabulary. Speech-evoked neural activity was time aligned with decoded outputs, suggesting a learned alignment between brain data and target speech.

Previous demonstrations of speech synthesis from cortical activity have primarily been conducted with participants who could speak and did not need assistive communication technology for speech (Angrick et al., 2021; Herff et al., 2019; Anumanchipalli et al., 2019; Angrick et al., 2021). Recent demonstrations for restoring lost speech have been promising but suffer from high latency (Metzger et al., 2023) or only predict text (Card et al., 2024; Metzger et al., 2022; Moses et al., 2021; Willett et al., 2023b; Luo et al., 2023). In particular, three recent publications have shown impressive advances in vocabulary size, decoding speeds and accuracy of text decoding (Metzger et al., 2023; Card et al., 2024; Willett et al., 2023b). However, all three incurred a lengthy wait time before producing audible speech at the end of the attempted production of a sentence, which was then synthesized via direct speech synthesis (Metzger et al., 2023) or TTS (Willett et al., 2023b; Card et al., 2024).

Improving speech synthesis latency and decoding speed is essential for dynamic conversation and fluent communication, which is compounded by the fact that speech synthesis requires additional time to play and for the user and listener to comprehend the synthesized audio. By contrast, decoded text can instantly be displayed on a screen, with rates already approaching the attempted speaking rate of study participants (Metzger et al., 2023; Card et al., 2024; Willett et al., 2023b). The latency of our speech synthesis approach was lower than the previous large-vocabulary speech synthesis BCI by a factor of 8, and communication rates were significantly improved (Metzger et al., 2023). Additionally, speech synthesis outputs were personalized to restore the participant's preinjury voice, a highly desired feature for this participant and others (Metzger et al., 2023; Yamagishi et al., 2012), and the participant self-reported more direct control over the streaming synthesizer than TTS synthesis. Furthermore, prior works relied on neural data associated with attempted vocalization to train a model (Moses et al., 2021; Willett et al., 2023b; Luo et al., 2023; Silva et al., 2024b; Card et al., 2024; Angrick et al., 2023) or did not stream speech synthesis (Metzger et al., 2023; 2022), whereas here, no attempted vocalization was required to train a streaming speech synthesizer. This suggests that this speech synthesis approach may be generalizable to situations where participants cannot produce overt speech during training or inference.

Offline, we also showed that the decoder could operate continuously for several minutes instead of single trials lasting several seconds. This allows for the continuous operation of speech neuroprostheses using a single model, similar to modern approaches for streaming ASR, which implicitly detect the user's speech and operate indefinitely (Zhang et al., 2020;

Shi et al., 2020; Graves, 2012; Li et al., 2020). By applying our model to extended blocks of neural activity, we took initial steps toward enabling long-form speech synthesis suitable for daily needs. The deep learning architectures demonstrate streamability and adaptability across different articulatory silent-speech interfaces, extending beyond just ECoG in a single participant. This is important because silent-speech synthesis from neural recordings without acoustic labels has only been demonstrated in a single participant using ECoG and should be generalizable to other articulatory recording interfaces.

A limitation of this study is that, although the architecture was shown to be generalizable offline to other participants and datasets, the online demonstrations were conducted with only a single participant. Although our synthesized speech waveforms were generally intelligible in this proof-of-principle demonstration, streaming speech synthesis performance was lower than what has been demonstrated via text-decoding methods (Metzger et al., 2023; Card et al., 2024; Willett et al., 2023b) and requires further improvement before the approach would be clinically viable as a speech neuroprosthesis (Silva et al., 2024a; Card et al., 2024). Further advances in electrode interfaces enabling higher spatiotemporal resolution should also continue to improve overall system performance, including latencies (Chiang et al., 2020; Duraivel et al., 2022). Future approaches could include frame-wise synthesizers, which process input neural data using a sliding window and predict a single acoustic output frame per window but require a very high signal-to-noise ratio and accurately aligned overt speech reference waveforms (Anumanchipalli et al., 2019; Wairagkar et al., 2023). Offline demonstrations showed that our approach could achieve above-chance generalization to unseen words during speech synthesis. This suggests that the RNN-T acoustic-speech unit decoding strategy is not limited to word-level vocabulary constraints. However, performance was not consistently as intelligible as for seen words.

Speaking seamlessly with real-time, low-latency communication at will is integral to our sense of identity and belonging, which is severely decreased in individuals with anarthria. Here, we have demonstrated a speech decoding approach that enables low-latency naturalistic spoken communication for speech and text outputs, a major step toward this goal. We show the direct synthesis of speech in synchrony with vocal intent in a paralyzed participant with anarthria and without relying on any vocalization during either training or testing. A major effort moving forward for researchers will be to continue refining the approach, which will ultimately inform the development of a speech neuroprosthesis suitable for daily use by individuals who cannot speak.

3.9 Methods

Clinical Trial Overview

This research was conducted as part of the BCI Restoration of Arm and Voice (BRAVO) clinical trial (ClinicalTrials.gov, NCT03698149). The main goal of the trial is to evaluate the long-term safety and efficacy of an ECoG-based interface, focusing on observing any treatment-related adverse events. The findings discussed are part of an ongoing exploratory clinical trial and are not intended to conclude the trial's primary safety outcomes. The trial commenced in November 2018, with the data gathered between 2022 and 2024. Informed consent was obtained from the participant, who was fully briefed on the details of the study enrollment and the potential risks associated with the study device through several discussions with the research and clinical team. The Food and Drug Administration approved the investigational use of the neural implant device, and the study's protocol was approved by the Institutional Review Board at the University of California, San Francisco. The study protocol was approved by the National Institute on Deafness and Communication Disorders at the National Institutes of Health.

Participant

The participant, a 47-year old at the time of her entry into the study, was diagnosed with quadriplegia and anarthria by neurologists and a speech language pathologist after a right pontine stroke in 2005. At the age of 30, while in good health, she experienced a sudden onset of symptoms, including dizziness, slurred speech, quadriplegia and weakness in the muscles used for speech. Medical investigations revealed a pontine stroke caused by a dissection of the left vertebral artery and an occlusion of the basilar artery. During her enrollment assessment, she scored 29/30 on the Mini-Mental State Exam, missing the total score solely because her paralysis prevented her from completing a drawing task. She is capable of making a limited range of monosyllabic sounds like 'ah' and 'ooh' but cannot form clear words. In clinical evaluations, a speech language pathologist asked her to attempt saying 58 words and ten phrases and to answer two open-ended questions within a structured dialogue. Analysis of audio and video recordings of her speech efforts indicated that her intelligibility was 5% for the words and 0% for phrases and open-ended responses. She cannot use speech for communication; instead, she uses a transparent letter board and a Tobii Dynavox device to express herself. She provided informed consent for participation in the study and for her images to be used in demonstration videos through her transparent letter board. She spelled out 'I consent' using her communication board to sign the consent documents formally and instructed her spouse to sign on her behalf. After implantation, she was compensated $50 per session up to an annual maximum of $5,000.

