Title

The speech neuroprosthesis.

Permalink

https://escholarship.org/uc/item/30w5t5pm

Journal

Nature Reviews Neuroscience, 25(7)

Authors

Silva, Alexander
Littlejohn, Kaylo
Liu, Jessie
et al.

Publication Date

2024-07-01

DOI

10.1038/s41583-024-00819-9

Peer reviewed

Published in final edited form as:
Nat Rev Neurosci. 2024 July ; 25(7): 473–492. doi:10.1038/s41583-024-00819-9.

The speech neuroprosthesis

Alexander B. Silva[1,2], **Kaylo T. Littlejohn**[1,2,3], **Jessie R. Liu**[1,2], **David A. Moses**[1,2], **Edward F. Chang**[1,2,✉]

[1]Department of Neurological Surgery, University of California, San Francisco, San Francisco, CA, USA.

[2]Weill Institute for Neuroscience, University of California, San Francisco, San Francisco, CA, USA.

[3]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA, USA.

Abstract

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. In this Review, we highlight such progress and how it has led to successful speech decoding, first in individuals implanted with intracranial electrodes for clinical epilepsy monitoring and subsequently in individuals with paralysis as part of early feasibility clinical trials to restore speech. We discuss high-spatiotemporal-resolution neural interfaces and the adaptation of state-of-the-art speech computational algorithms that have driven rapid and substantial progress in decoding neural activity into text, audible speech, and facial movements. 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. Given this accelerated rate of progress in the field, we propose key evaluation metrics for speed and accuracy, among others, to help standardize across studies. We finish by highlighting several directions to more fully explore the multidimensional feature space of speech and language, which will continue to accelerate progress towards a clinically viable speech neuroprosthesis.

✉ edward.chang@ucsf.edu .

Author contributions
E.F.C. and A.B.S. researched data for the article and contributed substantially to discussion of the content. All authors wrote the article and reviewed and/or edited the manuscript before submission.

Competing interests
D.A.M., J.R.L. and E.F.C. are inventors on a pending provisional UCSF patent application that is relevant to the neural-decoding approaches surveyed in this work. E.F.C. is an inventor on patent application PCT/US2020/028926, D.A.M. and E.F.C. are inventors on patent application PCT/US2020/043706 and E.F.C. is an inventor on patent US9905239B2, which are broadly relevant to the neural-decoding approaches surveyed in this work. EFC is co-founder of Echo Neurotechnologies, LLC. All other authors declare no competing interests.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41583-024-00819-9.

Introduction

Losing the ability to speak drastically hinders communication and, as a result, substantially reduces quality of life[1]. 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[2–4]. 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[5]. Communication neuroprostheses that decode cortical activity into attempted cursor movements[6–8] or handwriting[9] have made strides in achieving faster spelling in individuals with paralysis. However, communication rates for neuroprostheses controlled by attempted hand movements remain slower and less expressive than natural speech[10].

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[11]. Advancements in speech neuroscience, neural-interface technology and machine learning have accelerated progress towards the goal of a clinically viable speech neuroprosthesis (Fig. 1). 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[12]. Subsequent work has expanded this initial demonstration by directly decoding cortical activity into audible speech[13,14] and allowing more generalizable and rapid text decoding[14–16] in individuals with vocal-tract paralysis (Fig. 1).

In this Review, our main goal is to provide readers with an overview of the rapidly progressing field of speech neuroprosthetics. We first discuss characteristics and features of speech that set it apart from other modes of communication. Next, we briefly overview advances in the understanding of the cortical encoding of these speech features, which have driven much progress in the field. Then, we discuss in depth the leading approaches used in speech decoding and the corresponding metrics used to quantify their performance. Finally, we also propose evaluation guidelines and future directions towards a clinically viable speech neuroprosthesis.

Neurological speech disorders impairing communication

Speech enables efficient and rapid communication at rates of around 150–200 words per minute (WPM) in conversational settings[10]. This is driven by millisecond-level coordination of more than 100 vocal-tract muscles that articulate speech sounds to express language. Speech is distinct from language; speech refers to how vocal-tract muscles — speech articulators — are used to produce sounds, but it is part of the broader domain of language, which also includes meaning (semantics) and syntax. Current speech-decoding applications from brain activity are primarily aimed at restoring communication to individuals with both limb and vocal-tract paralysis[14–16]. For these individuals, their paralysis precludes typical communication methods, such as speaking, writing and typing, with AAC technology often being a slow and effortful means of communicating. Conditions that cause these deficits, including amyotrophic lateral sclerosis (ALS) and brainstem stroke (Box 1), disrupt descending motor-neuron pathways to cranial nerves that innervate muscles in the vocal tract (Box 1).

Individuals with these conditions are often diagnosed with dysarthria, a motor-speech disorder characterized by impaired neuromuscular control over the vocal tract that results in a limited ability to produce intelligible speech. In severe cases, they may be diagnosed with anarthria, which is the complete inability to articulate intelligible speech (Box 1). Importantly, in individuals with dysarthria or anarthria, cerebral cortex neuronal populations, which contain representations of articulatory movements used to produce sounds, may remain intact. These persisting neural representations provide a substrate for decoding intended articulation and vocalization, even if these motor commands no longer reach their destination in the vocal tract. Disorders of language (aphasias) can also affect the ability of a person to understand or produce speech and result from stroke in, lesion to or neurodegeneration of cortical regions, including the superior temporal gyrus (STG), supramarginal gyrus (SMG) and precentral gyrus, among others (Box 1). However, it is unclear whether intended articulatory representations can be recovered from the cortex of individuals with aphasia (see Box 1 for discussion of the challenges posed by differences in cortical representations in language versus speech-motor disorders).

Overview of speech features

Although a thorough discussion of speech features and the speech-production pathway is beyond the scope of this Review, we focus on the points that are critical for understanding current speech-decoding approaches (Fig. 2). Namely, we aim to provide the background necessary to understand the methodology and principles in decoding cortical articulatory representations into intended speech as discussed in the subsequent sections.

Neural populations in the ventral sensorimotor cortex (vSMC) and the middle precentral gyrus (midPrCG) control the millisecond-level movements of the vocal-tract articulators (lips, tongue, jaw and larynx) that give rise to speech sounds[17,18] (Fig. 2a). Axons from the vSMC and midPrCG descend through the corticobulbar tract to the brainstem, terminating in cranial nerve nuclei. Cranial nerves then project out of the brainstem to innervate speech articulator muscles, the final step in translating cortical commands into continuous muscle movements that give rise to speech (Fig. 2a). These articulator movements are coordinated with expiration of air to produce the sound waves of speech[19]. In natural conversation, meaning is primarily conveyed with speech. What features of speech are important for producing and conveying that information? We consider three broad categories of speech features: articulatory, acoustic and linguistic, emphasizing those that have informed speech-decoding approaches.

Articulatory features relate directly to how the vocal tract is shaped to transform airflow from the respiratory system into speech sounds[20]. During continuous speech, the vocal-tract muscles move rapidly between different configurations that may be grouped as labial, front tongue, back tongue and vocalic[21] (Fig. 2b) on the basis of where air is constricted in the vocal tract. For example, a labial configuration involves bringing the lips forward, together and then back to articulate a ‘/b/’, ‘/p/’ or ‘/m/’ sound. Similarly, a front-tongue configuration involves raising the tip of the tongue to the roof of the mouth and bringing it back to make, for example, a ‘/t/’, ‘/d/’, or ‘/n/’ sound (Fig. 2b). These configurations are not a complete description of vocal-tract dynamics; further distinctions in produced sound occur based on whether air passes through the nasal cavity (nasals; /n/, /m/) and whether the vocal tract is partially (‘/s/’, ‘/f/’) or completely (‘/t/’, ‘/p/’) closed. A complementary approach to defining discrete vocal-tract configurations is to measure the location of individual vocal-tract articulators continuously over time, using imaging techniques and biosensors[22] or inference from the produced sound[23–26] (Fig. 2c). These features can be continuously measured, defining kinematic trajectories for individual articulators or different points in the vocal tract.

Acoustic features relate directly to the characteristics of audible speech. Speech sound waves may be represented over time either as an acoustic waveform (amplitude-based) or a speech spectrogram (frequency-based). The envelope of the acoustic waveform illustrates the intensity of speech over time (Fig. 2d), which is an important measurement that correlates with speech rate, global stress patterns and loudness[11]. Often, a mel-frequency scale — in which frequency is logarithmically transformed — is used to visualize the speech spectrogram, as this emphasizes power in perceptually salient frequency bands that better correspond to how the human auditory system processes sound[27] (Fig. 2e). Individual speech sounds can be defined by their spectrotemporal (time–frequency) features in the mel-spectrogram (allowing visualization of recurring speech sound patterns). Another important acoustic feature is the fundamental frequency at which the laryngeal vocal folds resonate during speech, which forms the basis of pitch and can be estimated from the acoustic waveform, using algorithms that infer the periodicity of a temporal signal[28,29] (Fig. 2e). Pitch can be modulated over the course of a phrase to convey additional meaning to the produced words (phrasal prosody), such as whether an utterance is a statement or question (rise in pitch at the end of the phrase)[11]. In tonal languages, such as Mandarin, the meaning of individual words depends on their associated pitch contour (lexical tone)[30]. In addition, pitch is an important aspect of the uniqueness and expressivity of the voice of a person[31].

Linguistic features are created by assigning categorical labels to speech sound patterns and are abstracted from purely articulatory or acoustic features alone (Fig. 2f). Phonemes are defined as the smallest perceptually distinct units of sound that form a language (39 unstressed phonemes exist in English)[32,33]. Phonemes can be grouped on the basis of similarities in their sound patterns, and their articulatory features — the place of articulation of a phoneme in the vocal tract — provide one important grouping category[21]. Sequences of phonemes form the basis of words (~170,000 exist in English) and sentences, which have associated semantic meanings. Phonemes are particularly useful for text decoding, given that they form a discrete, low-dimensional representation of speech (~39D in English) that can be scaled to produce the much larger vocabularies of words and sentences[21].

Cortical encoding of speech features

Cortical and subcortical brain structures, known as the corticobulbar system, are involved in the neural control of speech, but an exhaustive discussion of these is beyond the scope of this Review. Thus, in this section, we focus on how cortical representations of the articulatory, acoustic and linguistic speech features previously described are encoded.

One particularly important region in the corticobulbar system is the speech sensorimotor cortex (SMC), composed of the precentral and postcentral gyri (separated by the central sulcus). Years of work focused on mapping premotor and primary motor representations on the precentral gyrus and central sulcus, revealing a somatotopic organization in which neural control of muscle groups is represented along the ventral–dorsal axis[34–38]. Specifically, the corticobulbar system is involved in controlling the speech articulators and has been localized to the vSMC and midPrCG. Located directly dorsal to this area is the region of cortex involved in controlling hand movements[18,39] (Fig. 2a). The postcentral gyrus, which is primarily responsible for encoding somatosensory information, is also somatotopically organized by body area; however, a growing body of work points to additional postcentral gyrus involvement in motor control[40–42].

Bouchard et al.[17] first demonstrated the dynamics of somatotopic speech-articulator representations in the SMC for the jaw, lips, tongue and larynx during speech production of syllables (Fig. 1; Bouchard et al. 2013). Subsequent work discovered that SMC populations encode articulatory kinematic trajectories and vocal-tract configurations[43] during continuous speech[24,39,44,45] (Fig. 1; Mugler et al. 2018, Chartier et al. 2018 and Dichter et al. 2018). Chartier et al.[24] discovered that the SMC encodes a large variety of low-dimensional movement trajectories of the vocal tract (also called gestures), which could be directly related to all consonants and vowels. Dichter et al.[44] defined dorsal and ventral laryngeal cortical representations, with the dorsal one controlling vocal pitch for prosodic intonation in English[44] and, interestingly, lexical tone production in Mandarin[46]. Furthermore, the midPrCG is involved across many speech functions and does not appear strongly tuned to — that is, consistently and exclusively activated by — a single articulatory group or movement (Fig. 2a), potentially instead having a critical role in speech-motor planning[47–50]. Interestingly, at the neuronal level, locations on the anterior precentral gyrus in orofacial or hand areas contain single neurons tuned to movements of the whole body, including key speech articulators[15,51,52]. Thus, although the ventral–dorsal axis of the SMC issomatotopically arranged in brain areas at the level of neural populations, at the single-neuron level, it may have a more heterogeneous and distributed arrangement.

In addition to articulatory features, acoustic features may be critical for not only speech perception but also speech production. Neural activity in the STG encodes acoustic features — for example, recurring speech sound patterns in the mel-spectrogram — during speech perception[53,54] and may be used to reconstruct perceived speech[55,56]. Beyond providing auditory feedback to one’s own vocalized speech (which may not be present for silently attempted speech), the role of the STG during speech production is less clear, although evidence suggests that acoustic representations may also be used to form and plan targets for speech production[57,58]. Along these lines, when able speakers are instructed to imagine speaking or hearing a word, neural activity in the STG can be used to decode those words with above-chance accuracy[59–63]. Furthermore, when an individual with anarthria silently attempted to speak — that is, was asked to attempt to ‘mime’ with no audible output from that attempt — select neural populations in the STG contributed to decoding performance[14]. The SMG is another region that may contribute to acoustic representations for speech production[64–66] (Fig. 1; Wandelt et al. 2022).

The cortical encoding of higher-order language features such as semantics and syntax is less clearly understood. Semantics appear to be encoded in a distributed manner, across many cortical regions including prefrontal, speech and language association cortices[67,68] (Fig. 1; Tang et al. 2023). Converging evidence suggests that Broca’s area, although classically posited as critical for speech production, may not have a critical role in speech articulation[15,69,70] but may instead be involved in word retrieval and grammar[71,72], although this is an active research area[71]. Thus, speech and language features encoded in the STG, SMG and Broca’s area could potentially be leveraged for speech decoding in disorders in which articulatory representations on the SMC are not intact (Box 1).