Neural Implant

The device this research uses incorporates a high-density ECoG array (PMT) and a percutaneous pedestal connector provided by Blackrock Microsystems. The ECoG array is designed with 253 disk-shaped electrodes, organized in a grid pattern with 3-mm center-to-center spacing. Each electrode features a recording contact diameter of 1mm and an overall diameter of 2mm. This array was positioned subdurally on the pial surface of the brain's left hemisphere, strategically covering areas involved in speech production and comprehension, predominantly the precentral and postcentral gyri and a portion of the STG. The percutaneous pedestal connector, secured to the skull in the same surgical procedure, facilitates the transmission of electrical signals from the ECoG array to an external digital headstage (CerePlex E256, Blackrock Microsystems). Implantation occurred in September 2022 at the University of California, San Francisco, Medical Center and was completed without surgical complications.

Signal processing

We used the same signal processing pipeline detailed in our previous work (Metzger et al., 2022; 2023) to extract and process common average referenced HGA (Crone et al., 1998) and LFS from the ECoG signals at a 200-Hz sampling rate. We applied a 30-s sliding window z score in real time to each ECoG channel's HGA and LFS features. All training data and online demos were collected at the participant's residence. We used a custom Python online data collection and task management software (rtNSR (Metzger et al., 2022; Moses et al., 2021; Metzger et al., 2023)), which we continually maintain and use in our work. PyTorch 1.13.1 was used for model-level transformations and signal processing.

Task Design

Experimental Paradigm

The participant silently attempted to speak phrases with short syllable-length pauses between each word precisely as instructed in our previous work (Metzger et al., 2023). Specifically, the participant attempted to articulate the target sentences (that is, physically moving her vocal tract muscles) but did not vocalize any sound. This behavior is akin to 'mouthing' or 'miming' the target sentence, albeit in a minimal capacity, given the participant's lack of the ability to coordinate articulation. This strategy more accurately reflects the use case of speech neuroprostheses for this participant and others who cannot efficiently or reliably produce vocalizations, as vocal speech attempts are prohibitively laborious for her. The prompted text was flanked by three dots on each side, vanishing one at a time in sequence to serve as a countdown. Following the disappearance of the last dot, the text changed to green,

signaling the GO cue and prompting the participant to begin silently attempting to speak the target phrase. Following a short pause, the screen cleared, moving on to the subsequent trial.

During online speech synthesis and text decoding, the decoder began receiving neural features 500 ms before the GO cue. The speech synthesizer streamed predicted speech as she began attempting to speak the sentence silently. Meanwhile, the most recent text outputs were displayed on the screen. The effects of providing explicit instructions to participants to adjust their strategy during online decoding have resulted in conflicting results (Sitaram et al., 2017), so we instructed the participant to silently attempt to speak exactly as done during training regardless of the decoded outputs.

Sentence Sets

We used a 50-phrase-AAC sentence set and a 1,024-word-General sentence set in this work, with the curation being the same as in our previous work (Metzger et al., 2023). The 50-phrase-AAC contained 50 sentences of 119 unique words, with sentences chosen for clinical relevance and everyday dialogue. The 1,024-word-General set contained 13,463 sentences sampled from Twitter and movie transcriptions containing 1,024 unique words. The 1,024-word-General set corpus was designed to facilitate training for large-vocabulary decoding, as in English, the most frequent 1,000 words can cover more than 85% of the content in spoken sentences (Adolphs & Schmitt, 2003). For the 50-phrase-AAC set, we used 11,700 trials collected over 58 weeks. For the 1,024-word-General set, we used 23,378 trials (12,379 unique sentences) collected over 65 weeks of training. Performance was not significantly different when using data collected over the most recent 17 weeks; however, we decided to allow the model to learn from all available training samples. We held out 50 trials as a development set to evaluate performance and choose hyperparameters before online testing. We randomly selected 100 sentences from the 200 1,024-word-General sentences used in our synthesis test set from previous work (Metzger et al., 2023) and used these during our test trials. For the training and testing using the 1,024-word-General sentence set, to assist the decoding models in identifying word boundaries from the neural signals without greatly compromising speed and fluency, we instructed the participant to include brief pauses of syllable length (about 300–500 ms) between words in her silent-speech attempts.

Modeling

Bimodal Decoding

The RNN-T is a sequence-to-sequence model designed for ASR tasks (Graves, 2012). It enables dynamic learning of alignments between variable-length input and output sequences.

The RNN-T comprises three main modules: the neural encoder, the language model and the joiner. The neural encoder, or transcriber, models the posterior of the targets from input neural signals. The language model, or predictor, is an autoregressive model that learns relationships within the target data. Finally, the joiner module combines the neural encoder and language model outputs, forming a $T \times L$ grid, where T and L are the lengths of the input neural data and target sequences, respectively. Each point (t, l) on the grid represents the emission probability of the l-th target given the t-th input neural data. During inference, multiple hypotheses are generated for plausible paths through the $T \times L$ grid, which are then searched by an RNN-T beam search algorithm (Graves, 2012).

Our bimodal decoder has two target modalities, specifically data-driven acoustic-speech units from HuBERT (Hsu et al., 2021) and text byte pair encodings (Kudo & Richardson, 2018). We trained a single shared neural encoder for both modalities and a separate joiner and language model for each modality. A blank token is added for each modality to allow for no emissions when the model is not confident in predicting speech. We use a convolutional neural network for the neural encoder, followed by an RNN. The architecture comprises two one-dimensional convolutional layers with 512 kernels, a kernel size of 7, a stride of 4 and three layers of unidirectional gated recurrent units with 512 hidden units and a dropout rate of 0.5. This effectively downsamples the signal by a factor of 16 to an inference rate of 12.5 Hz or one prediction per 80 ms.

For each modality, the language model comprises four Long Short-Term Memory (LSTM) RNN layers, each with 512 hidden units, a dropout rate of 0.3 and layer normalization. The language model outputs are then added to these transformed features and passed to the joiner with a nonlinear function (hyperbolic tangent), followed by a linear projection to the target space. This resulted in an output distribution over 101 classes (100 acoustic-speech units and 1 blank token) for speech and over 4,097 classes (4,096 subword tokens and 1 blank token) for text. We pretrained the language model from a large speech dataset (960 h of LibriSpeech audio). When training the models using neural activity, the language model parameters were not updated.

When streaming, ECoG features are passed to the neural encoder in 80-ms 'chunks'. As model predictions were being made, we used RNN-T beam search to keep the top K hypotheses for each iteration. Although K determines the search space across L, the system can also track multiple hypotheses over T, which is helpful for text decoding. However, the model only keeps the most likely hypothesis for each time point for speech synthesis because audio chunks cannot be replayed during inference. The outputs of the RNN-T beam search are stored in internal buffers for each modality. For acoustic-speech units, every 80 ms, the speech synthesizer synthesizes audio from four acoustic-speech units from the speech buffer, which are streamed directly to a sound card via the PyAudio Python package (version 0.2.14). For text encodings, the entire buffer is converted to text; during online decoding, it is displayed on the monitor.

Streaming Speech Synthesizer

We trained and designed a speech synthesizer, or acoustic-speech unit vocoder, to synthesize the predicted acoustic-speech units into personalized, intelligible speech in increments of 80 ms. The speech synthesizer is implemented using a custom HiFi-CAR, a generative adversarial network model designed to generate high-fidelity speech waveforms from articulatory or acoustic features (Wu et al., 2022). Our speech synthesizer is trained on a large single-speaker dataset (LJSpeech (Ito & Johnson, 2017)), for which we applied voice conversion to each waveform to convert the default voice into personalized audio. Specifically, we used a short voice clip of the participant recorded before she lost her speech ability to condition a voice conversion module (YourTTS (Casanova et al., 2022)), which converted the LJSpeech audio waveform into personalized speech.

Incremental TTS

As an alternative to continuous speech synthesis, we developed an incremental TTS decoding system for playing back decoded speech to the participant. We used a pretrained TTS model, VITS (Kim et al., 2021), to synthesize the speech one word at a time as the text was predicted. During online decoding, once the text decoder predicted a new word, we passed the complete predicted phrase into the TTS system to generate an output waveform. We then used VITS's internal word duration predictor model to identify the waveform segment corresponding to the newly predicted word. This waveform was played back to the participant via the sound card using the same method for continuous speech synthesis. The 1024-word-General sentence set real-time neural decoder was used; however, the continuous speech synthesis language model and joiner network were disabled to speed up inference. We tailored the TTS output speed to the participant's preferred speaking rate.

System Evaluation

Error Rate Calculation and Perceptual Evaluations

The error rate of a sequence is typically defined as the minimum number of deletions, insertions and substitutions needed to convert a decoded transcript into the target transcript divided by the number of words in the target transcript. The units of such operations are words, characters and phonemes for WER, CER and PER, respectively. Single-trial error rates are quite variable due to differences in the lengths of the sentences, so BCI error rates are typically assessed and reported over sets of sentences rather than individual trials (Moses et al., 2021; Metzger et al., 2023). We sequentially parceled sentences into 'pseudoblocks' of ten sentences before computing metrics over their distributions, which is done by summing

the edit distances between each pair of sentences and dividing by the total number of tokens (that is, words, characters or phonemes) in the target sentences. We determined phoneme sequences using g2p-en (Park & Kim, 2019), a grapheme-to-phoneme model.

We conducted a perceptual assessment for online speech synthesis using crowdsourced workers from Amazon Mechanical Turk. Each of the 250 online trials was independently evaluated by 12 workers (except for 26 trials where only 11 workers completed their evaluations and 5 trials where only 10 workers completed their evaluations). Evaluators listened to the decoded speech and transcribed what they heard. The instructions and Mechanical Turk setup were identical to our previous work (Metzger et al., 2023). To control for outlier evaluator performance, for each trial, we used the transcript corresponding to the median CER across evaluators as the final predicted transcript before computing error rate distributions.

Perceptual transcriptions can be costly to implement at scale, however. Because ASR models are now beginning to match human-level performance (Radford et al., 2022), we used ASR for offline speech synthesis evaluations. We used a state-of-the-art large ASR model known as Whisper (Radford et al., 2022) to transcribe the decoded speech for offline analyses using speech transcripts. For online speech synthesis, we computed word, character and PERs from transcripts generated from Whisper and found no significant difference in performance compared to perceptual transcriptions. This indicates that, for our study, leading ASR models are suitable for speech synthesis evaluation. However, more work remains to characterize its suitability in other speech synthesis systems with different artifact profiles or speech synthesis quality (Varshney et al., 2023).

Speech Detection

To estimate when the participant was silently attempting to speak, we trained a speech detection model from neural features, with modeling parameters identical to our previous demonstrations (Moses et al., 2021; Metzger et al., 2022; 2023). These predictions were used for decoding speed and latency calculations and were not used online.

The speech detection model was trained on the last calendar month of data from the training sets used to train the online speech synthesis models. The three class labels (silence, speech and preparation) were automatically labeled for all data based on the task structure. The model used both LFS and HGA, causally processing these features to continuously predict the likelihood of speech at a rate of 200 Hz using three unidirectional LSTM layers containing 128, 96 and 32 hidden units, respectively, followed by a softmax. The probability distribution over the three types of speech events was then smoothed using a running window before being converted into a binary signal (identified as either speech or not) based on a probability threshold of 0.5. A time thresholding technique was also applied, requiring a minimum duration (500 ms) for changes to be recognized as an event's start (onset) or end (offset). These parameters were chosen based on ideal parameters from our prior work (Metzger et al.,

2023).

Decoding Speed and Latency

The onsets and offsets of the synthesized speech were determined by the first and last time points with speech energy above a threshold of 0.1 (Jongseo Sohn et al., 1999). Speech energy was calculated from a z-scored waveform by summing the squared amplitude within a sliding 20-ms window. Text onset was defined as the emission time of the first word(s) emitted. Similarly, text offset was defined as the emission time of the last word(s) emitted.

To measure the synthesized words per minute (WPM) observed during online testing, we used the formula:

$$\text{WPM} = \frac{N}{T}$$

where N is the number of words in the synthesized waveform, and T is the time (in minutes) that the participant attempted to speak.

We computed T by calculating the elapsed time between the participant's detected silent-speech attempt time, t_{attempt}, and the offset time of the synthesized speech, $t_{\text{synthesis\_offset}}$, yielding the final rate:

$$\text{WPM} = \frac{N}{t_{\text{synthesis\_offset}} - t_{\text{attempt}}}$$

For latency measurements, we computed the absolute elapsed time between the participant's detected start of the silent-speech attempt and the onset of the synthesized speech. The offset latency was defined as the absolute elapsed time between the participant's detected end of the silent-speech attempt and the offset of the synthesized speech. We also computed the latency between the GO cue and the onset of the synthesized speech. We only considered speech synthesis for WPM, as rapid text decoding has already been demonstrated (Metzger et al., 2023; Willett et al., 2023b). For latency, we considered the speech latency and text latency independently. Five trials and two trials for the 50-phrase-AAC sentence set and 1,024-word-General sentence set, respectively, had detection times occurring at least 500 s before the GO cue and were excluded from latency analyses. These trials would have occurred before the model could have received neural samples. For the 50-phrase-AAC sentence set, one trial where no audio was emitted was also excluded.

Long-Form Decoding

For the long-form decoding experiments, we passed entire blocks of neural data into the 1,024-word-General sentence set model to perform continuous streaming speech synthesis and text decoding. Because our online demonstrations used four blocks, four blocks of neural activity were independently passed into the model offline, autoregressively, and in 80-ms chunks. Each block lasted approximately 5.9 min. The initial input to the decoder was the data sample recorded 1 s before the GO cue of the block's first trial. We found it beneficial to reset the hidden states of the RNNs (neural encoder and language model) if the output text encodings repeated for more than four consecutive seconds, which indicates that the model has yet to generate additional outputs. We used voice activity detection to synthesize only the active region of speech and used the same 0.1 thresholds as we did for decoding latency calculations. Our internal speech buffer enforces speech output in 80-ms increments, meaning that the synthesized speech sounds the same as if generated in real time.

Neural Feature Evaluation

To quantify the contribution of each electrode to the online decoding, we applied an ablation-based salience mapping method via determining the loss objective difference induced by occluding a specific channel in the input neural data. For each of the 253 channels, both HGA and LFS signals were set as 0 for all 1,024-word-General test trials. The per-channel contribution value was then measured by calculating the difference in loss compared to the original RNN-T loss, which was calculated with complete coverage of the channels. These differences were then averaged across trials within conditions.