In summary, early neurosurgical-stimulation studies provided some of the first insights into the somatotopic layout of the sensorimotor cortex. Recent studies using direct intracranial recordings have been instrumental to define a functional organization of the SMC and the precise cortical dynamics for speech movements[17,24,43,44] (Fig. 1; Bouchard et al. 2013, Mugler et al. 2018, Chartier et al. 2018 and Dichter et al. 2018). At the cellular level, the anterior precentral gyrus in hand and orofacial cortex contains single neurons tuned to multiple vocal-tract articulator movements[15,51,52]. Together, these studies all highlighted a large region of cortex, primarily on the SMC, that has been subsequently targeted in the first clinical trials of speech neuroprostheses.

Decoding speech from articulatory features

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 (Fig. 3). This necessitates choices regarding the neural-recording interface, targeted speech features and nature of the final decoded communication units. In this section, we first discuss neural-recording interfaces that have been used to capture neural activity needed to decode speech features. Then, we review two commonly used approaches to modelling speech features from neural activity: decoding words and phrases (text decoding) and directly decoding the speech waveform (speech synthesis).

Neural-recording interfaces

An ideal neural interface for a speech neuroprosthesis (Fig. 3a) should have three key characteristics. First, it should be safe to implant (and explant) with minimal damage to neural tissue and minimal risk to the patient. Second, it should have sufficient spatiotemporal resolution to facilitate high-performance speech decoding. Third, it should enable stable acquisition of neural signals over the course of several years, ideally decades. Together, neural interfaces with these three characteristics could facilitate speech-decoding systems moving from academic research environments to widespread clinical use.

Non-invasive technologies, such as scalp electroencephalography (EEG), magnetoencephalography (MEG) and functional magnetic resonance imaging (MRI), do not require surgical implantation and therefore are very safe for users. EEG and MEG use non-invasive sensors to measure electric and magnetic fields, respectively, that are generated by neural activity in the brain. These techniques have relatively high temporal resolution but lack spatial specificity, with each sensor recording from a large area of the brain[73]. Functional MRI, by contrast, has higher spatial specificity but lower temporal resolution than EEG and MEG as it measures changes in cerebral blood flow as a correlate of neural activity[74]. Although these non-invasive techniques have facilitated decoding of limited vocabularies of imagined words, syllables and speech sounds[75,76] and reconstruction of semantic content of speech[68] (Fig. 1; Tang et al. 2023), they are not portable — with the exception of EEG — and low spatial or temporal resolution has hindered scaling decoding to larger vocabularies and longer speech segments (sentences). Although efforts to mitigate these limitations and increase portability are ongoing[77–80], this Review focuses on invasive recording techniques that have demonstrated spatiotemporal resolution sufficient to enable high-performance speech decoding.

Invasive recording techniques require implantation surgery and directly record extracellular electric field potentials from the cortex via electrodes, affording both high spatial and temporal resolution. However, all invasive recording techniques share risks related to the anaesthesia, craniotomy and dural opening in the implantation surgery and the possibility of infection[81]. Exact placement of invasive electrodes depends on the technique: intracortical microelectrode arrays (MEAs) are inserted within cortical tissue whereas electrocorticography (ECoG) electrode arrays are placed directly on the brain surface, without penetrating the pial or arachnoid meningeal layers[82].

Intracortical MEA recordings are used to measure neural activity at the level of single neurons or neuron clusters[83]. MEAs typically record high-fidelity signals from a small (a few square millimetres) area of cortex. The first efforts towards a speech neuroprosthesis for an individual with vocal-tract paralysis, developed by Phil Kennedy and Frank Guenther, used an intracortical neurotrophic microelectrode placed in the precentral gyrus, which allowed above-chance synthesis of vowel formants[84] and decoding of phonemes[85] (Fig. 1; Guenther et al. 2009 and Brumberg et al. 2011).

ECoG electrodes are placed directly on the cortex and record local field potentials from spatially distinct neural populations. The high-γ band of the local field potential (70–150 Hz) correlates with neuronal firing rates underneath each ECoG electrode[86,87] and captures cortical representations of acoustic and articulatory features relevant to speech[17,88,89] (Fig. 1; Crone et al. 1998, 2001). ECoG implantation can cover broad areas of cortex relevant for speech, such as the vSMC, midPrCG and STG, with a single array.

ECoG has been used clinically for decades for localizing seizures and for brain mapping in surgical work-up for epilepsy treatment[90]. The discovery of the high-γ band catalysed the use of the ECoG for neuro-scientific research in individuals with epilepsy[88], leading to advances in understanding cortical representations of speech (see 'Cortical encoding of speech features') and the first successful demonstrations of speech decoding from brain activity (see 'Modelling speech features from neural activity'). Importantly, those demonstrations in able speakers with epilepsy served as a proof-of-principle for future applications in individuals with vocal-tract paralysis that led to the first clinical trial for a speech neuroprosthesis using a chronically implanted 128-electrode ECoG array[12]. Previously, data on the long-term safety and stability of ECoG primarily came from its application in epilepsy[82,90–92], including the paddle-style surface strip electrodes with the Neuropace responsive neurostimulation system, which has demonstrated safety and signal stability for at least 9 years[92]. Although ECoG electrodes do not penetrate the cortex, they may be associated with subdural fibrosis[93] with minimal or no cortical injury. Despite this, durable stability has enabled ECoG-based neuroprostheses without the need for daily calibration by the user[12,14,94,95]. Furthermore, a 253-electrode ECoG array with increased channel count and density led to substantial gains towards high-performance speech decoding[14]. Thus, the stable acquisition, broad coverage and high performance of ECoG make it a strong candidate for a clinically viable neuroprosthesis.

On the basis of findings from a first proof-of-concept chronic ECoG trial to restore speech to a person with vocal-tract paralysis[12], intracortical MEA-based approaches have been re-visited with higher channel counts in speech-motor cortex using the Utah MEA[96,97]. Recent studies have also achieved high-performance speech-decoding results[15,16] by measuring relevant single-unit and multi-unit neuronal activity from multiple small (4 × 4 mm^2 (ref. 83)) cortical regions on the anterior precentral gyrus[15,16,52,98]. Previous demonstrations of intracortical insertion of MEAs showed injury and neuronal loss at the implantation location[99], as well as signal instabilities contributing to signal loss and performance decline over time[100,101]. Many efforts are ongoing to improve the safety and stability of MEAs, along with promising work in computational techniques to recalibrate data across days[102–105]. Although more work is needed to validate long-term robustness and signal stability with MEAs, their potential for high-performance speech decoding in combination with computational recalibration for stable acquisition offers potential for future use as a clinically viable neuroprosthesis.

Modelling speech features from neural activity

Text (Fig. 3b) and synthesized speech (Fig. 3c) are two outputs commonly used in speech-decoding systems. 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 (Fig. 3b). 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 (Fig. 3c). In our discussion of text decoding and speech synthesis mentioned subsequently, we follow the progression highlighted by the literature in Fig. 1. For text decoding, we first discuss studies in able speakers, undergoing surgical work-up for epilepsy, that demonstrated approaches to decode phonemes[106,107], words and sentences[108,109] from cortical activity. We next discuss adaptations of these approaches to decode text in people with vocal-tract paralysis who cannot speak[12,14–16]. Finally, we highlight three recent studies that have demonstrated high-performance decoding of sentences with large vocabularies closer to natural speaking rates[14–16]. For speech synthesis, we similarly first discuss approaches developed in able speakers[110–112] before moving to how these approaches have been adapted for individuals with vocal-tract paralysis[13,14]. Given the strong link between vocal-tract articulatory movements and produced speech, we end the speech-synthesis section by discussing standalone approaches to decode articulation[14,52,113,114].

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[115,116]. 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[117]. To build on this result and move beyond direct classification of words or sentences, Makin et al.[108] 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 (Fig. 1; Makin et al. 2020). 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 (Fig. 3b). Multiple studies have demonstrated that isolated phonemes can be decoded from neural activity[17,106,107,118] (Fig. 1; Mugler et al. 2014 and Herff et al. 2015). Herff et al.[106] further demonstrated that, during continuous speech, phoneme sequences could be decoded provided that synchronized produced acoustics were available to aid model training[106]. A language model applied to the decoded phoneme sequences could then generate sentences. Subsequent work built on this finding, applying additional ASR techniques to mitigate the need to synchronize linguistic subunits directly with the neural data (Fig. 3b). By using the connectionist temporal classification (CTC) loss[119], 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.[109] trained an RNN to decode character sequences from brain activity during sentence production. Language models similarly converted the decoded character sequences into sentences (Fig. 1; Sun et al. 2020). 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.[12] instructed a participant with anarthria to attempt to speak sentences word-by-word (with pauses between words). 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[12] (Fig. 1; Moses et al. 2021; Supplementary Table 1). 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[120]. 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 (Supplementary Table 1). 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[15,16] and one with ECoG[14]) 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 (Fig. 1; Metzger et al. 2023, Willett et al. 2023 and Card et al. 2023). Language models then mapped decoded phoneme sequences into words and sentences (Fig. 3b and Supplementary Table 1). 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.

MEA-based decoders, including text decoders[9,15], have traditionally required day-to-day recalibration owing to signal instabilities and drift, requiring additional dedicated time from the user. However, computational techniques are actively being developed to counter this challenge. One line of work involves developing models that learn an underlying manifold (low-dimensional spaces that capture neural population dynamics) structure for a given behaviour[103,104,121], which can then be used to stabilize MEA activity over time. This approach has primarily been applied in the context of upper limb decoders[103] but may scale to speech neuroprostheses in the future. A second approach, specific to text decoding, involves leveraging large language models to automatically recalibrate MEA-based text neuroprostheses. In this approach, MEA activity is decoded into subword units and converted into sentences using language models; this language model output is used as a ‘pseudo-label’ to update the RNN decoding model[9,16,105]. This allows the RNN to be automatically recalibrated without additional training labels or dedicated time on the part of an individual with paralysis. Thus, promising avenues exist to create MEA-based text neuroprostheses that do not require daily recalibration time from the user.

Speech synthesis.—An alternative to decoding text is to synthesize audible speech from brain activity (Fig. 3c). 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[122,123]. 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[124]. Finally, an individual may feel greater embodiment when using a speech neuroprosthesis with a personalized voice that reflects their likeness[14]. 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[110–112,125]. Herff et al.[112] developed a concatenative synthesizer to transform neural activity recorded from SEEG into audible speech (Fig. 1; 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[112]. 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.[111] and Anumanchipalli et al.[110], involved a two-stage decoding process. In the first stage, a deep-learning model is trained to regress neural activity (recorded from SEEG[111] or ECoG[110]) to a time series of acoustic features, such as the mel-spectrogram (Fig. 3b). In the second stage, a speech synthesizer — a vocoder — is used to convert the acoustic representation into an audible speech waveform (Fig. 3b). 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. 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[110]. 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. also computed a human-transcribed WER by asking volunteers to transcribe decoded speech to text, which can then be compared with the ground truth[110].

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[13] (Fig. 1; Angrick et al. 2023; and Supplementary Table 1). 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. Metzger et al. 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[14]. 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 (Supplementary Table 1). 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. Future approaches may explore integrating word-level alignment into speech-synthesis models, using either speech detection or microphone and/or video annotation. Word-level alignment may allow regression-based approaches, tailored for low-latency streaming[13], to scale to individuals with vocal-tract paralysis.

Because speech is variable across people and expressive, neuroprostheses that synthesize speech are well suited for personalization (Fig. 3c). 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[126]; however, larger personalized training datasets can be leveraged if the individual has additional pre-injury audio recordings. Metzger et al.[14] first used this voice-conversion approach to personalize synthesized speech for an individual with vocal-tract paralysis using speech samples recorded before the injury. Card et al.[16] trained a personalized TTS model on a small corpus of pre-injury speech, which is an alternative to using a voice-conversion model. 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[127].

Given the strong link between speech acoustics and underlying vocal-tract movements, articulation constitutes another relevant feature space (Fig. 3d) for speech decoding and speech synthesis[128,129]. 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)[110]. In addition, articulatory features may be useful as a standalone output space for both speech and non-speech orofacial gestures[14,52,113,114] (Fig. 3d). 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[130,131], that provides increased clarity[132]. Studies with both ECoG[14] and MEAs[15,16,52] in individuals with paralysis have demonstrated the feasibility of decoding articulatory features, and Metzger et al. 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[14] (Fig. 3d).

Best practices for evaluating and implementing speech neuroprostheses

The past decade has seen rapid progress in speech decoding from the neural activity of both able speakers and individuals with vocal-tract paralysis. As the field continues to accelerate, it is important to define best practices for evaluating speech-decoding performance and implementing practical, user-controlled speech neuroprostheses. In this section, we propose key metrics for quantifying performance of text-decoding and speech-synthesis systems. We also discuss considerations for scaling speech-decoding research into practical speech-neuroprosthetic systems.

Evaluation and standardization

When evaluating speech-decoding performance in individuals with vocal-tract paralysis, it is important to contextualize findings with the disease aetiology of the individual. For example, brainstem-stroke survivors[4,133,134] and individuals with ALS[135] can retain varying degrees of control over orofacial movements and vocalization. Given that most early studies in individuals with vocal-tract paralysis involve only single participants[12–16,64,95,120], great care is needed when comparing results, as the broad spectrum of disease presentations is not sampled (see Supplementary Table 1 for a high-level overview of current speech neuroprosthesis studies in individuals with paralysis). By contrast, in previous multiparticipant studies in able speakers, participants had near-identical speech capacity[106,108,136–138]. Although disease aetiologies, research goals and stimulus sets are likely to remain variable across speech-decoding studies, the metrics and methods used to quantify performance should be standardized.

Speech instruction type.—In speech-decoding studies, the type of speech instructions given to participants must be carefully considered and reported. Common types of speech instructions include imagined speech (in which participants imagine hearing or saying utterances), silently attempted speech (in which participants attempt to move their orofacial muscles but do not vocalize) and attempted speech (in which participants attempt to move their orofacial muscles and to vocalize; Fig. 4a). Together, imagined, silently attempted and attempted speech reflect intended speech from the participant, which is distinct from internal monologues or thoughts[139]. When attempting to speak, any audible vocalizations that an individual with anarthria produces are generally unintelligible[140] owing to their lack of coordinated vocal-tract control (Fig. 4a). In addition to reporting the type of speech instructions, we propose a few performance metrics to become standard for speech neuroprostheses.

Text-decoding performance.—For text-decoding systems, these metrics draw heavily on precedents set in ASR[141,142]. One such precedent is to report the WER and any explicitly decoded subword unit error rates, which are commonly PERs[14–16,106] or CERs[109,120] (Fig. 4b). Often in ASR, to prevent equal weighing of short and long sentences, error rates are computed over paragraphs or longer segments of text[143,144] rather than across individual sentences. Correspondingly, in text-decoding speech neuroprostheses, error rates should be computed over a series of sequential trials[9,12,15,120]. In ASR, a 5% WER, over longer segments of text, is standard for professional-level transcriptions[141] and 25% or less is considered acceptable[142]. The field of speech neuroprosthetics should adopt the aforediscussed precedents for reporting unit error rates, ensuring proper trial weighing and defined benchmarks.

Speech-decoding systems often use recent advances in artificial intelligence and natural language processing to improve decoding performance. Most prominently, language models can be used to rescore sequences of linguistic features (such as phonemes or words) that are decoded from neural activity. Language models perform this re-scoring by evaluating how likely decoded sequences are to occur in natural language. Thus, by being trained to capture statistical patterns in language, language models can prioritize which decoded sequences from neural activity are most probable. Generative pretrained transformers[145] are a class of language models that are particularly powerful for this application. Although decoding systems should take advantage of error rate improvements that come from using language models, it is important to disambiguate pure neural-decoding performance, which could result from different signal features or decoding models, from improvements that come from leveraging statistical priors in natural language. To this end, decoding metrics, such as subword unit error rates such as CERs or PERs, should be computed with and without the use of language models. A complementary perspective to assess the information content in the neural activity alone may be to investigate the neural encoding of the speech features relevant for decoding[14,15]; however, the specific approach may vary across studies.

In addition to error rates, the decoding speed is a critical text-decoding metric that may be quantified as the decoded WPM (Fig. 4b). The WPM can be computed for any utterance decoded by a speech neuroprosthesis, using a simple procedure. The number of decoded words can be divided by the elapsed time between the onset of a speech attempt and the time the decoding model processed the final sample of neural data (Fig. 4b). The onset of a speech attempt may be measured by a go-cue or an algorithm that detects speech attempts from neural activity. Able speakers often speak at rates of around 150 WPM during conversational speech[10]; however, for individuals with vocal-tract paralysis, their maximal rate of attempted speech can vary considerably and may depend on their residual articulatory capacities and degree of comfort[4,133,134]. For these reasons, a decoding speed standard is difficult to define. Instead, a reasonable goal would be to decode at the same rate that the participant can perform their preferred intended speech type (Fig. 4a). This would likely provide substantial improvements over AAC, which typically do not exceed 15 WPM (refs. 146–148).

A final important text-decoding metric is the vocabulary size used to define a lexicon constraint — the list of valid words that may be output by the model — and train a language model (Fig. 4b). This is distinct from the word set used during text-decoding model training or evaluation, which would be a subset of the lexicon and language model vocabulary (see ‘Vocabulary metrics‘ for further discussion).

In summary, text-decoding systems should report the WER and PER (or other relevant subword error rates such as CERs) with and without a language model, the decoded WPM and the vocabulary size of the lexicon and/or language model.

Speech-synthesis performance.—A strong precedent for evaluating neural speech-synthesis systems can be found in modern speech-processing algorithms, specifically TTS systems. The performance of TTS systems is often described using mel-cepstral distortion (MCD) and open-ended human-transcribed WER[149,150]. MCD is a measure of distortion between perceptually salient mel-scaled spectrotemporal features of the ground truth and decoded waveforms (Fig. 4c). It is preferred to correlation-based metrics, such as computing a Pearson correlation between ground truth and decoded acoustic waveforms, as MCD weighs the features in the produced sound that are most perceptually relevant for conveying information[151]. Open-ended human-transcribed WER complements MCD by providing a more interpretable metric derived from the perceptual intelligibility of the synthesized speech. These industry-standard metrics should also be standard for quantifying neural speech-synthesis performance.

The use of the MCD metric is, however, complicated by the lack of precisely aligned ground-truth speech when working with individuals who cannot speak. To resolve this, proxy waveforms, such as waveforms generated from TTS or recorded by able speakers and used during model training, can be used as references in MCD computations[14]. Separately, the open-ended human-transcribed WER might not be possible in all situations. Most previous neural speech-synthesis systems have not been highly intelligible with an open-ended vocabulary[152] and, thus, have used forced-choice listening tests or word-level classification accuracies instead of open-ended WER. In cases of unintelligibility, these are suitable alternatives to using open-ended vocabularies. ASR can also be used to transcribe synthesized speech into text to compute WERs; however, ASR systems are typically trained to recognize natural waveforms that are not artificially synthesized[153]. Hence, ASR-based WERs should currently be seen as complementary, but not a substitute, to human-transcribed WERs. This may change as work to tailor ASR algorithms to synthesized speech is ongoing[152].

Decoding latency is a final metric of particular importance for speech synthesis, which may be quantified in a few ways. We propose that the system latency be defined as the elapsed time from the onset of an intended speech attempt (or go cue) to the onset of synthesized audio. In addition, the rate at which the speech-synthesis system can output new segments of audio — referred to as ‘buffer size’ in speech processing — should be quantified. Delayed auditory feedback disrupts speech production, so an ideal streaming neural-speech synthesizer would output sound with negligible latencies similar to the self-perception of able speech (within a few milliseconds). Although this is a long-term goal, current streaming neural-speech synthesizers should aim for latencies below 200 ms to minimize effects of delayed auditory feedback[154]. Low-latency feedback may also improve speech neuroprosthetic control through feedback learning[124].

Vocabulary metrics.—Common to both speech-synthesis and text-decoding systems, vocabulary metrics should also be reported. A language’s vocabulary is characterized by Zipf’s law, from which it follows that the most common words in a language make up a disproportionate amount of spoken content in that language[155]. For example, in English, the most common 1,000 words can cover more than 85% of the content in spoken sentences[156]. However, lower frequency words remain an important component of natural language because they facilitate expressivity[156]. Thus, vocabulary metrics should provide insight into the frequency and breadth of words and subword sequences used during speech-decoding model training and evaluation (Fig. 4d). Importantly, three vocabulary sizes should be reported (even if subword units are first decoded): the number of unique words in the utterance set used to train the model, to evaluate the model performance and to define a lexicon or language model (only applicable for text decoders), the latter of which may be increased without substantially limiting performance[14,15]. For large-vocabulary sizes, it is not feasible to evaluate performance on every word. However, a strong indicator that a speech neuroprosthesis would allow the user to produce words beyond the training vocabulary is high performance on an evaluation set that is reflective of natural language, containing words not seen by the speech-decoding model during training. Thus, a long-term goal for both text decoding and speech synthesis is to achieve WERs lower than 5% on natural sentences, freely chosen by the user, containing words that were not seen during model training. As part of this long-term goal, performance should be demonstrated over the course of years and, ideally, decades.

Training times.—A final class of metrics that should be reported across speech-decoding studies are related to the time required to train speech decoders. Specifically, we propose two metrics. First, the amount of training data required to achieve usable system performance, which could be defined as the 25% WER threshold for a minimally viable system and the 5% WER threshold for the performance of an ideal system. Second, the number of days the decoder can maintain this threshold of usable system performance without necessitating dedicated re-training time with the user — either by not re-training or by using automatic recalibration techniques which may be simulated offline. This can highlight the tradeoff between absolute performance and day-to-day stability (Fig. 4e). Here, the time at which system evaluation occurred relative to the time of device implantation (for example, measured in days; Supplementary Table 1) should be reported and may give insight into longevity of the system. These two training-time metrics will help the field monitor progress towards scalable and robust speech-neuroprosthetic solutions in the future, which ideally would be usable after only a brief initial training period and require very-little-to-no explicit recalibration periods on a day-to-day basis.

Practical implementation of a speech neuroprosthesis

In this section, we focus on practical speech neuroprosthesis implementations tailored to work in daily-life settings. Although studying isolated aspects of speech decoding such as text or synthesis with single words or sentences is appropriate for research, an ideal speech neuroprosthesis in daily life would have the capacity to leverage numerous communication modes based on the setting and desire of users. For example, in some situations, users may prefer to limit their output space of speech to single, high-utility words and/or phrases (such as ‘no’, ‘yes’, ‘thank you’, ‘not now’ or ‘bring that’) in exchange for near-perfect decoding performance[157]. A similar tradeoff can be facilitated by having a spelling framework available on the neuroprosthesis, either using attempted speech[120] or handwriting[9], which would allow users to spell out arbitrary sentences with high accuracy (albeit with substantially lower speeds). Although spelling and classification approaches may be critical to a practical system for daily use, they do not address users’ desire for high-speed and natural communication.

Thus, algorithms to directly decode neural activity into sentence-level speech, with a large vocabulary and at natural conversational speeds, are critical in a speech neuroprosthesis. Ideally, these algorithms would be designed to capture the multimodal output space of speech. For example, text outputs may be preferred in web-based or computer-based browsing applications, whereas speech synthesis may be ideal during interpersonal interactions. Virtual interactions can further be facilitated by facial animation accompanying speech synthesis and can be used to express non-verbal gestures such as laughing, smiling and frowning. Synthesized speech and facial animation should be customized to the likeness of the participant, using techniques common in speech and avatar processing.

In developing speech-decoding systems for use in daily life, a few additional user concerns must be considered. As long and recurring training periods for speech neuroprostheses might discourage long-term use[158,159], it is essential to develop decoders that are stable over time or can be quickly recalibrated[102,104,105] to avoid dedicated and high-effort re-training periods inconveniencing users. Transfer learning, either between individuals[108] or different types of speech within the same individual[120], may also be used to expedite decoder training. Another critical concern of neuroprosthesis users is whether thoughts or ideas not intended to be conveyed might be output by the system (see Box 2 for ethical considerations). Although current speech neuroprostheses do not decode or target internal monologue (Fig. 4a), an additional safeguard is to use a speech-detection algorithm to gate any decoding process, which is highly successful in correctly identifying volitional attempts of users to overtly and silently speak[12,120]. In the future, speech-detection algorithms can be trained using neural activity collected during internal monologues or the presence of external sensory stimuli (such as watching TV, listening to a podcast or reading an article) as negative examples to further refine specificity for intended speech.

Importantly, many of the aforementioned user concerns — ease of use, re-training periods and data privacy — are not unique to neuroprostheses but rather apply to many machine-learning-based applications. A growing literature aims to define general machine-learning principles to encourage the development of scalable, reliable and efficient systems[160]. As the field of speech neuroprosthetics grows, adopting shared principles with other machine-learning systems may become increasingly important.

Future directions

In this Review, we have discussed progress towards the development of clinically viable speech neuroprostheses. We now focus on three overarching future directions that will accelerate the field towards achieving that goal: understanding the cortical encoding of speech production, developing better engineering techniques to sample and decode neural activity in daily-life settings and scaling to different clinical aetiologies of speech loss.

Attempts to better understand the cortical control of speech production can provide insight into optimal electrode array placements for speech neuroprostheses. Further understanding of somatotopy and articulatory control on the SMC, both in able speakers[37–39,161] and in individuals with vocal-tract paralysis[14,15], could accelerate progress, which may mirror previous investigations of somatotopy in upper limb representations[8,162–164]. However, it is also vital to better understand non-direct articulatory control of speech, such as which cortical regions are important for speech planning[47,165,166], including semantic[68,167] and phonological representations[58,168] that may be important for forming imagined or attempted speech targets. Such regions and representations may inform the development of speech neuroprostheses to treat individuals with speech disorders in which the SMC is damaged but cognition is otherwise intact. The supplementary motor area (SMA) is another important region in the speech production network[165,169] that is involved in speech initiation[170]. A recent non-peer-reviewed preprint found that neural populations in the SMA are among the earliest predictors of speech onset and may remain strongly activated in imagined and attempted speech[170]. Thus, the SMA may be a particularly useful target for future speech detection algorithms. A parallel line of research should investigate the speech features encoded by single neurons versus neuronal populations, as different sampling scales may offer complementary information, even in the same general anatomical area. Finally, speech decoders have yet to fully leverage paradigms of feedback-based learning that have the potential to entrain new patterns of neural activity for better control[171–174].

Numerous engineering advancements are required to enable daily-life use of speech neuroprostheses. For example, higher-density cortical sampling — via an increased number of surface or penetrating electrodes — has been shown to increase the performance of decoding models[14–16,175,176]. Different approaches to increase the density and lower the invasiveness of ECoG are currently being developed for use in humans and animal models[176–182]. Similarly, both in academic and industry settings, approaches are being developed to improve throughput, sampling, biocompatibility and stability of MEAs[183–186]. In addition, although not ready for chronic implantation, NeuroPixels, or similar probes that sample laminar cortex, may offer a method to record high-density single units across cortical layers in humans[187,188]. Separately, the feasibility of non-invasive methods for communication brain computer interfaces is being actively investigated[77–80], with the goal of improving signal-to-noise and spatial resolution. Overall, a positive correlation exists between decoder performance and electrode density and/or the relevant coverage afforded by the corresponding neural interface, emphasizing that improvements to these two hardware aspects can lead to improved speech brain computer interfaces, regardless of the specific recording modality.

In addition, algorithms for decoding speech from neural activity may be improved. One potential avenue is to leverage advancements in self-supervised learning[189] to learn latent representations that best capture behaviour-related dynamics, even when noise is present in the neural activity[121,190–194]. This may improve the decoding of speech features, such as phonemes or the mel-spectrogram, and lower the amount of training data required. Another approach to improve decoding performance is to leverage large language models, which are currently being developed in industry[145,195], to score the likelihood of decoded sentences from neural activity. Importantly, these large language models have the capacity to integrate conversational context into their predictions, which has been shown to improve neural-decoding pipelines[117]. Contemporary machine-learning architectures, tailored for streaming, may also reduce the latency at which speech-synthesis and text-decoding systems operate[196,197].

System-design advancements are also necessary to scale speech neuroprostheses for daily use. To this end, one promising research avenue involves developing recording technologies that can be wireless, portable and fully implanted[6,198–203]. Complementary work may focus on integrating the speech neuroprosthesis with the phone of users and computer operating systems to allow seamless communication via text, e-mail and web services. In developing these approaches, value-aligned development, in which user feedback is used to guide advancements, will be essential and will help ethically design speech neuroprostheses for clinical use (Box 2). This may involve frequently soliciting feedback from users of speech neuroprostheses, potentially through standardized survey mechanisms. A complementary approach is to survey[158,159] the population of potential users for their desired features in a speech neuroprosthesis.