The no-feedback condition consisted of 50 trials, collected on the same day as the 1,024-word-General online demonstrations. The task setup for these trials was the same as during the online demonstrations, except no text decoding or speech synthesis was decoded as feedback. These sentences corresponded with the first 50 sentences from the online 1,024-word-General demonstrations. In the auditory feedback condition, we only used the first 50 of the 100 test trials to control for variations in the sentence content. For the WER evaluation, we used ten pseudoblocks of five-sentence segments.

Cross-Recording Modality Training and Evaluation

We constrained the test datasets from each recording modality to contain only data recorded during silent-speech attempts to evaluate the model's generalizability for silent-speech synthesis. For MEA, we began with the data split from a recent publication demonstrating high-performance open-vocabulary text decoding with a person with paralysis (Willett et al.,

2023a). We then moved all nonsilent utterances from the test set to the training or validation set. For EMG, which uses electrodes to measure the voltage potentials on the surface of the vocal tract apparatus as a person attempts to speak, we used a dataset from a single healthy speaker (Gaddy, 2020). We also filtered out utterances lacking alphanumeric characters and removed utterances containing words that begin or end with apostrophes (which only affected the training utterances). No vocabulary constraints were applied for the MEA and EMG datasets. For ECoG, we used our speech synthesis test set from our previous 1,024-word vocabulary dataset with this participant (Metzger et al., 2023).

Bibliography

Unreal engine, 2020. URL https://www.unrealengine.com/en-US.

Adolphs, S. and Schmitt. Lexical Coverage of Spoken Discourse. *Applied Linguistics*, 24(4):425–438, December 2003. ISSN 0142-6001, 1477-450X. doi: 10.1093/applin/24.4.425. URL https://academic.oup.com/applij/article-lookup/doi/10.1093/applin/24.4.425.

Angrick, M., Herff, C., Mugler, E., Tate, M. C., Slutzky, M. W., Krusienski, D. J., and Schultz, T. Speech synthesis from ECoG using densely connected 3D convolutional neural networks. *Journal of Neural Engineering*, 16(3):036019, June 2019. ISSN 1741-2560, 1741-2552. doi: 10.1088/1741-2552/ab0c59. URL https://iopscience.iop.org/article/10.1088/1741-2552/ab0c59.

Angrick, M., Ottenhoff, M. C., Diener, L., Ivucic, D., Ivucic, G., Goulis, S., Saal, J., Colon, A. J., Wagner, L., Krusienski, D. J., Kubben, P. L., Schultz, T., and Herff, C. Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity. *Communications Biology*, 4(1):1055, December 2021. ISSN 2399-3642. doi: 10.1038/s42003-021-02578-0. URL https://www.nature.com/articles/s42003-021-02578-0.

Angrick, M., Luo, S., Rabbani, Q., Candrea, D. N., Shah, S., Milsap, G. W., Anderson, W. S., Gordon, C. R., Rosenblatt, K. R., Clawson, L., Maragakis, N., Tenore, F. V., Fifer, M. S., Hermansky, H., Ramsey, N. F., and Crone, N. E. Online speech synthesis using a chronically implanted brain-computer interface in an individual with ALS. preprint, Neurology, July 2023. URL http://medrxiv.org/lookup/doi/10.1101/2023.06.30.23291352.

Anumanchipalli, G. K., Chartier, J., and Chang, E. F. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, April 2019. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-019-1119-1. URL http://www.nature.com/articles/s41586-019-1119-1.

Arce, F. I., Lee, J.-C., Ross, C. F., Sessle, B. J., and Hatsopoulos, N. G. Directional information from neuronal ensembles in the primate orofacial sensorimotor cortex. *American Journal of*

Physiology-Heart and Circulatory Physiology, June 2013. doi: 10.1152/jn.00144.2013. URL https://journals.physiology.org/doi/10.1152/jn.00144.2013. Publisher: The American Physiological Society.

Berezutskaya, J., Freudenburg, Z. V., Vansteensel, M. J., Aarnoutse, E. J., Ramsey, N. F., and van Gerven, M. A. Direct Speech Reconstruction from Sensorimotor Brain Activity with Optimized Deep Learning Models. preprint, Neuroscience, August 2022. URL http://biorxiv.org/lookup/doi/10.1101/2022.08.02.502503.

Berger, M. A., Hofer, G., and Shimodaira, H. Carnival—Combining Speech Technology and Computer Animation. *IEEE Computer Graphics and Applications*, 31(5):80–89, September 2011. ISSN 1558-1756. doi: 10.1109/MCG.2011.71. Conference Name: IEEE Computer Graphics and Applications.

Berndt, D. J. and Clifford, J. Using Dynamic Time Warping to Find Patterns in Time Series.

Beukelman, D. R., Mirenda, P., and others. *Augmentative and alternative communication*. Paul H. Brookes Baltimore, 1998.

Binder, J. R. The Wernicke area. *Neurology*, 85(24):2170–2175, December 2015. ISSN 0028-3878. doi: 10.1212/WNL.0000000000002219. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4691684/.

Bird, S. and Loper, E. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pp. 214–217, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/P04-3031.

Bouchard, K. E., Mesgarani, N., Johnson, K., and Chang, E. F. Functional organization of human sensorimotor cortex for speech articulation. *Nature*, 495(7441):327–332, March 2013. ISSN 1476-4687 (Electronic)\r0028-0836 (Linking). doi: 10.1038/nature11911. URL https://www.nature.com/articles/nature11911.

Brady, P. T. Effects of Transmission Delay on Conversational Behavior on Echo-Free Telephone Circuits. *Bell System Technical Journal*, 50(1):115–134, January 1971. ISSN 00058580. doi: 10.1002/j.1538-7305.1971.tb02538.x. URL https://ieeexplore.ieee.org/document/6771855.

Branco, M. P., Pels, E. G. M., Sars, R. H., Aarnoutse, E. J., Ramsey, N. F., Vansteensel, M. J., and Nijboer, F. Brain-Computer Interfaces for Communication: Preferences of Individuals With Locked-in Syndrome. *Neurorehabilitation and Neural Repair*, 35(3):267–279, March 2021. ISSN 1552-6844. doi: 10.1177/1545968321989331.

Breshears, J. D., Molinaro, A. M., and Chang, E. F. A probabilistic map of the human ventral sensorimotor cortex using electrical stimulation. *Journal of Neurosurgery*, 123(2):340–349, August 2015. ISSN 1933-0693. doi: 10.3171/2014.11.JNS14889.

Bibliography

Brumberg, J. S., Pitt, K. M., and Burnison, J. D. A Noninvasive Brain-Computer Interface for Real-Time Speech Synthesis: The Importance of Multimodal Feedback. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(4):874–881, April 2018. ISSN 1558-0210. doi: 10.1109/TNSRE.2018.2808425. Conference Name: IEEE Transactions on Neural Systems and Rehabilitation Engineering.

Bruurmijn, M. L. C. M., Pereboom, I. P. L., Vansteensel, M. J., Raemaekers, M. A. H., and Ramsey, N. F. Preservation of hand movement representation in the sensorimotor areas of amputees. *Brain*, 140(12):3166–3178, 2017. doi: 10.1093/brain/awx274.

Card, N. S., Wairagkar, M., Iacobacci, C., Hou, X., Singer-Clark, T., Willett, F. R., Kunz, E. M., Fan, C., Nia, M. V., Deo, D. R., Srinivasan, A., Choi, E. Y., Glasser, M. F., Hochberg, L. R., Henderson, J. M., Shahlaie, K., Stavisky, S. D., and Brandman, D. M. An accurate and rapidly calibrating speech neuroprosthesis. *New England Journal of Medicine*, 391(7):609–618, 2024. doi: 10.1056/NEJMoa2314132. URL https://www.nejm.org/doi/full/10.1056/NEJMoa2314132.

Carey, D., Krishnan, S., Callaghan, M. F., Sereno, M. I., and Dick, F. Functional and Quantitative MRI Mapping of Somatomotor Representations of Human Supralaryngeal Vocal Tract. *Cerebral Cortex (New York, N.Y.: 1991)*, 27(1):265–278, January 2017. ISSN 1460-2199. doi: 10.1093/cercor/bhw393.

Casanova, E., Weber, J., Shulby, C., Junior, A. C., Gölge, E., and Ponti, M. A. YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone, February 2022. URL http://arxiv.org/abs/2112.02418. arXiv:2112.02418 [cs, eess].

Chang, E. F. and Anumanchipalli, G. K. Toward a speech neuroprosthesis. *JAMA*, 323(5):413–414, February 2020.

Chang, E. F., Niziolek, C. A., Knight, R. T., Nagarajan, S. S., and Houde, J. F. Human cortical sensorimotor network underlying feedback control of vocal pitch. *Proceedings of the National Academy of Sciences*, 110(7):2653–2658, February 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1216827110. URL https://pnas.org/doi/full/10.1073/pnas.1216827110.

Chartier, J., Anumanchipalli, G. K., Johnson, K., and Chang, E. F. Encoding of Articulatory Kinematic Trajectories in Human Speech Sensorimotor Cortex. *Neuron*, 98(5):1042–1054.e4, 2018. doi: 10.1016/j.neuron.2018.04.031. URL https://doi.org/10.1016/j.neuron.2018.04.031.

Cheung, C., Hamilton, L. S., Johnson, K., and Chang, E. F. The auditory representation of speech sounds in human motor cortex. *eLife*, 5:e12577, 2016. doi: 10.7554/eLife.12577. Erratum in: eLife. 2016 Apr 27;5:e17181. doi: 10.7554/eLife.17181.

Chiang, C.-H., Won, S. M., Orsborn, A. L., Yu, K. J., Trumpis, M., Bent, B., Wang, C., Xue, Y., Min, S., Woods, V., Yu, C., Kim, B. H., Kim, S. B., Huq, R., Li, J., Seo, K. J., Vitale, F., Richardson, A., Fang, H., Huang, Y., Shepard, K., Pesaran, B., Rogers, J. A., and Viventi, J. Development of a neural interface for high-definition, long-term recording in rodents and nonhuman primates. *Science Translational Medicine*, 12(538):eaay4682, April 2020. doi: 10.1126/scitranslmed.aay4682. URL https://www.science.org/doi/full/10.1126/scitranslmed.aay4682. Publisher: American Association for the Advancement of Science.

Cho, C. J., Wu, P., Mohamed, A., and Anumanchipalli, G. K. Evidence of Vocal Tract Articulation in Self-Supervised Learning of Speech. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, June 2023. doi: 10.1109/icassp49357.2023.10094711. URL https://doi.org/10.1109%2Ficassp49357.2023.10094711.

Collobert, R., Puhrsch, C., and Synnaeve, G. Wav2Letter: an End-to-End ConvNet-based Speech Recognition System, September 2016. URL http://arxiv.org/abs/1609.03193. arXiv:1609.03193 [cs].

Crone, N. E., Miglioretti, D. L., Gordon, B., and Lesser, R. P. Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band. *Brain*, 121(12):2301–2315, December 1998. ISSN 0006-8950. doi: 10.1093/brain/121.12.2301. URL https://doi.org/10.1093/brain/121.12.2301.

Danescu-Niculescu-Mizil, C. and Lee, L. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs, June 2011. URL http://arxiv.org/abs/1106.3077. arXiv:1106.3077 [physics].

Dichter, B. K., Breshears, J. D., Leonard, M. K., and Chang, E. F. The control of vocal pitch in human laryngeal motor cortex. *Cell*, 174(1):1–11, 2018. doi: 10.1016/j.cell.2018.05.016. URL https://doi.org/10.1016/j.cell.2018.05.016.

Duraivel, S., Rahimpour, S., Chiang, C.-H., Trumpis, M., Wang, C., Barth, K., Lad, S. P., Friedman, A. H., Southwell, D. G., Sinha, S. R., Viventi, J., and Cogan, G. High-resolution neural recordings improve the accuracy of speech decoding. preprint, Neuroscience, May 2022. URL http://biorxiv.org/lookup/doi/10.1101/2022.05.19.492723.

Eichert, N., Papp, D., Mars, R. B., and Watkins, K. E. Mapping Human Laryngeal Motor Cortex during Vocalization. *Cerebral Cortex*, 30(12):6254–6269, November 2020. ISSN 1047-3211. doi: 10.1093/cercor/bhaa182. URL https://doi.org/10.1093/cercor/bhaa182.

Eichert, N., Watkins, K. E., Mars, R. B., and Petrides, M. Morphological and functional variability in central and subcentral motor cortex of the human brain. *Brain Structure &*

Function, 226(1):263–279, 2021. ISSN 1863-2653. doi: 10.1007/s00429-020-02180-w. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7817568/.

Ekman, P. and Friesen, W. V. Facial Action Coding System, January 2019. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/t27734-000.

Felgoise, S. H., Zaccheo, V., Duff, J., and Simmons, Z. Verbal communication impacts quality of life in patients with amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 17(3-4):179–183, May 2016. ISSN 2167-8421, 2167-9223. doi: 10.3109/21678421.2015.1125499. URL https://www.tandfonline.com/doi/full/10.3109/21678421.2015.1125499.

Gaddy, D. Silent Speech EMG, October 2020. URL https://zenodo.org/record/4064409.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., and others. MEG and EEG data analysis with MNE-Python. *Frontiers in neuroscience*, 7, 2013. doi: 10.3389/fnins.2013.00267. Publisher: Frontiers.

Graves, A. Sequence Transduction with Recurrent Neural Networks. 2012. doi: 10.48550/ARXIV.1211.3711. URL https://arxiv.org/abs/1211.3711. Publisher: arXiv Version Number: 1.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*, pp. 369–376, Pittsburgh, Pennsylvania, 2006. ACM Press. ISBN 978-1-59593-383-6. doi: 10.1145/1143844.1143891. URL http://portal.acm.org/citation.cfm?doid=1143844.1143891.

Graves, A., Mohamed, A.-r., and Hinton, G. Speech recognition with deep recurrent neural networks. In Ward, R. and Deng, L. (eds.), *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6645–6649, 2013. doi: 10.1109/ICASSP.2013.6638947.

Guenther, F. H., Brumberg, J. S., Wright, E. J., Nieto-Castanon, D. A., Tourville, J. A., Panko, S. F., Law, M., Siebert, H., Bartels, W. S., Andreasen, P. R., Bouchard, P. L., Kennedy, L. E., Madsen, R. A., Blabe, S. S., Walsh, E. F., Schalk, L. R., and Kennedy, P. R. A wireless brain-machine interface for real-time speech synthesis. *PLoS ONE*, 4(12):e8218, 2009. doi: 10.1371/journal.pone.0008218. URL https://doi.org/10.1371/journal.pone.0008218.