A final direction towards an optimized speech neuroprosthesis is to explore decoding in the context of a number of different speech disorders (Box 1). This may start by further assessing decoding performance in individuals with anarthria that have varying degrees of vocal-tract paralysis. Specifically, high-performance speech decoding has yet to be demonstrated in an individual that is fully locked-in with negligible residual motor function. Speech-decoding studies might also consider targeting cortical forms of dysarthria (caused by injury to the vSMC)[204,205] by recording from intact speech production regions such as the posterior STG, SMG or preserved aspects of the precentral gyrus. Apraxia of speech and aphasias (Box 1) may similarly be targeted in the future by decoding from intact regions, upstream or parallel in the production process from the lesion. Overall, we believe these directions will accelerate the field of speech decoding towards its ultimate goal: restoring natural and expressive communication to all who have lost it.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors are incredibly grateful to the many people who enrolled in the aforedescribed studies. A.B.S. was supported by the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health under award number F30DC021872. K.T.L. is supported by the National Science Foundation GRFP. J.R.L. and D.A.M. were supported by the National Institutes of Health grant U01 DC018671-01A1.

Glossary

Anarthria

Speech-motor disorder referring to an inability to move the vocal-tract muscles to articulate speech.

Aphasias

A disorder of understanding or expressing language.

Attempted speech

This is an instruction given to individuals with vocal-tract paralysis to attempt to speak the best they can, despite lack of the attempt being intelligible.

Concatenative synthesizer

A speech-synthesis approach that relies on matching neural activity with discrete units of a speech waveform that are then concatenated together.

Corticobulbar system

The pathway through which motor commands from the cortex reach the muscles of the vocal tract. At a high level, cortical motor neurons send axons via the corticobulbar tract which terminate in cranial nerve nuclei in the brainstem. Second-order motor neurons in the cranial nerve nuclei then send axons, that bundle and form cranial nerves, to innervate the muscles of the vocal tract.

Formants

The preferred resonating frequencies of the vocal tract that are critical for forming different vowel sounds.

Language models

Models that are trained to capture the statistical patterns of word occurrences in natural language.

Locked-in syndrome

This refers to a clinical condition in which a participant retains cognitive capacity but has limited voluntary motor function. Locked-in syndrome is a spectrum, ranging from fully locked in states (no residual voluntary motor function) to partially locked in states (some residual voluntary motor function such as head movements).

Mime

An attempt to move vocal-tract muscles without attempting to vocalize.

Sensorimotor cortex

This area of the cortex is composed of the precentral and postcentral gyri, primarily responsible for motor control and sensation, respectively.

Silently attempted speech

This is an instruction given to individuals with vocal-tract paralysis to attempt to speak the best they can, but without vocalizing.

Speech articulators

The vocal-tract muscle groups that are important for producing (articulating) speech, including the lips, jaw, tongue and larynx.

Syntax

The arrangement and structure of words to form coherent sentences.

Vocal-tract paralysis

An inability to contract and move the speech articulators, often caused by injury to descending motor-neuron tracts in the brainstem.

Zipf's law

The law that generally proposes that the frequencies of items are inversely proportional to their ranks.

References

1. Felgoise SH, Zaccheo V, Duff J & Simmons Z Verbal communication impacts quality of life in patients with amyotrophic lateral sclerosis. Amyotroph. Lateral Scler. Front. Degener 17, 179–183 (2016).
2. Das JM, Anosike K & Asuncion RMD Locked-in syndrome. StatPearls https://www.ncbi.nlm.nih.gov/books/NBK559026/ (StatPearls, 2021).
3. Lulé D et al. Life can be worth living in locked-in syndrome. Prog. Brain Res 177, 339–351 (2009). [PubMed: 19818912]
4. Pels EGM, Aarnoutse EJ, Ramsey NF & Vansteensel MJ Estimated prevalence of the target population for brain–computer interface neurotechnology in the Netherlands. Neurorehabil. Neural Repair 31, 677–685 (2017). [PubMed: 28639486]
5. Koch Fager S, Fried-Oken M, Jakobs T & Beukelman DR New and emerging access technologies for adults with complex communication needs and severe motor impairments: state of the science. Augment. Altern. Commun. Baltim. MD 1985 35, 13–25 (2019).
6. Vansteensel MJ et al. Fully implanted brain–computer interface in a locked-in patient with ALS. N. Engl. J. Med 375, 2060–2066 (2016). [PubMed: 27959736]
7. Utsumi K et al. Operation of a P300-based brain–computer interface in patients with Duchenne muscular dystrophy. Sci. Rep 8, 1753 (2018). [PubMed: 29379140]
8. Pandarinath C et al. High performance communication by people with paralysis using an intracortical brain–computer interface. eLife 6, e18554 (2017). [PubMed: 28220753]
9. Willett FR, Avansino DT, Hochberg LR, Henderson JM & Shenoy KV High-performance brain-to-text communication via handwriting. Nature 593, 249–254 (2021). [PubMed: 33981047]
10. Chang EF & Anumanchipalli GK Toward a speech neuroprosthesis. JAMA 323, 413–414 (2020). [PubMed: 31880768]

11. Bull P & Frederikson L in Companion Encyclopedia of Psychology (Routledge, 1994).

12. Moses DA et al. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. N. Engl. J. Med 385, 217–227 (2021). [PubMed: 34260835] The authors first demonstrated speech decoding in a person with vocal-tract paralysis by decoding cortical activity word-by-word into sentences, using a vocabulary of 50 words at a rate of 15 wpm.

13. Angrick M et al. Online speech synthesis using a chronically implanted brain–computer interface in an individual with ALS. Preprint at medRxiv 10.1101/2023.06.30.23291352 (2023).The authors demonstrated speech synthesis of single words from cortical activity during attempted speech in a person with vocal-tract paralysis.

14. Metzger SL et al. A high-performance neuroprosthesis for speech decoding and avatar control. Nature 10.1038/s41586-023-06443-4 (2023).The authors reported demonstrations of speech synthesis and avatar animation (orofacial-movement decoding), along with improved text-decoding vocabulary size and speed, by using connectionist temporal classification loss to train models to map persistent-somatotopic representations on the sensorimotor cortex into sentences during silent speech (a large vocabulary was used at a speech rate of 78 wpm).

15. Willett FR et al. A high-performance speech neuroprosthesis. Nature 10.1038/s41586-023-06377-x (2023).The authors improved text decoding to an expansive vocabulary size at 62 wpm, by training models with connectionist temporal classification loss to decode sentences from multiunit activity from microelectrode arrays on precentral gyrus while a person with dysarthria silently attempted to speak.

16. Card NS et al. An Accurate and Rapidly Calibrating Speech Neuroprosthesis 10.1101/2023.12.26.23300110 (2023).The authors used a similar approach to Willett et al. (2023), demonstrating that doubling the number of microelectrode arrays in the precentral gyrus further improved text-decoding accuracy with a rate of 33 wpm.

17. Bouchard KE, Mesgarani N, Johnson K & Chang EF Functional organization of human sensorimotor cortex for speech articulation. Nature 495, 327–332 (2013). [PubMed: 23426266] Here, the authors demonstrated the dynamics of somatotopic organization and speech-articulator representations for the jaw, lips, tongue and larynx during production of syllables, directly connecting phonetic production with speech-motor control of vocal-tract movements.

18. Carey D, Krishnan S, Callaghan MF, Sereno MI & Dick F Functional and quantitative MRI mapping of somatomotor representations of human supralaryngeal vocal tract. Cereb. Cortex N. Y. N 1991 27, 265–278 (2017).

19. Ludlow CL Central nervous system control of the laryngeal muscles in humans. Respir. Physiol. Neurobiol 147, 205–222 (2005). [PubMed: 15927543]

20. Browman CP & Goldstein L Articulatory gestures as phonological units. Phonology 6, 201–251 (1989).

21. Ladefoged P & Johnson K A Course in Phonetics (Cengage Learning, 2014).

22. Berry JJ Accuracy of the NDI wave speech research system. J. Speech Lang. Hear. Res 54, 1295–1301 (2011). [PubMed: 21498575]

23. Liu P et al. A deep recurrent approach for acoustic-to-articulatory inversion. In 2015 IEEE International Conf. Acoustics, Speech and Signal Processing (ICASSP) 10.1109/ICASSP.2015.7178812 (2015).

24. Chartier J, Anumanchipalli GK, Johnson K & Chang EF Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex. Neuron 98, 1042–1054.e4 (2018). [PubMed: 29779940] The authors demonstrated that, during continuous speech in able speakers, cortical activity on the ventral sensorimotor cortex encodes coordinated kinematic trajectories of speech articulators and gives rise to a low-dimensional representation of consonants and vowels.

25. Illa A & Ghosh PK Representation learning using convolution neural network for acoustic-to-articulatory inversion. In ICASSP 2019 — 2019 IEEE International Conf. Acoustics, Speech and Signal Processing (ICASSP) 10.1109/ICASSP.2019.8682506 (2019).

26. Shahrebabaki AS, Salvi G, Svendsen T & Siniscalchi SM Acoustic-to-articulatory mapping with joint optimization of deep speech enhancement and articulatory inversion models. IEEEACM Trans. Audio Speech Lang. Process 30, 135–147 (2022).

27. Tychtl Z & Psutka J Speech production based on the mel-frequency cepstral coefficients. In 6th European Conf. Speech Communication and Technology (Eurospeech 1999) 10.21437/Eurospeech.1999-510 (ISCA, 1999).

28. Belyk M & Brown S The origins of the vocal brain in humans. Neurosci. Biobehav. Rev 77, 177–193 (2017). [PubMed: 28351755]

29. Simonyan K & Horwitz B Laryngeal motor cortex and control of speech in humans. Neuroscientist 17, 197–208 (2011). [PubMed: 21362688]

30. McCawley JD in Tone (ed. Fromkin VA) 113–131 (Academic, 1978).

31. Murray IR & Arnott JL Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. J. Acoust. Soc. Am 93, 1097–1108 (1993). [PubMed: 8445120]

32. Chomsky N & Halle M The Sound Pattern of English (Harper, 1968).

33. Baddeley A Working Memory xi, 289 (Clarendon/Oxford Univ. Press, 1986).

34. Penfield W & Boldrey E Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. Brain 60, 389–443 (1937).The authors demonstrated evidence of somatotopy on sensorimotor cortex by localizing cortical-stimulation-induced movement and sensation for individual muscle groups.

35. Penfield W & Roberts L Speech and Brain-Mechanisms (Princeton Univ. Press, 1959).This study provided insights into cortical control of speech and language through neurosurgical cases, including cortical resection, direct-cortical stimulation and seizure mapping.

36. Cushing H A note upon the Faradic stimulation of the postcentral gyrus in conscious patients. Brain 32, 44–53 (1909).This study was one of the first that applied direct-cortical stimulation to localize function on the sensorimotor cortex.

37. Roux F-E, Niare M, Charni S, Giussani C & Durand J-B Functional architecture of the motor homunculus detected by electrostimulation. J. Physiol 598, 5487–5504 (2020). [PubMed: 32857862]

38. Jensen MA et al. A motor association area in the depths of the central sulcus. Nat. Neurosci 26, 1165–1169 (2023). [PubMed: 37202552]

39. Eichert N, Papp D, Mars RB & Watkins KE Mapping human laryngeal motor cortex during vocalization. Cereb. Cortex 30, 6254–6269 (2020). [PubMed: 32728706]

40. Umeda T, Isa T & Nishimura Y The somatosensory cortex receives information about motor output. Sci. Adv 5, eaaw5388 (2019). [PubMed: 31309153]

41. Murray EA & Coulter JD Organization of corticospinal neurons in the monkey. J. Comp. Neurol 195, 339–365 (1981). [PubMed: 7251930]

42. Arce FI, Lee J-C, Ross CF, Sessle BJ & Hatsopoulos NG Directional information from neuronal ensembles in the primate orofacial sensorimotor cortex. Am. J. Physiol. Heart Circ. Physiol 10.1152/jn.00144.2013 (2013).

43. Mugler EM et al. Differential representation of articulatory gestures and phonemes in precentral and inferior frontal gyri. J. Neurosci 4653, 1206–1218 (2018).The authors demonstrated that the ventral sensorimotor cortex, not Broca’s area in the inferior frontal gyrus, best represents speech-articulatory gestures.

44. Dichter BK, Breshears JD, Leonard MK & Chang EF The control of vocal pitch in human laryngeal motor cortex. Cell 174, 21–31.e9 (2018). [PubMed: 29958109] The authors uncovered the causal role of the dorsal laryngeal motor cortex in controlling vocal pitch through feedforward motor commands, as well as additional auditory properties.

45. Belyk M, Eichert N & McGettigan C A dual larynx motor networks hypothesis. Philos. Trans. R. Soc. B 376, 20200392 (2021).

46. Lu J et al. Neural control of lexical tone production in human laryngeal motor cortex. Nat. Commun 14, 6917 (2023). [PubMed: 37903780]

47. Silva AB et al. A neurosurgical functional dissection of the middle precentral gyrus during speech production. J. Neurosci 42, 8416–8426 (2022). [PubMed: 36351829]

48. Itabashi R et al. Damage to the left precentral gyrus is associated with apraxia of speech in acute stroke. Stroke 47, 31–36 (2016). [PubMed: 26645260]

49. Chang EF et al. Pure apraxia of speech after resection based in the posterior middle frontal gyrus. Neurosurgery 87, E383–E389 (2020). [PubMed: 32097489]

50. Levy DF et al. Apraxia of speech with phonological alexia and agraphia following resection of the left middle precentral gyrus: illustrative case. J. Neurosurg. Case Lessons 5, CASE22504 (2023). [PubMed: 37014023]

51. Willett FR et al. Hand knob area of premotor cortex represents the whole body in a compositional way. Cell 181, 396–409.e26 (2020). [PubMed: 32220308]

52. Stavisky SD et al. Neural ensemble dynamics in dorsal motor cortex during speech in people with paralysis. eLife 8, e46015 (2019). [PubMed: 31820736] The authors demonstrated that, at single locations on the dorsal precentral gyrus (hand area), neurons are tuned to movements of each key speech articulator.