Hannun, A. Y., Maas, A. L., Jurafsky, D., and Ng, A. Y. First-Pass Large Vocabulary Continuous Speech Recognition using Bi-Directional Recurrent DNNs. *arXiv:1408.2873 [cs]*, December 2014. URL http://arxiv.org/abs/1408.2873. arXiv: 1408.2873.

He, Y., Sainath, T. N., Prabhavalkar, R., McGraw, I., Alvarez, R., Zhao, D., Rybach, D., Kannan, A., Wu, Y., Pang, R., Liang, Q., Bhatia, D., Shangguan, Y., Li, B., Pundak, G., Sim, K. C., Bagby, T., Chang, S.-y., Rao, K., and Gruenstein, A. Streaming End-to-end Speech Recognition For Mobile Devices, November 2018. URL http://arxiv.org/abs/1811.06621. arXiv:1811.06621 [cs].

Heafield, K. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 187–197, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. URL https://aclanthology.org/W11-2123.

Herff, C. and Schultz, T. Automatic Speech Recognition from Neural Signals: A Focused Review. *Frontiers in Neuroscience*, 10, 2016. ISSN 1662-453X. URL https://www.frontiersin.org/articles/10.3389/fnins.2016.00429.

Herff, C., Diener, L., Angrick, M., Mugler, E., Tate, M. C., Goldrick, M. A., Krusienski, D. J., Slutzky, M. W., and Schultz, T. Generating Natural, Intelligible Speech From Brain Activity in Motor, Premotor, and Inferior Frontal Cortices. *Frontiers in Neuroscience*, 13:1267, 2019. ISSN 1662-453X. doi: 10.3389/fnins.2019.01267. URL https://www.frontiersin.org/article/10.3389/fnins.2019.01267.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. June 2021.

Huggins, J. E., Wren, P. A., and Gruis, K. L. What would brain-computer interface users want? Opinions and priorities of potential users with amyotrophic lateral sclerosis. *Amyotrophic Lateral Sclerosis*, 12(5):318–324, September 2011. ISSN 1748-2968, 1471-180X. doi: 10.3109/17482968.2011.572978. URL http://www.tandfonline.com/doi/full/10.3109/17482968.2011.572978.

Ito, K. and Johnson, L. The LJ Speech Dataset, 2017. URL https://keithito.com/LJ-Speech-Dataset/.

Jia, J., Wang, X., Wu, Z., Cai, L., and Meng, H. Modeling the correlation between modality semantics and facial expressions. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1–10, December 2012.

Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1):1–3, January 1999. ISSN 1070-9908, 1558-2361. doi: 10.1109/97.736233. URL http://ieeexplore.ieee.org/document/736233/.

Jurafsky, D. and Martin, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson Education, 2009.

Kim, J., Kong, J., and Son, J. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. 2021. doi: 10.48550/ARXIV.2106.06103. URL https://arxiv.org/abs/2106.06103. Publisher: arXiv Version Number: 1.

King, D. E. Dlib-ml: A Machine Learning Toolkit.

King, D. E. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

Kneser, R. and Ney, H. Improved backing-off for m-gram language modeling. In Sanei, S. and Hanzo, L. (eds.), *Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pp. 181–184. IEEE, 1995.

Krauss, R. M. and Bricker, P. D. Effects of Transmission Delay and Access Delay on the Efficiency of Verbal Communication. *The Journal of the Acoustical Society of America*, 41(2):286–292, February 1967. ISSN 0001-4966, 1520-8524. doi: 10.1121/1.1910338. URL https://pubs.aip.org/jasa/article/41/2/286/675454/Effects-of-Transmission-Delay-and-Access-Delay-on.

Kubichek, R. F. Megcepstral distance measure for objective speech quality assessment.

Kudo, T. and Richardson, J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. 2018. doi: 10.48550/ARXIV.1808.06226. URL https://arxiv.org/abs/1808.06226. Publisher: arXiv Version Number: 1.

Lakhotia, K., Kharitonov, E., Hsu, W.-N., Adi, Y., Polyak, A., Bolte, B., Nguyen, T.-A., Copet, J., Baevski, A., Mohamed, A., and Dupoux, E. On Generative Spoken Language Modeling from Raw Audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354, 2021. doi: 10.1162/tacl_a_00430. URL https://aclanthology.org/2021.tacl-1.79. Place: Cambridge, MA Publisher: MIT Press.

Lee, A., Chen, P.-J., Wang, C., Gu, J., Popuri, S., Ma, X., Polyak, A., Adi, Y., He, Q., Tang, Y., Pino, J., and Hsu, W.-N. Direct Speech-to-Speech Translation With Discrete Units. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3327–3339, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.235. URL https://aclanthology.org/2022.acl-long.235.

Li, J., Zhao, R., Meng, Z., Liu, Y., Wei, W., Parthasarathy, S., Mazalov, V., Wang, Z., He, L., Zhao, S., and Gong, Y. Developing RNN-T models surpassing high-performance hybrid models with customization capability. 2020.

Littlejohn, K. T., Cho, C. J., R. Liu, J., B. Silva, A., Yu, B., R. Anderson, V., M. Kurtz-Miott, C., Brosler, S., Kashyap, A. P., P. Hallinan, I., Shah, A., Tu-Chan, A., Ganguly, K., A. Moses, D., F. Chang, E., and Anumanchipalli, G. K. A streaming silent-speech

neuroprosthesis for naturalistic voice restoration, 2024. URL https://dataverse.harvard.edu/citation?persistentId=doi:10.7910/DVN/8TQKC8.

Lotte, F., Brumberg, J. S., Brunner, P., Gunduz, A., Ritaccio, A. L., Guan, C., and Schalk, G. Electrocorticographic representations of segmental features in continuous speech. *Frontiers in Human Neuroscience*, 9:97, February 2015. doi: 10.3389/fnhum.2015.00097.

Luo, S., Angrick, M., Coogan, C., and et al. Stable decoding from a speech bci enables control for an individual with als without recalibration for 3 months. *Advanced Science*, 10 (35):e2304853, 2023. doi: 10.1002/advs.202304853.

Makin, J. G., Moses, D. A., and Chang, E. F. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature Neuroscience*, 23(4):575–582, April 2020. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-020-0608-8. URL http://www.nature.com/articles/s41593-020-0608-8.

Martin, S., Iturrate, I., Millán, J. d. R., Knight, R. T., and Pasley, B. N. Decoding inner speech using electrocorticography: Progress and challenges toward a speech prosthesis. *Front. Neurosci.*, 12, June 2018.

Mehrabian, A. *Silent messages: implicit communication of emotions and attitudes*. 2 edition, 1981.

Mermelstein, P. Articulatory model for the study of speech production. *The Journal of the Acoustical Society of America*, 53(4):1070–1082, April 1973. ISSN 0001-4966, 1520-8524. doi: 10.1121/1.1913427. URL https://pubs.aip.org/jasa/article/53/4/1070/681858/Articulatory-model-for-the-study-of-speech.