53. Venezia JH, Thurman SM, Richards VM & Hickok G Hierarchy of speech-driven spectrotemporal receptive fields in human auditory cortex. NeuroImage 186, 647–666 (2019). [PubMed: 30500424]

54. Mesgarani N, Cheung C, Johnson K & Chang EF Phonetic feature encoding in human superior temporal gyrus. Science 343, 1006–1010 (2014). [PubMed: 24482117]

55. Akbari H, Khalighinejad B, Herrero JL, Mehta AD & Mesgarani N Towards reconstructing intelligible speech from the human auditory cortex. Sci. Rep 9, 874 (2019). [PubMed: 30696881]

56. Pasley BN et al. Reconstructing speech from human auditory cortex. PLOS Biol. 10, e1001251 (2012). [PubMed: 22303281]

57. Binder JR The Wernicke area. Neurology 85, 2170–2175 (2015). [PubMed: 26567270]

58. Binder JR Current controversies on Wernicke’s area and its role in language. Curr. Neurol. Neurosci. Rep 17, 58 (2017). [PubMed: 28656532]

59. Martin S et al. Word pair classification during imagined speech using direct brain recordings. Sci. Rep 6, 25803 (2016). [PubMed: 27165452]

60. Pei X, Barbour D, Leuthardt EC & Schalk G Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. J. Neural Eng 8, 046028 (2011). [PubMed: 21750369]

61. Martin S et al. Decoding spectrotemporal features of overt and covert speech from the human cortex. Front. Neuroeng 10.3389/fneng.2014.00014 (2014).

62. Proix T et al. Imagined speech can be decoded from low- and cross-frequency intracranial EEG features. Nat. Commun 13, 48 (2022). [PubMed: 35013268]

63. Simanova I, Hagoort P, Oostenveld R & van Gerven MAJ Modality-independent decoding of semantic information from the human brain. Cereb. Cortex 24, 426–434 (2014). [PubMed: 23064107]

64. Wandelt SK et al. Online internal speech decoding from single neurons in a human participant. Preprint at medRxiv 10.1101/2022.11.02.22281775 (2022).The authors decoded neuronal activity from a microelectrode array in the supramarginal gyrus into a set of eight words while the participant in their study imagined speaking.

65. Acharya AB & Maani CV Conduction aphasia. StatPearls https://www.ncbi.nlm.nih.gov/books/NBK537006/ (StatPearls, 2023).

66. Price CJ, Moore CJ, Humphreys GW & Wise RJ Segregating semantic from phonological processes during reading. J. Cogn. Neurosci 9, 727–733 (1997). [PubMed: 23964595]

67. Huth AG, de Heer WA, Griffiths TL, Theunissen FE & Gallant JL Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–458 (2016). [PubMed: 27121839]

68. Tang J, LeBel A, Jain S & Huth AG Semantic reconstruction of continuous language from non-invasive brain recordings. Nat. Neurosci 26, 858–866 (2023). [PubMed: 37127759] The authors developed an approach to decode functional MRI activity during imagined speech into sentences with preserved semantic meaning, although word-by-word accuracy was limited.

69. Andrews JP et al. Dissociation of Broca’s area from Broca’s aphasia in patients undergoing neurosurgical resections. J. Neurosurg 10.3171/2022.6.JNS2297 (2022).

70. Mohr JP et al. Broca aphasia: pathologic and clinical. Neurology 28, 311–324 (1978). [PubMed: 565019]

71. Matchin W & Hickok G The cortical organization of syntax. Cereb. Cortex 30, 1481–1498 (2020). [PubMed: 31670779]
72. Chang EF, Kurteff G & Wilson SM Selective interference with syntactic encoding during sentence production by direct electrocortical stimulation of the inferior frontal gyrus. J. Cogn. Neurosci 30, 411–420 (2018). [PubMed: 29211650]
73. Thukral A, Ershad F, Enan N, Rao Z & Yu C Soft ultrathin silicon electronics for soft neural interfaces: a review of recent advances of soft neural interfaces based on ultrathin silicon. IEEE Nanotechnol. Mag 12, 21–34 (2018).
74. Chow MSM, Wu SL, Webb SE, Gluskin K & Yew DT Functional magnetic resonance imaging and the brain: a brief review. World J. Radiol 9, 5–9 (2017). [PubMed: 28144401]
75. Panachakel JT & Ramakrishnan AG Decoding covert speech from EEG — a comprehensive review. Front. Neurosci 15, 642251 (2021). [PubMed: 33994922]
76. Lopez-Bernal D, Balderas D, Ponce P & Molina A A state-of-the-art review of EEG-based imagined speech decoding. Front. Hum. Neurosci 16, 867281 (2022). [PubMed: 35558735]
77. Rabut C et al. A window to the brain: ultrasound imaging of human neural activity through a permanent acoustic window. Preprint at bioRxiv 10.1101/2023.06.14.544094 (2023).
78. Kwon J, Shin J & Im C-H Toward a compact hybrid brain–computer interface (BCI): performance evaluation of multi-class hybrid EEG-fNIRS BCIs with limited number of channels. PLOS ONE 15, e0230491 (2020). [PubMed: 32187208]
79. Wittevrongel B et al. Optically pumped magnetometers for practical MEG-based brain–computer interfacing. In Brain–Computer Interface Research: A State-of-the-Art Summary 10 (eds Guger C, Allison BZ & Gunduz A) 10.1007/978-3-030-79287-9_4 (Springer International, 2021).
80. Zheng H et al. The emergence of functional ultrasound for noninvasive brain–computer interface. Research 6, 0200 (2023). [PubMed: 37588619]
81. Fernández-de Thomas RJ, Munakomi S & De Jesus O Craniotomy. StatPearls https://www.ncbi.nlm.nih.gov/books/NBK560922/ (StatPearls, 2024).
82. Parvizi J & Kastner S Promises and limitations of human intracranial electroencephalography. Nat. Neurosci 21, 474–483 (2018). [PubMed: 29507407]
83. Rubin DB et al. Interim safety profile from the feasibility study of the BrainGate Neural Interface system. Neurology 100, e1177–e1192 (2023). [PubMed: 36639237]
84. Guenther FH et al. A wireless brain–machine interface for real-time speech synthesis. PLoS ONE 4, e8218 (2009). [PubMed: 20011034] The authors demonstrated above-chance online synthesis of formants, but not words or sentences, from neural activity recorded with an intracortical neurotrophic microelectrode in the precentral gyrus of an individual with anarthria.
85. Brumberg J, Wright E, Andreasen D, Guenther F & Kennedy P Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech motor cortex. Front. Neurosci 10.3389/fnins.2011.00065 (2011).In a follow-up study to Guenther et al. (2009), the authors demonstrated the above-chance classification accuracy of phonemes.
86. Ray S & Maunsell JHR Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. PLOS Biol. 9, e1000610 (2011). [PubMed: 21532743]
87. Ray S, Crone NE, Niebur E, Franaszczuk PJ & Hsiao SS Neural correlates of high-gamma oscillations (60–200 Hz) in macaque local field potentials and their potential implications in electrocorticography. J. Neurosci 28, 11526–11536 (2008). [PubMed: 18987189]
88. Crone NE, Boatman D, Gordon B & Hao L Induced electrocorticographic gamma activity during auditory perception. Clin. Neurophysiol 112, 565–582 (2001). [PubMed: 11275528]
89. Crone NE, Miglioretti DL, Gordon B & Lesser RP Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization gamma band. Brain 121, 2301–2315 (1998). [PubMed: 9874481]
90. Vakani R & Nair DR in Handbook of Clinical Neurology Vol. 160 (eds Levin KH & Chauvel P) Ch. 20, 313–327 (Elsevier, 2019). [PubMed: 31277857]
91. Lee AT et al. Modern intracranial electroencephalography for epilepsy localization with combined subdural grid and depth electrodes with low and improved hemorrhagic complication rates. J. Neurosurg 1, 1–7 (2022).

92. Nair DR et al. Nine-year prospective efficacy and safety of brain-responsive neurostimulation for focal epilepsy. Neurology 95, e1244–e1256 (2020). [PubMed: 32690786]

93. Degenhart AD et al. Histological evaluation of a chronically-implanted electrocorticographic electrode grid in a non-human primate. J. Neural Eng 13, 046019 (2016). [PubMed: 27351722]

94. Silversmith DB et al. Plug-and-play control of a brain–computer interface through neural map stabilization. Nat. Biotechnol 39, 326–335 (2021). [PubMed: 32895549]

95. Luo S et al. Stable decoding from a speech BCI enables control for an individual with ALS without recalibration for 3 months. Adv. Sci. Weinh. Baden-Wurtt. Ger 10.1002/advs.202304853 (2023).The authors demonstrated stability of electrocorticography-based speech decoding in a person with dysarthria by showing that, despite not re-training a model over the course of months, performance did not drop off.

96. Nordhausen CT, Maynard EM & Normann RA Single unit recording capabilities of a 100 microelectrode array. Brain Res. 726, 129–140 (1996). [PubMed: 8836553]

97. Normann RA & Fernandez E Clinical applications of penetrating neural interfaces and Utah Electrode Array technologies. J. Neural Eng 13, 061003 (2016). [PubMed: 27762237]

98. Wilson GH et al. Decoding spoken English from intracortical electrode arrays in dorsal precentral gyrus. J. Neural Eng 17, 066007 (2020). [PubMed: 33236720]

99. Patel PR et al. Utah array characterization and histological analysis of a multi-year implant in non-human primate motor and sensory cortices. J. Neural Eng 20, 014001 (2023).

100. Barrese JC et al. Failure mode analysis of silicon-based intracortical microelectrode arrays in non-human primates. J. Neural Eng 10, 066014 (2013). [PubMed: 24216311]

101. Woeppel K et al. Explant analysis of Utah electrode arrays implanted in human cortex for brain–computer-interfaces. Front. Bioeng. Biotechnol 10.3389/fbioe.2021.759711 (2021).

102. Wilson GH et al. Long-term unsupervised recalibration of cursor BCIs. Preprint at bioRxiv 10.1101/2023.02.03.527022 (2023).

103. Degenhart AD et al. Stabilization of a brain–computer interface via the alignment of low-dimensional spaces of neural activity. Nat. Biomed. Eng 4, 672–685 (2020). [PubMed: 32313100]

104. Karpowicz BM et al. Stabilizing brain–computer interfaces through alignment of latent dynamics. Preprint at bioRxiv 10.1101/2022.04.06.487388 (2022).

105. Fan C et al. Plug-and-play stability for intracortical brain–computer interfaces: a one-year demonstration of seamless brain-to-text communication. Preprint at bioRxiv 10.48550/arXiv.2311.03611 (2023).

106. Herff C et al. Brain-to-text: decoding spoken phrases from phone representations in the brain. Front. Neurosci 10.3389/fnins.2015.00217 (2015).The authors demonstrated that sequences of phonemes can be decoded from cortical activity in able speakers and assembled into sentences using language models, albeit with high error rates on increased vocabulary sizes.

107. Mugler EM et al. Direct classification of all American English phonemes using signals from functional speech motor cortex. J. Neural Eng 11, 035015 (2014). [PubMed: 24836588] The authors demonstrated that all English phonemes can be decoded from cortical activity of able speakers.

108. Makin JG, Moses DA & Chang EF Machine translation of cortical activity to text with an encoder–decoder framework. Nat. Neurosci 23, 575–582 (2020). [PubMed: 32231340] The authors developed a recurrent neural network-based approach to decode cortical activity from able speakers word-by-word into sentences, with word error rates as low as 3%.

109. Sun P, Anumanchipalli GK & Chang EF Brain2Char: a deep architecture for decoding text from brain recordings. J. Neural Eng 17, 066015 (2020).The authors trained a recurrent neural network with connectionist temporal classification loss to decode cortical activity from able speakers into sequences of characters, which were then built into sentences using language models, achieving word error rates as low as 7% with an over 1,000-word vocabulary.

110. Anumanchipalli GK, Chartier J & Chang EF Speech synthesis from neural decoding of spoken sentences. Nature 568, 493–498 (2019). [PubMed: 31019317] The authors developed a biomimetic approach to synthesize full sentences from cortical activity in able speakers:

articulatory kinematics were first decoded from cortical activity and an acoustic waveform was subsequently synthesized from this intermediate representation.

111. Angrick M et al. Speech synthesis from ECoG using densely connected 3D convolutional neural networks. J. Neural Eng 16, 036019 (2019). [PubMed: 30831567] The authors developed a neural-network-based approach to synthesize single words from cortical activity in able speakers.

112. Herff C et al. Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices. Front. Neurosci 10.3389/fnins.2019.01267 (2019).The authors developed a concatenative speech-synthesis approach for single words in healthy speakers, tailored to limited-sized datasets.

113. Salari E et al. Classification of articulator movements and movement direction from sensorimotor cortex activity. Sci. Rep 9, 14165 (2019). [PubMed: 31578420]

114. Salari E, Freudenburg ZV, Vansteensel MJ & Ramsey NF Classification of facial expressions for intended display of emotions using brain–computer interfaces. Ann. Neurol 88, 631–636 (2020). [PubMed: 32548859]

115. Berezutskaya J et al. Direct speech reconstruction from sensorimotor brain activity with optimized deep learning models. Preprint at bioRxiv 10.1101/2022.08.02.502503 (2022).

116. Martin S et al. Decoding inner speech using electrocorticography: progress and challenges toward a speech prosthesis. Front. Neurosci 10.3389/fnins.2018.00422 (2018).

117. Moses DA, Leonard MK, Makin JG & Chang EF Real-time decoding of question-and-answer speech dialogue using human cortical activity. Nat. Commun 10, 3096 (2019). [PubMed: 31363096]

118. Ramsey NF et al. Decoding spoken phonemes from sensorimotor cortex with high-density ECoG grids. NeuroImage 180, 301–311 (2018). [PubMed: 28993231]

119. Graves A, Fernández S, Gomez F & Schmidhuber J Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proc. 23rd Int. Conf. Machine Learning — ICML ‘06 10.1145/1143844.1143891 (ACM Press, 2006).