Metzger, S. L., Liu, J. R., Moses, D. A., Dougherty, M. E., Seaton, M. P., Littlejohn, K. T., Chartier, J., Anumanchipalli, G. K., Tu-Chan, A., Ganguly, K., and Chang, E. F. Generalizable spelling using a speech neuroprosthesis in an individual with severe limb and vocal paralysis. *Nature Communications*, 13(1):6510, November 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-33611-3. URL https://www.nature.com/articles/s41467-022-33611-3.

Metzger, S. L., Littlejohn, K. T., Silva, A. B., Moses, D. A., Seaton, M. P., Wang, R., Dougherty, M. E., Liu, J. R., Wu, P., Berger, M. A., Zhuravleva, I., Tu-Chan, A., Ganguly, K., Anumanchipalli, G. K., and Chang, E. F. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620(7976):1037–1046, August 2023. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-023-06443-4. URL https://www.nature.com/articles/s41586-023-06443-4.

Moses, D. A., Leonard, M. K., and Chang, E. F. Real-time classification of auditory sentences using evoked cortical activity in humans. *Journal of Neural Engineering*, 15(3):036005, June 2018. doi: 10.1088/1741-2552/aaab6f. Epub 2018 Jan 30.

Moses, D. A., Metzger, S. L., Liu, J. R., Anumanchipalli, G. K., Makin, J. G., Sun, P. F., Chartier, J., Dougherty, M. E., Liu, P. M., Abrams, G. M., Tu-Chan, A., Ganguly, K., and Chang, E. F. Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria. *New England Journal of Medicine*, 385(3):217–227, July 2021. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMoa2027540. URL http://www.nejm.org/doi/10.1056/NEJMoa2027540.

Mugler, E. M., Tate, M. C., Livescu, K., Templer, J. W., Goldrick, M. A., and Slutzky, M. W. Differential representation of articulatory gestures and phonemes in precentral and inferior frontal gyri. *Journal of Neuroscience*, 38(46):9803–9813, 2018. doi: 10.1523/JNEUROSCI.1206-18.2018. URL https://doi.org/10.1523/JNEUROSCI.1206-18.2018.

Murray, E. A. and Coulter, J. D. Organization of corticospinal neurons in the monkey. *Journal of Comparative Neurology*, 195(2):339–365, 1981. ISSN 1096-9861. doi: 10.1002/cne.901950212. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cne.901950212. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cne.901950212.

Müllner, D. Modern hierarchical, agglomerative clustering algorithms, 2011. URL https://arxiv.org/abs/1109.2378.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio, September 2016. URL http://arxiv.org/abs/1609.03499. arXiv:1609.03499 [cs].

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. fairseq: A Fast, Extensible Toolkit for Sequence Modeling, April 2019. URL http://arxiv.org/abs/1904.01038. arXiv:1904.01038 [cs].

Ozker, M., Yu, L., Dugan, P., Doyle, W., Friedman, D., Devinsky, O., and Flinker, A. Speech-induced suppression and vocal feedback sensitivity in human cortex, December 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.12.08.570736.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, April 2015. doi: 10.1109/ICASSP.2015.7178964. ISSN: 2379-190X.

Pandarinath, C., Nuyujukian, P., Blabe, C. H., Sorice, B. L., Saab, J., Willett, F. R., Hochberg, L. R., Shenoy, K. V., and Henderson, J. M. High performance communication by people with paralysis using an intracortical brain-computer interface. *Elife*, 6, February 2017.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. Specaugment: A simple data augmentation method for automatic speech recognition. In Kubin, G. and Kačič, Z. (eds.), *Proceedings of Interspeech 2019*, pp. 2613–2617, 2019.

doi: 10.21437/Interspeech.2019-2680. URL https://doi.org/10.21437/Interspeech.2019-2680.

Park, K. and Kim, J. g2pE, 2019. URL https://github.com/Kyubyong/g2p.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Édouard Duchesnay. Scikit-learn: Machine learning in python, 2018. URL https://arxiv.org/abs/1201.0490.

Peters, B., Bieker, G., Heckman, S., Huggins, J., Wolf, C., Zeitlin, D., and Fried-Oken, M. 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, March 2015. ISSN 0003-9993. doi: 10.1016/j.apmr.2014.03.037. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4383178/.

Peters, B., O'Brien, K., and Fried-Oken, M. A recent survey of augmentative and alternative communication use and service delivery experiences of people with amyotrophic lateral sclerosis in the united states. *Disability and Rehabilitation: Assistive Technology*, 19(4): 1121–1134, 2024. doi: 10.1080/17483107.2022.2149866.

Prenger, R., Valle, R., and Catanzaro, B. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2019.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

Rousseau, M.-C., Baumstarck, K., Alessandrini, M., Blandin, V., Billette de Villemeur, T., and Auquier, P. Quality of life in patients with locked-in syndrome: Evolution over a 6-year period. *Orphanet journal of rare diseases*, 10:88–88, 2015. doi: 10.1186/s13023-015-0304-z.

Sadikaj, G. and Moskowitz, D. S. I hear but I don't see you: Interacting over phone reduces the accuracy of perceiving affiliation in the other. *Computers in Human Behavior*, 89:140–147, December 2018. ISSN 0747-5632. doi: 10.1016/j.chb.2018.08.004. URL https://www.sciencedirect.com/science/article/pii/S0747563218303790.

Sadtler, P. T., Quick, K. M., Golub, M. D., Chase, S. M., Ryu, S. I., Tyler-Kabara, E. C., Yu, B. M., and Batista, A. P. Neural constraints on learning. *Nature*, 512

(7515):423–426, August 2014. ISSN 1476-4687. doi: 10.1038/nature13665. URL https://www.nature.com/articles/nature13665. Number: 7515 Publisher: Nature Publishing Group.

Salari, E., Freudenburg, Z. V., Vansteensel, M. J., and Ramsey, N. F. Classification of Facial Expressions for Intended Display of Emotions Using Brain–Computer Interfaces. *Annals of Neurology*, 88(3):631–636, 2020. ISSN 1531-8249. doi: 10.1002/ana.25821. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.25821. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ana.25821.

Schoenenberg, K., Raake, A., and Koeppe, J. Why are you so slow? – Misattribution of transmission delay to attributes of the conversation partner at the far-end. *International Journal of Human-Computer Studies*, 72(5):477–487, May 2014. ISSN 10715819. doi: 10.1016/j.ijhcs.2014.02.004. URL https://linkinghub.elsevier.com/retrieve/pii/S1071581914000287.

Seabold, S. and Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. In Walt, S. v. d. and Millman, J. (eds.), *Proceedings of the 9th Python in Science Conference*, pp. 92 – 96, 2010. doi: 10.25080/Majora-92bf1922-011.

Shi, B., Hsu, W.-N., Lakhotia, K., and Mohamed, A. Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction, March 2022. URL http://arxiv.org/abs/2201.02184. arXiv:2201.02184 [cs, eess].

Shi, Y., Wang, Y., Wu, C., Yeh, C.-F., Chan, J., Zhang, F., Le, D., and Seltzer, M. Emformer: Efficient Memory Transformer Based Acoustic Model For Low Latency Streaming Speech Recognition. 2020. doi: 10.48550/ARXIV.2010.10759. URL https://arxiv.org/abs/2010.10759. Publisher: arXiv Version Number: 4.