120. Metzger SL et al. Generalizable spelling using a speech neuroprosthesis in an individual with severe limb and vocal paralysis. Nat. Commun 13, 6510 (2022). [PubMed: 36347863]

121. Pandarinath C et al. Latent factors and dynamics in motor cortex and their application to brain–machine interfaces. J. Neurosci 38, 9390–9401 (2018). [PubMed: 30381431]

122. Parrell B & Houde J Modeling the role of sensory feedback in speech motor control and learning. J. Speech Lang. Hear. Res 62, 2963–2985 (2019). [PubMed: 31465712]

123. Houde J & Nagarajan S Speech production as state feedback control. Front. Hum. Neurosci 10.3389/fnhum.2011.00082 (2011).

124. Sitaram R et al. Closed-loop brain training: the science of neurofeedback. Nat. Rev. Neurosci 18, 86–100 (2017). [PubMed: 28003656]

125. Wairagkar M, Hochberg LR, Brandman DM & Stavisky SD Synthesizing speech by decoding intracortical neural activity from dorsal motor cortex. In 2023 11th Int. IEEE/EMBS Conf. Neural Engineering (NER) 10.1109/NER52421.2023.10123880 (IEEE, 2023).

126. Casanova E et al. YourTTS: towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In Proc. 39th Int. Conf. Machine Learning (eds Chaudhuri K et al.) Vol. 162, 2709–2720 (PMLR, 2022).

127. Peters B, O’Brien K & 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. Disabil. Rehabil. Assist. Technol 10.1080/17483107.2022.2149866 (2022).

128. Wu P, Watanabe S, Goldstein L, Black AW & Anumanchipalli GK Deep speech synthesis from articulatory representations. In Proc. Interspeech 2022, 779–783 (2022). 10.21437/Interspeech.2022-10892.

129. Cho CJ, Wu P, Mohamed A & Anumanchipalli GK 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, 2023). 10.1109/icassp49357.2023.10094711.

130. Mehrabian A Silent Messages: Implicit Communication of Emotions and Attitudes (Wadsworth, 1981).

131. Jia J, Wang X, Wu Z, Cai L & Meng H Modeling the correlation between modality semantics and facial expressions. In Proc. 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference 1–10 (2012).

132. Sumby WH & Pollack I Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am 26, 212–215 (1954).

133. Branco MP et al. Brain–computer interfaces for communication: preferences of individuals with locked-in syndrome. Neurorehabil. Neural Repair 35, 267–279 (2021). [PubMed: 33530868]

134. Patterson JR & Grabois M Locked-in syndrome: a review of 139 cases. Stroke 17, 758–764 (1986). [PubMed: 3738962]

135. Tomik B & Guiloff RJ Dysarthria in amyotrophic lateral sclerosis: a review. Amyotroph. Lateral Scler 11, 4–15 (2010). [PubMed: 20184513]

136. Thomas TM et al. Decoding articulatory and phonetic components of naturalistic continuous speech from the distributed language network. J. Neural Eng 20, 046030 (2023).

137. Flinker A et al. Redefining the role of Broca's area in speech. Proc. Natl Acad. Sci. USA 112, 2871–2875 (2015). [PubMed: 25730850]

138. Cogan GB et al. Sensory–motor transformations for speech occur bilaterally. Nature 507, 94–98 (2014). [PubMed: 24429520]

139. Rainey S, Martin S, Christen A & Mégevand P & Fourneret E Brain recording, mind-reading, and neurotechnology: ethical issues from consumer devices to brain-based speech decoding. Sci. Eng. Ethics 26, 2295–2311 (2020). [PubMed: 32356091]

140. Nip I & Roth CR in Encyclopedia of Clinical Neuropsychology (eds Kreutzer J, DeLuca J & Caplan B) (Springer International, 2017).

141. Xiong W et al. Toward human parity in conversational speech recognition. IEEEACM Trans. Audio Speech Lang. Process 25, 2410–2423 (2017).

142. Munteanu C, Penn G, Baecker R, Toms E & James D Measuring the acceptable word error rate of machine-generated webcast transcripts. In Interspeech 2006 10.21437/Interspeech.2006-40 (2006).

143. Panayotov V, Chen G, Povey D & Khudanpur S Librispeech: an ASR corpus based on public domain audio books. In 2015 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP) 10.1109/ICASSP.2015.7178964 (IEEE, 2015).

144. Godfrey JJ, Holliman EC & McDaniel J SWITCHBOARD: telephone speech corpus for research and development. In Proc. ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing Vol. 1, 517–520 (1992).

145. OpenAI. GPT-4 Technical Report. Preprint at https://arxiv.org/abs/2303.08774 (2023).

146. Trnka K, Yarrington D, McCaw J, McCoy KF & Pennington C The effects of word prediction on communication rate for AAC. In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers 173–176 (Association for Computational Linguistics, 2007).

147. Venkatagiri H Effect of window size on rate of communication in a lexical prediction AAC system. Augment. Altern. Commun 10, 105–112 (1994).

148. Trnka K, Mccaw J, Mccoy K & Pennington C in Human Language Technologies 2007 173–176 (2008).

149. Kayte SN, Mal M, Gaikwad S & Gawali B Performance evaluation of speech synthesis techniques for English language. In Proc. Int. Congress on Information and Communication Technology (eds Satapathy SC, Bhatt YC, Joshi A & Mishra DK) 253–262 10.1007/978-981-10-0755-2_27 (Springer, 2016).

150. Wagner P et al. Speech synthesis evaluation — state-of-the-art assessment and suggestion for a novel research program. In 10th ISCA Workshop on Speech Synthesis (SSW 10) 10.21437/SSW.2019-19 (ISCA, 2019).

151. Kubichek R Mel-cepstral distance measure for objective speech quality assessment. In Proc. IEEE Pacific Rim Conf. Communications Computers and Signal Processing Vol. 1, 125–128 (1993).

152. Varshney S, Farias D, Brandman DM, Stavisky SD & Miller LM Using automatic speech recognition to measure the intelligibility of speech synthesized from brain signals. In 2023 11th

Int. IEEE/EMBS Conf. Neural Engineering (NER) 10.1109/NER52421.2023.10123751 (IEEE, 2023).

153. Radford A et al. Robust speech recognition via large-scale weak supervision. Preprint at http://arxiv.org/abs/2212.04356 (2022).

154. Yates AJ Delayed auditory feedback. Psychol. Bull 60, 213–232 (1963). [PubMed: 14002534]

155. Zanette D Statistical patterns in written language. Preprint at https://arxiv.org/abs/1412.3336v1 (2014).

156. Adolphs S & Schmitt N Lexical coverage of spoken discourse. Appl. Linguist 24, 425–438 (2003).

157. Laureys S et al. The locked-in syndrome: what is it like to be conscious but paralyzed and voiceless? in Progress in Brain Research Vol. 150 (ed. Laureys S) 495–611 (Elsevier, 2005). [PubMed: 16186044]

158. Peters B et al. Brain–computer interface users speak up: the Virtual Users' Forum at the 2013 International Brain–Computer Interface Meeting. Arch. Phys. Med. Rehabil 96, S33–S37 (2015). [PubMed: 25721545]

159. Huggins JE, Wren PA & Gruis KL What would brain–computer interface users want? Opinions and priorities of potential users with amyotrophic lateral sclerosis. Amyotroph. Lateral Scler 12, 318–324 (2011). [PubMed: 21534845]

160. Kreuzberger D, Kühl N & Hirschl S Machine learning operations (MLOps): overview, definition, and architecture. IEEE Access. 11, 31866–31879 (2023).

161. Gordon EM et al. A somato-cognitive action network alternates with effector regions in motor cortex. Nature 10.1038/s41586-023-05964-2 (2023).

162. Degenhart AD et al. Remapping cortical modulation for electrocorticographic brain–computer interfaces: a somatotopy-based approach in individuals with upper-limb paralysis. J. Neural Eng 15, 026021 (2018). [PubMed: 29160240]

163. Kikkert S, Pfyffer D, Verling M, Freund P & Wenderoth N Finger somatotopy is preserved after tetraplegia but deteriorates over time. eLife 10, e67713 (2021). [PubMed: 34665133]

164. Bruurmijn MLCM, Pereboom IPL, Vansteensel MJ, Raemaekers MAH & Ramsey NF Preservation of hand movement representation in the sensorimotor areas of amputees. Brain 140, 3166–3178 (2017). [PubMed: 29088322]

165. Guenther FH Neural Control of Speech (MIT Press, 2016).

166. Castellucci GA, Kovach CK, Howard MA, Greenlee JDW & Long MA A speech planning network for interactive language use. Nature 602, 117–122 (2022). [PubMed: 34987226]

167. Murphy E et al. The spatiotemporal dynamics of semantic integration in the human brain. Nat. Commun 14, 6336 (2023). [PubMed: 37875526]

168. Ozker M, Doyle W, Devinsky O & Flinker A A cortical network processes auditory error signals during human speech production to maintain fluency. PLOS Biol. 20, e3001493 (2022). [PubMed: 35113857]

169. Quirarte JA et al. Language supplementary motor area syndrome correlated with dynamic changes in perioperative task-based functional MRI activations: case report. J. Neurosurg 134, 1738–1742 (2020). [PubMed: 32502992]

170. Bullock L, Forseth KJ, Woolnough O, Rollo PS & Tandon N Supplementary motor area in speech initiation: a large-scale intracranial EEG evaluation of stereotyped word articulation. Preprint at bioRxiv 10.1101/2023.04.04.535557 (2023).

171. Oby ER et al. New neural activity patterns emerge with long-term learning. Proc. Natl Acad. Sci. USA 116, 15210–15215 (2019). [PubMed: 31182595]

172. Luu TP, Nakagome S, He Y & Contreras-Vidal JL Real-time EEG-based brain–computer interface to a virtual avatar enhances cortical involvement in human treadmill walking. Sci. Rep 7, 8895 (2017). [PubMed: 28827542]

173. Alimardani M et al. Brain–Computer Interface and Motor Imagery Training: The Role of Visual Feedback and Embodiment. Evolving BCI Therapy — Engaging Brain State Dynamics10.5772/intechopen.78695 (IntechOpen, 2018).

174. Orsborn AL et al. Closed-loop decoder adaptation shapes neural plasticity for skillful neuroprosthetic control. Neuron 82, 1380–1393 (2014). [PubMed: 24945777]

175. Muller L et al. Thin-film, high-density micro-electrocorticographic decoding of a human cortical gyrus. In 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 10.1109/EMBC.2016.7591001 (2016).

176. Duraivel S et al. High-resolution neural recordings improve the accuracy of speech decoding. Nat. Commun 14, 6938 (2023). [PubMed: 37932250]

177. Kaiju T, Inoue M, Hirata M & Suzuki T High-density mapping of primate digit representations with a 1152-channel μECoG array. J. Neural Eng 18, 036025 (2021).

178. Woods V et al. Long-term recording reliability of liquid crystal polymer μECoG arrays. J. Neural Eng 15, 066024 (2018). [PubMed: 30246690]

179. Rachinskiy I et al. High-density, actively multiplexed μECoG array on reinforced silicone substrate. Front. Nanotechnol 10.3389/fnano.2022.837328 (2022).

180. Sun J et al. Intraoperative microseizure detection using a high-density micro-electrocorticography electrode array. Brain Commun. 4, fcac122 (2022). [PubMed: 35663384]

181. Ho E et al. The layer 7 cortical interface: a scalable and minimally invasive brain–computer interface platform. Preprint at bioRxiv 10.1101/2022.01.02.474656 (2022).

182. Oxley TJ et al. Motor neuroprosthesis implanted with neurointerventional surgery improves capacity for activities of daily living tasks in severe paralysis: first in-human experience. J. NeuroIntervent. Surg 13, 102–108 (2021).

183. Chen R, Canales A & Anikeeva P Neural recording and modulation technologies. Nat. Rev. Mater 2, 1–16 (2017).

184. Hong G & Lieber CM Novel electrode technologies for neural recordings. Nat. Rev. Neurosci 20, 330–345 (2019). [PubMed: 30833706]

185. Sahasrabuddhe K et al. The Argo: a high channel count recording system for neural recording in vivo. J. Neural Eng 18, 015002 (2021). [PubMed: 33624614]

186. Musk E & Neuralink. An integrated brain–machine interface platform with thousands of channels. J. Med. Internet Res 21, e16194 (2019). [PubMed: 31642810]

187. Paulk AC et al. Large-scale neural recordings with single neuron resolution using neuropixels probes in human cortex. Nat. Neurosci 25, 252–263 (2022). [PubMed: 35102333]

188. Chung JE et al. High-density single-unit human cortical recordings using the neuropixels probe. Neuron 110, 2409–2421.e3 (2022). [PubMed: 35679860]

189. Kingma DP & Welling M An introduction to variational autoencoders. Found. Trends Mach. Learn 12, 307–392 (2019).

190. Schneider S, Lee JH & Mathis MW Learnable latent embeddings for joint behavioural and neural analysis. Nature 617, 360–368 (2023). [PubMed: 37138088]

191. Liu R et al. Drop, swap, and generate: a self-supervised approach for generating neural activity. Preprint at http://arxiv.org/abs/2111.02338 (2021).

192. Cho CJ, Chang E & Anumanchipalli G Neural latent aligner: cross-trial alignment for learning representations of complex, naturalistic neural data. In Proc. 40th Int. Conf. Machine Learning 5661–5676 (PMLR, 2023).

193. Keshtkaran MR et al. A large-scale neural network training framework for generalized estimation of single-trial population dynamics. Nat. Methods 19, 1572–1577 (2022). [PubMed: 36443486]

194. Berezutskaya J et al. Direct speech reconstruction from sensorimotor brain activity with optimized deep learning models. J. Neural Eng 20, 056010 (2023).

195. Touvron H et al. LLaMA: Open and Efficient Foundation Language Models. Preprint at 10.48550/arXiv.2302.13971 (2023).

196. Graves A Sequence transduction with recurrent neural networks. Preprint at 10.48550/arXiv.1211.3711 (2012).

197. Shi Y et al. Emformer: efficient memory transformer based acoustic model for low latency streaming speech recognition. Preprint at 10.48550/arXiv.2010.10759 (2020).