Silva, A. B., Littlejohn, K. T., Liu, J. R., Moses, D. A., and Chang, E. F. The speech neuroprosthesis. *Nature Reviews Neuroscience*, 25(7):473–492, 2024a. doi: 10.1038/s41583-024-00819-9.

Silva, A. B., Liu, J. R., Metzger, S. L., Bhaya-Grossman, I., Dougherty, M. E., Seaton, M. P., Littlejohn, K. T., Tu-Chan, A., Ganguly, K., Moses, D. A., and Chang, E. F. A bilingual speech neuroprosthesis driven by cortical articulatory representations shared between languages. *Nat. Biomed. Eng.*, 8(8):977–991, August 2024b.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In Yoshua Bengio and Yann LeCun, Bengio, Y., and LeCun, Y. (eds.), *Workshop at International Conference on Learning Representations*, 2014.

Sitaram, R., Ros, T., Stoeckel, L., and et al. Closed-loop brain training: the science of neurofeedback. *Nature Reviews Neuroscience*, 18(2):86–100, 2017. doi: 10.1038/nrn.

2016.164. Published correction appears in Nat Rev Neurosci. 2019 May;20(5):314. doi: 10.1038/s41583-019-0161-1.

Sumby, W. H. and Pollack, I. Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, 26(2):212–215, March 1954. ISSN 0001-4966. doi: 10.1121/1.1907309. URL http://asa.scitation.org/doi/10.1121/1.1907309.

Sun, P., Anumanchipalli, G. K., and Chang, E. F. Brain2Char: a deep architecture for decoding text from brain recordings. *Journal of Neural Engineering*, 17(6):066015, December 2020. ISSN 1741-2560, 1741-2552. doi: 10.1088/1741-2552/abc742. URL https://iopscience.iop.org/article/10.1088/1741-2552/abc742.

Umeda, T., Isa, T., and Nishimura, Y. The somatosensory cortex receives information about motor output. *Science Advances*, 5(7):eaaw5388, July 2019. doi: 10.1126/sciadv.aaw5388. URL https://www.science.org/doi/10.1126/sciadv.aaw5388. Publisher: American Association for the Advancement of Science.

van den Oord, A., Vinyals, O., and Kavukcuoglu, K. Neural Discrete Representation Learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6309–6318, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 978-1-5108-6096-4. event-place: Long Beach, California, USA.

Vansteensel, M. J., Pels, E. G., Bleichner, M. G., Branco, M. P., Denison, T., Freudenburg, Z. V., Gosselaar, P., Leinders, S., Ottens, T. H., Boom, M. A. V. D., Rijen, P. C. V., Aarnoutse, E. J., and Ramsey, N. F. Fully implanted brain–computer interface in a locked-in patient with als. *New England Journal of Medicine*, 375(21):2060–2066, 2016. doi: 10.1056/NEJMoa1608085. URL https://www.nejm.org/doi/full/10.1056/NEJMoa1608085.

Varshney, S., Farias, D., Brandman, D. M., Stavisky, S. D., and Miller, L. M. Using automatic speech recognition to measure the intelligibility of speech synthesized from brain signals. In *2023 International IEEE/EMBS Conference on Neural Engineering (NER)*, 2023. doi: 10.1109/NER52421.2023.10123751.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, , Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., and van Mulbregt, P. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, March 2020. ISSN 1548-7105. doi: 10.1038/s41592-019-0686-2. URL https://www.nature.com/articles/s41592-019-0686-2. Number: 3 Publisher: Nature Publishing Group.

Wairagkar, M., Hochberg, L., Brandman, D., and Stavisky, S. Synthesizing speech by decoding intracortical neural activity from dorsal motor cortex. pp. 1–4, 04 2023. doi: 10.1109/NER52421.2023.10123880.

Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.*, 6(60):3021, April 2021.

Watanabe, S., Delcroix, M., Metze, F., and Hershey, J. R. *New era for robust speech recognition: exploiting deep learning*. Springer-Verlag, Berlin, Germany, 2017. ISBN 978-3-319-64680-0. URL http://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=5117928.

Willett, F., Kunz, E., Fan, C., Avansino, D., Wilson, G., Choi, E. Y., Kamdar, F., Glasser, M., Hochberg, L., Druckmann, S., Shenoy, K., and Henderson, J. Data for: A high-performance speech neuroprosthesis, June 2023a. URL https://datadryad.org/stash/dataset/doi:10.5061/dryad.x69p8czpq. Artwork Size: 80440635830 bytes Pages: 80440635830 bytes.

Willett, F. R., Avansino, D. T., Hochberg, L. R., Henderson, J. M., and Shenoy, K. V. High-performance brain-to-text communication via handwriting. *Nature*, 593(7858):249–254, May 2021.

Willett, F. R., Kunz, E. M., Fan, C., Avansino, D. T., Wilson, G. H., Choi, E. Y., Kamdar, F., Glasser, M. F., Hochberg, L. R., Druckmann, S., Shenoy, K. V., and Henderson, J. M. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, August 2023b.

Wolters, M. K., Isaac, K. B., and Renals, S. Evaluating speech synthesis intelligibility using amazon mechanical turk. In Sagisaka, Y. and Tokuda, K. (eds.), *Proceedings of the 7th ISCA Workshop on Speech Synthesis (SSW-7)*, pp. 136–141, 2010.

Wu, P., Watanabe, S., Goldstein, L., Black, A. W., and Anumanchipalli, G. K. Deep Speech Synthesis from Articulatory Representations. 2022. doi: 10.48550/ARXIV.2209.06337. URL https://arxiv.org/abs/2209.06337. Publisher: arXiv Version Number: 1.

Yamagishi, J., Usabaev, B., King, S., Watts, O., Dines, J., Tian, J., Guan, Y., Hu, R., Oura, K., Wu, Y.-J., Tokuda, K., Karhila, R., and Kurimo, M. 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): 984–1004, July 2010. ISSN 1558-7916, 1558-7924. doi: 10.1109/TASL.2010.2045237. URL http://ieeexplore.ieee.org/document/5431023/.

Yamagishi, J., Veaux, C., King, S., and Renals, S. Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. *Acoustical Science and Technology*, 33(1):1–5, 2012. ISSN 1346-3969, 1347-5177. doi: 10.1250/ast.33.1. URL http://www.jstage.jst.go.jp/article/ast/33/1/33_1_1/_article.

Yang, Y.-Y., Hira, M., Ni, Z., Astafurov, A., Chen, C., Puhrsch, C., Pollack, D., Genzel, D., Greenberg, D., Yang, E. Z., Lian, J., Hwang, J., Chen, J., Goldsborough, P., Narenthiran, S., Watanabe, S., Chintala, S., and Quenneville-Bélair, V. Torchaudio: Building Blocks for Audio and Speech Processing. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6982–6986, May 2022. doi: 10.1109/ICASSP43922.2022.9747236. ISSN: 2379-190X.

Zhang, Q., Lu, H., Sak, H., Tripathi, A., McDermott, E., Koo, S., and Kumar, S. Transformer Transducer: A Streamable Speech Recognition Model with Transformer Encoders and RNN-T Loss. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7829–7833, Barcelona, Spain, May 2020. IEEE. ISBN 978-1-5090-6631-5. doi: 10.1109/ICASSP40776.2020.9053896. URL https://ieeexplore.ieee.org/document/9053896/.