198. Rapeaux AB & Constandinou TG Implantable brain machine interfaces: first-in-human studies, technology challenges and trends. Curr. Opin. Biotechnol 72, 102–111 (2021). [PubMed: 34749248]

199. Matsushita K et al. A fully implantable wireless ECoG 128-channel recording device for human brain–machine interfaces: W-HERBS. Front. Neurosci 12, 511 (2018). [PubMed: 30131666]

200. Cajigas I et al. Implantable brain–computer interface for neuroprosthetic-enabled volitional hand grasp restoration in spinal cord injury. Brain Commun. 3, fcab248 (2021). [PubMed: 34870202]

201. Jarosiewicz B & Morrell M The RNS system: brain-responsive neurostimulation for the treatment of epilepsy. Expert Rev. Med. Dev 18, 129–138 (2021).

202. Lorach H et al. Walking naturally after spinal cord injury using a brain–spine interface. Nature 618, 126–133 (2023). [PubMed: 37225984]

203. Weiss JM, Gaunt RA, Franklin R, Boninger ML & Collinger JL Demonstration of a portable intracortical brain–computer interface. Brain-Comput. Interfaces 6, 106–117 (2019).

204. Kim JS, Kwon SU & Lee TG Pure dysarthria due to small cortical stroke. Neurology 60, 1178–1180 (2003). [PubMed: 12682329]

205. Urban PP et al. Left-hemispheric dominance for articulation: a prospective study on acute ischaemic dysarthria at different localizations. Brain 129, 767–777 (2006). [PubMed: 16418180]

206. Wu P et al. Speaker-independent acoustic-to-articulatory speech inversion. Preprint at 10.48550/arXiv.2302.06774 (2023).

207. Oppenheim AV, Schafer RW & Schafer RW Discrete-Time Signal Processing (Pearson, 2014).

208. Kim JW, Salamon J, Li P & Bello JP CREPE: a convolutional representation for pitch estimation. Preprint at 10.48550/arXiv.1802.06182 (2018).

209. Park K & Kim J g2pE. Github https://github.com/Kyubyong/g2p (2019).

210. Duffy JR Motor Speech Disorders: Substrates, Differential Diagnosis, and Management (Elsevier Health Sciences, 2019).

211. Basilakos A, Rorden C, Bonilha L, Moser D & Fridriksson J Patterns of poststroke brain damage that predict speech production errors in apraxia of speech and aphasia dissociate. Stroke 46, 1561–1566 (2015). [PubMed: 25908457]

212. Berthier ML Poststroke aphasia: epidemiology, pathophysiology and treatment. Drugs Aging 22, 163–182 (2005). [PubMed: 15733022]

213. Wilson SM et al. Recovery from aphasia in the first year after stroke. Brain 146, 1021–1039 (2022).

214. Marzinske M Help for speech, language disorders. Mayo Clinic Health System https://www.mayoclinichealthsystem.org/hometown-health/speaking-of-health/help-is-available-for-speech-and-language-disorders (2022).

215. Amyotrophic lateral sclerosis. CDC https://www.cdc.gov/als/WhatisALS.html (CDC, 2022).

216. Sokolov A Inner Speech and Thought (Springer Science & Business Media, 2012).

217. Alderson-Day B & Fernyhough C Inner speech: development, cognitive functions, phenomenology, and neurobiology. Psychol. Bull 141, 931–965 (2015). [PubMed: 26011789]

218. Sankaran N, Moses D, Chiong W & Chang EF Recommendations for promoting user agency in the design of speech neuroprostheses. Front. Hum. Neurosci 17, 1298129 (2023). [PubMed: 37920562]

219. Sun X & Ye B The functional differentiation of brain–computer interfaces (BCIs) and its ethical implications. Humanit. Soc. Sci. Commun 10, 1–9 (2023).

220. Ienca M, Haselager P & Emanuel EJ Brain leaks and consumer neurotechnology. Nat. Biotechnol 36, 805–810 (2018). [PubMed: 30188521]

221. Yuste R Advocating for neurodata privacy and neurotechnology regulation. Nat. Protoc 18, 2869–2875 (2023). [PubMed: 37697107]

222. Kamal AH et al. A person-centered, registry-based learning health system for palliative care: a path to coproducing better outcomes, experience, value, and science. J. Palliat. Med 21, S–61 (2018).

223. Alford J The multiple facets of co-production: building on the work of Elinor Ostrom. Public. Manag. Rev 16, 299–316 (2014).

224. Institute of Medicine (US) Roundtable on Value & Science-Driven Health Care. Clinical Data as the Basic Staple of Health Learning: Creating and Protecting a Public Good: Workshop Summary (National Academies Press, 2011).

Box 1

Disorders of speech and language

In discussing different types of speech and language disorders that may be treated with a speech neuroprosthesis, it is helpful to simplify the speech-production process to a few representational levels (although more detailed levels may exist and we do not suggest that these levels occur in complete isolation from one another). These include accessing conceptual language representations, generating and coordinating the proper vocal-tract motor plans to reach intended speech targets, executing these motor plans, transmitting these signals through the brainstem and peripheral nerves and properly contracting the corresponding vocal-tract muscles.

Dysarthria is a speech-production disorder in which the ability of an individual to control their vocal-tract articulators is degraded but their ability to comprehend speech and language is preserved[210]. Anarthria is the most severe form of dysarthria and refers to the case in which the ability to coordinate articulation is fully lost[140]. Dysarthria can result from an injury that disrupts one of the final three stages of the speech-production process (executing the motor plan onwards). Injuries that selectively impair the ability of an individual to contract their vocal-tract muscles, such as a laryngectomy or facial nerve injury, may cause dysarthria. Another common cause of dysarthria, and often anarthria, is injury to the descending motor pathways in the brainstem. Past speech-neuroprosthesis studies have generally focused on individuals with this type of injury, often owing to amyotrophic lateral sclerosis, a neurodegenerative disease that selectively targets motor neurons, or to a stroke that causes irreversible injury to the brainstem tissue. In these cases, speech is rarely the only motor function lost, with descending motor neurons that innervate the upper and lower limbs also often injured. Individuals who have lost all motor control, except for certain eye or head movements, are referred to as being ‘locked in’[2]. Finally, dysarthria may be caused by injury to the motor cortex itself, referred to as cortical anarthria or dysarthria[204,205]. In such cases, the motor cortex cannot implement the proper motor commands and transmit them to the peripheral nerves to contract the vocal-tract articulators.

Beyond anarthria and dysarthria, apraxia of speech (AOS) and aphasia are disorders of speech and language, respectively, that may be treated with a speech neuroprosthesis in the future. AOS is a disorder of speech-motor planning and coordination in which articulatory muscle strength and language comprehension are preserved[210,211]. In contrast to dysarthria, AOS is characterized by a deficit in generating and coordinating complex speech-motor sequences rather than a direct deficit in motor control of vocal-tract movements. Individuals with AOS tend to have inconsistent speech errors that occur more often as the complexity of an utterance increases, whereas individuals with dysarthria are consistently unable to finely control their vocal tract and orofacial movements[210]. AOS often occurs with lesions to the language-dominant left hemisphere[212], and recent studies have localized AOS to damage in the middle precentral gyrus[48–50,211]. Language disorders, or aphasias, affect the ability of an individual to comprehend or express speech, disrupting the conceptual-language system of the brain, and generally result from injury to cortical regions in the language-dominant left hemisphere[213]. Treatment of AOS or aphasia has not been explicitly explored with a neuroprosthesis. Future studies should investigate whether intact cortical regions in individuals with aphasia and AOS contain representations that may be decoded into intended speech. Although AOS (without coincident aphasia) disrupts the ability of an individual to implement and plan coordinated speech-articulatory movements, language representations (lexico-semantics and syntax), and probably representations corresponding to intended speech targets, remain intact[210]. For aphasia, the nature of preserved representations is less consistent across individuals and may depend on whether the ability of an individual to comprehend or express speech is impacted. Approaches to restore speech in individuals with AOS or aphasia is likely to require electrode arrays to be placed in areas beyond the motor cortex.

Given the broad range of clinical disorders that may be treated with a speech neuroprosthesis, it is challenging to define the number of individuals who may benefit from the technology. An estimated 5–10% of people in the USA live with a communication disorder and approximately 1 million people in the USA live with aphasia specifically[214]. According to the US Centers for Disease Control, each year there are an estimated 5,000 newly diagnosed cases of amyotrophic lateral sclerosis[215]. However, the worldwide proportion of these cases in which a speech neuroprosthesis may be needed or desired is unclear. Thus, although speech neuroprostheses have the potential to improve quality of life for many individuals, future work is needed to precisely define the user population. Beyond measuring the incidence of conditions resulting in communication disorders, this should involve estimating the proportion of cases in which intact cortical representations persist that can be decoded into intended speech. Furthermore, the performance and medical risk at which an individual would be willing to use the technology, given different severity levels of underlying communication impairment, should be estimated through survey. These efforts should go hand in hand with work to better solicit feedback and design considerations from user populations (Box 2).

Box 2

Ethical considerations for the development of speech neuroprostheses

Speech neuroprostheses have progressed rapidly over the past decade, demonstrating immense potential to restore agency, expression and autonomy to individuals with severe paralysis. Although these research advances have given hope to individuals with severe paralysis, they have also raised important ethical concerns for researchers to address primarily surrounding mental and data privacy.

A contributing factor to ethical concerns held by individuals with severe paralysis is how advances in speech neuroprostheses are communicated to the public through the media. Given that neuroprostheses and human–machine interactions are prevalent in popular culture, the presentation of current scientific work can be misunderstood. Indeed, some recent speech neuroprostheses advances[14,15] have been portrayed as mind reading — the reading out of internal thoughts or inner monologue that one did not intend to make public (referred to as internal speech in this Review) — when, in reality, they rely on volitional speech attempts by the user and not internal speech. Thus, researchers have the responsibility to clearly convey their work to help address mental privacy concerns, and we hope that standardizing speech instruction reporting will help towards this end.

However, mind reading is still a pervasive ethical concern. In this Review, we distinguish internal speech from intended speech, as intended speech is either matched by a behaviour, such as trying to mime (silently attempted speech) or trying to speak aloud (attempted speech), or the user is imagining that they are saying or hearing their intended speech output. Confusing these concepts is understandable, given that our internal thoughts and ideas can be conceptualized through language, and language is conveyed with speech[216]. Current speech neuroprostheses do not target internal speech representations, as their goal has been to decode volitional and intended speech from users with paralysis. In addition, current speech-decoding approaches are rooted in years of work understanding intact speech production, specifically the volitional control of articulation encoded on the sensorimotor cortex. By contrast, the understanding of how the brain represents internal monologue and thoughts is relatively limited[217]. However, although the field may not currently be close to decoding internal thoughts and monologue, it is possible that this could change in the future. A potential safeguard would be to chain any speech decoding to an initial speech-detection module, trained to identify volitional speech attempts[218]. Such a model has been highly successful in preliminary work and could be augmented to explicitly distinguish types of intended speech from types of internal monologue and thought.

In addition to mind reading, the rapid expansion of speech neuroprostheses has raised data-related ethical concerns over ownership[219], leaks[220] and privacy[221] of one’s neural activity. For example, it is not currently clear what rights speech-neuroprosthesis users will have over their neural activity data. It is critical that the future of speech neuroprostheses is driven by constant communication and collaboration between engineers, physicians, neuroscientists, neuroethicists, regulators, legislators and, importantly, current and potential users and their advocacy groups. Value-aligned development, in which innovation is guided by user input, has proven successful in many health-care fields by creating more sustainable and efficient co-designed systems[222–224]. Soliciting consistent feedback from users may optimize the development of desired speech neuroprostheses features and encourage long-term and widespread system adoption.

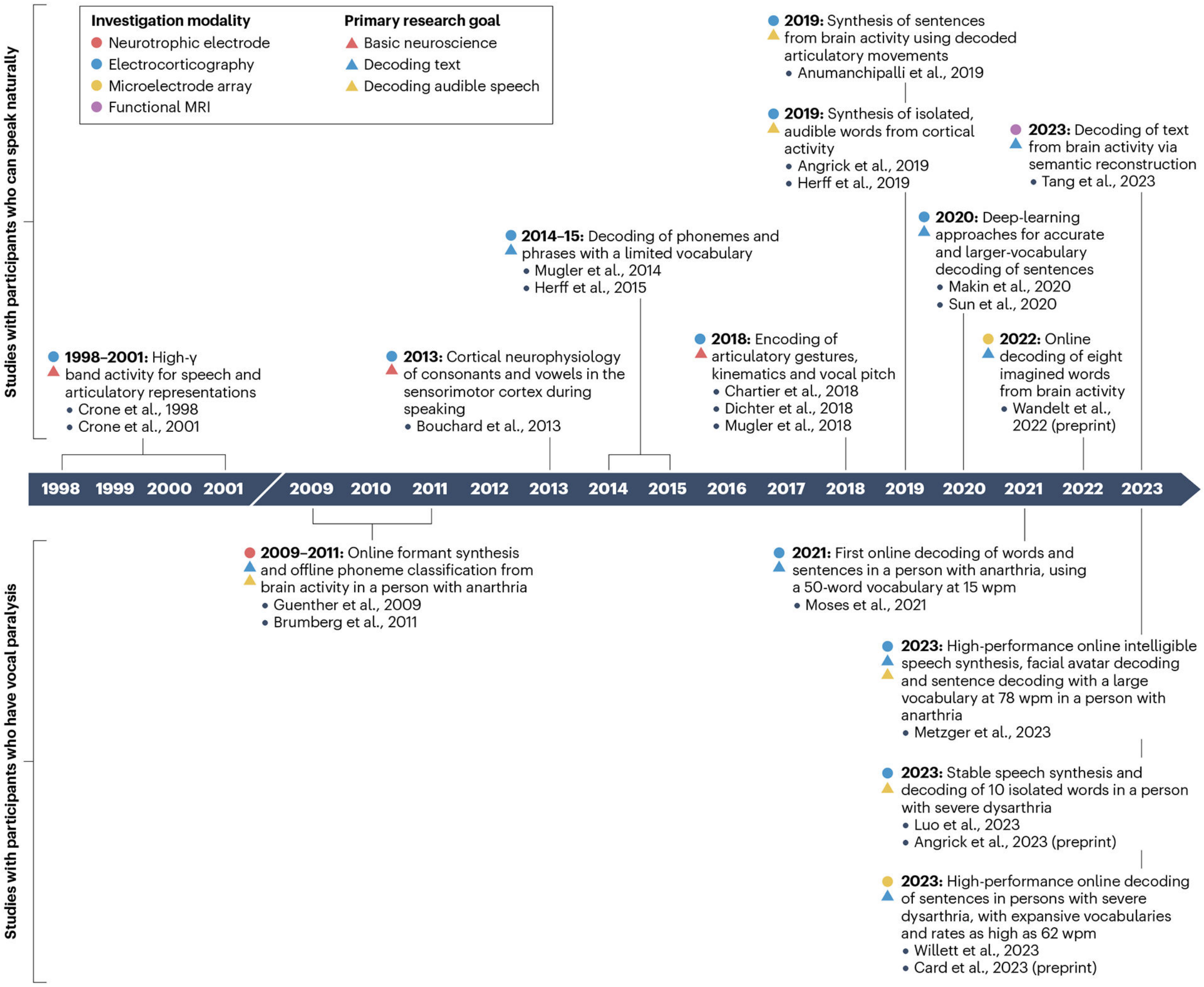
Investigation modality
● Neurotrophic electrode
● Electrocorticography
● Microelectrode array
● Functional MRI
Primary research goal
▲ Basic neuroscience
▲ Decoding text
▲ Decoding audible speech
Studies with participants who can speak naturally
1998–2001: High-γ band activity for speech and articulatory representations
• Crone et al., 1998
• Crone et al., 2001
2013: Cortical neurophysiology of consonants and vowels in the sensorimotor cortex during speaking
• Bouchard et al., 2013
2014–15: Decoding of phonemes and phrases with a limited vocabulary
• Mugler et al., 2014
• Herff et al., 2015
2018: Encoding of articulatory gestures, kinematics and vocal pitch
• Chartier et al., 2018
• Dichter et al., 2018
• Mugler et al., 2018
2019: Synthesis of sentences from brain activity using decoded articulatory movements
• Anumanchipalli et al., 2019
2019: Synthesis of isolated, audible words from cortical activity
• Angrick et al., 2019
• Herff et al., 2019
2020: Deep-learning approaches for accurate and larger-vocabulary decoding of sentences
• Makin et al., 2020
• Sun et al., 2020
2022: Online decoding of eight imagined words from brain activity
• Wandelt et al., 2022 (preprint)
2023: Decoding of text from brain activity via semantic reconstruction
• Tang et al., 2023
1998 1999 2000 2001 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023
Studies with participants who have vocal paralysis
2009–2011: Online formant synthesis and offline phoneme classification from brain activity in a person with anarthria
• Guenther et al., 2009
• Brumberg et al., 2011
2021: First online decoding of words and sentences in a person with anarthria, using a 50-word vocabulary at 15 wpm
• Moses et al., 2021
2023: High-performance online intelligible speech synthesis, facial avatar decoding and sentence decoding with a large vocabulary at 78 wpm in a person with anarthria
• Metzger et al., 2023
2023: Stable speech synthesis and decoding of 10 isolated words in a person with severe dysarthria
• Luo et al., 2023
• Angrick et al., 2023 (preprint)
2023: High-performance online decoding of sentences in persons with severe dysarthria, with expansive vocabularies and rates as high as 62 wpm
• Willett et al., 2023
• Card et al., 2023 (preprint)

Fig. 1 |. Key milestones in speech decoding.

Timeline of key advancements that have ultimately led to proof-of-concept speech neuroprostheses for individuals with paralysis. Advancements are labelled based on their study population, neural-recording technology and research goal[12–17,24,43,44,64,68,84,85,88,89,95,106–112].

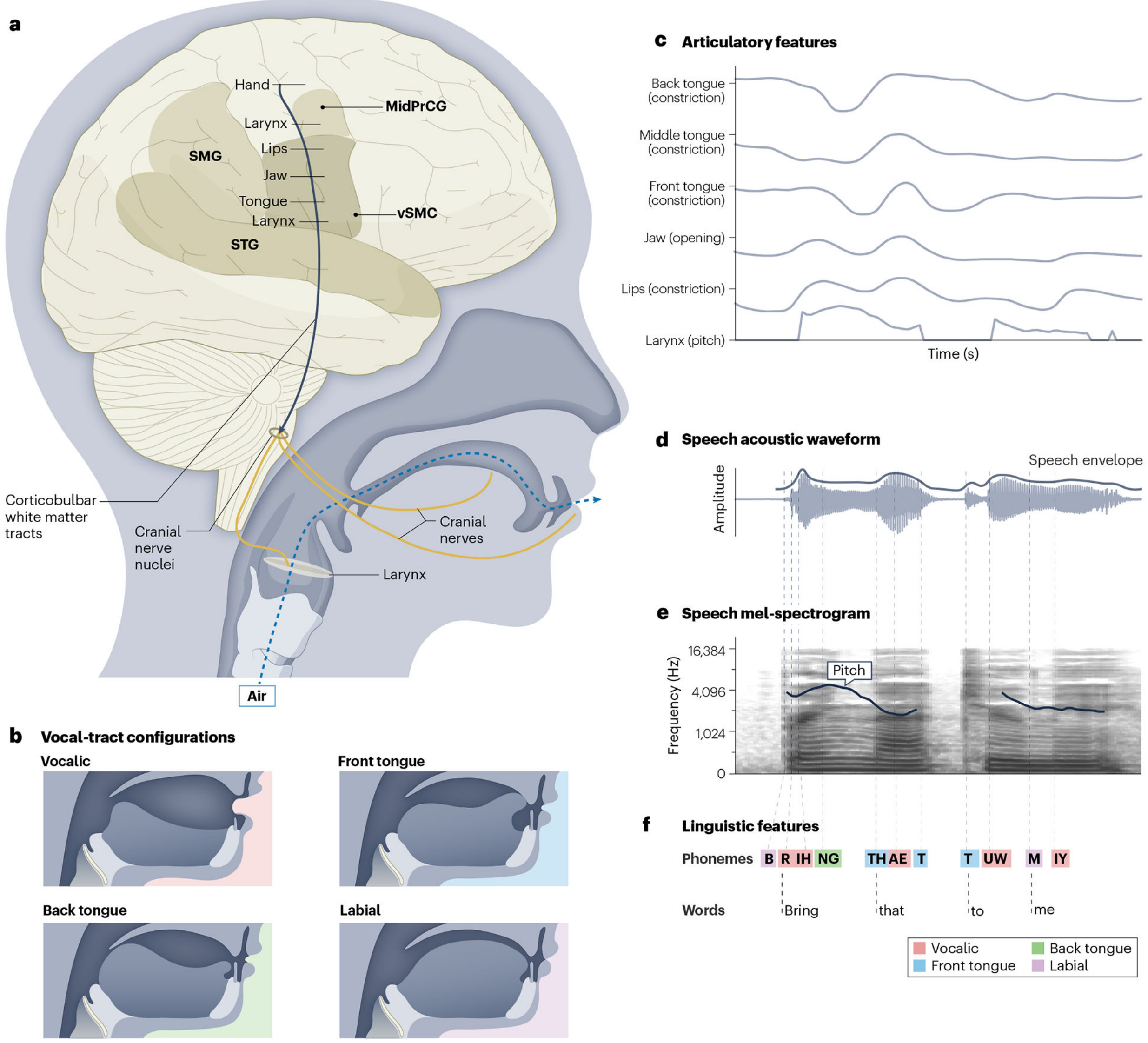

Fig. 2 |. Articulatory control of speech.

Speech articulation relies on the corticobulbar system. At a broad level, this system is composed of cortical neuronal populations that project axons to the brainstem, where activations are relayed through cranial nerves to the speech articulators (muscles). **a**, Neural populations, arranged somatotopically on the ventral sensorimotor cortex (vSMC) and middle precentral gyrus (midPrCG), control the movements of key vocal-tract articulators, such as the larynx, tongue, jaw and lips. These neural populations may receive input from other regions involved in speech, a few of which are highlighted (superior temporal gyrus (STG) and the supramarginal gyrus (SMG)). The vSMC and midPrCG send motor-control signals through their axons, which bundle and form the corticobulbar white matter tract, that terminate in cranial nerve nuclei in the brainstem. Neurons in cranial nerve nuclei then send axons, which bundle and form cranial nerves, that innervate the speech articulators (larynx, tongue, jaw and lips). **b**, Cortical-activity patterns, ultimately transmitted by the cranial nerves, lead to contraction of the vocal-tract articulators and defined vocal-tract configurations that can broadly be grouped based on the place of air constriction into four classes: vocalic, back tongue, front tongue and labial. **c**, Continuous movements of the vocal-tract articulators between these configurations, along with air from respiratory structures, turn neural activity related to intended speech into vocalized sound waves. Continuous articulatory features can be measured for different landmarks in the vocal tract over time. Here, the visualized articulatory features are inferred from the produced acoustic waveform[206]. **d**, The produced speech can be represented as an acoustic waveform (amplitude over time). The envelope can be estimated from the acoustic waveform[207] and represents the intensity of speech over time, an important measurement that correlates with speech rate, stress patterns and loudness **e**, Speech can also be represented as a mel-spectrogram, in which the power of different perceptually salient frequency bands is shown over time. Pitch can be computed by computational algorithms that estimate the fundamental frequency of a signal[208]. **f**, Defined patterns in the produced sound, visible on a mel-spectrogram, form the basis of meaning. Phonemes are a type of linguistic feature and refer to the smallest perceptually distinct units of sound (epochs denoted by dotted lines in panels **d** and **e**) that form a language. Phonemes, along with words, can be annotated and inferred based on the produced sound during continuous speech[209]. In addition, the vocal-tract configuration that gives rise to a distinct unit of sound can be used to group phonemes. Visualization of articulatory control of speech in panels **c**–**f** from produced sound was created using algorithms from refs. 206–209.

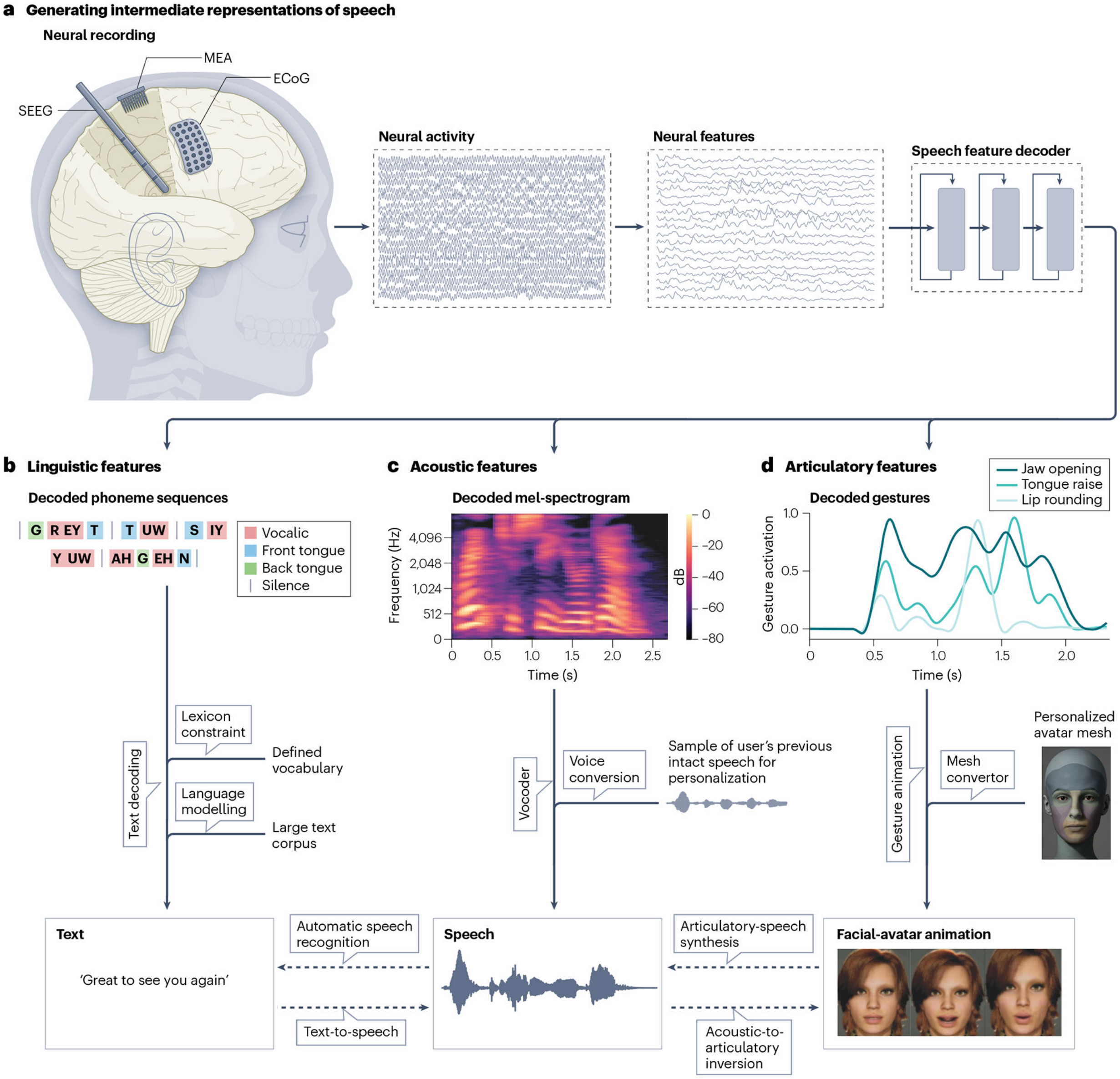

Fig. 3 |. Decoding speech from neural activity.

Speech-decoding systems follow a similar heuristic; neural activity during intended speech is captured with an interface of choice and relevant features are extracted and processed by a decoding model. This decoding model can be trained to transform neural activity into text, audible speech or orofacial movements. **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 (colours indicate their vocal-tract configuration), and then a defined vocabulary (via a lexicon constraint) 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 (gesture activation). 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). Visualization of acoustic and articulatory features in panels **b–d** was created using data and algorithms from ref. 14. The personalized avatar mesh animations in panel **d** were generated using Unreal Engine animation software (Speech Graphics, Edinburgh, UK). Panel **d** adapted from ref. 14, Springer Nature Limited.

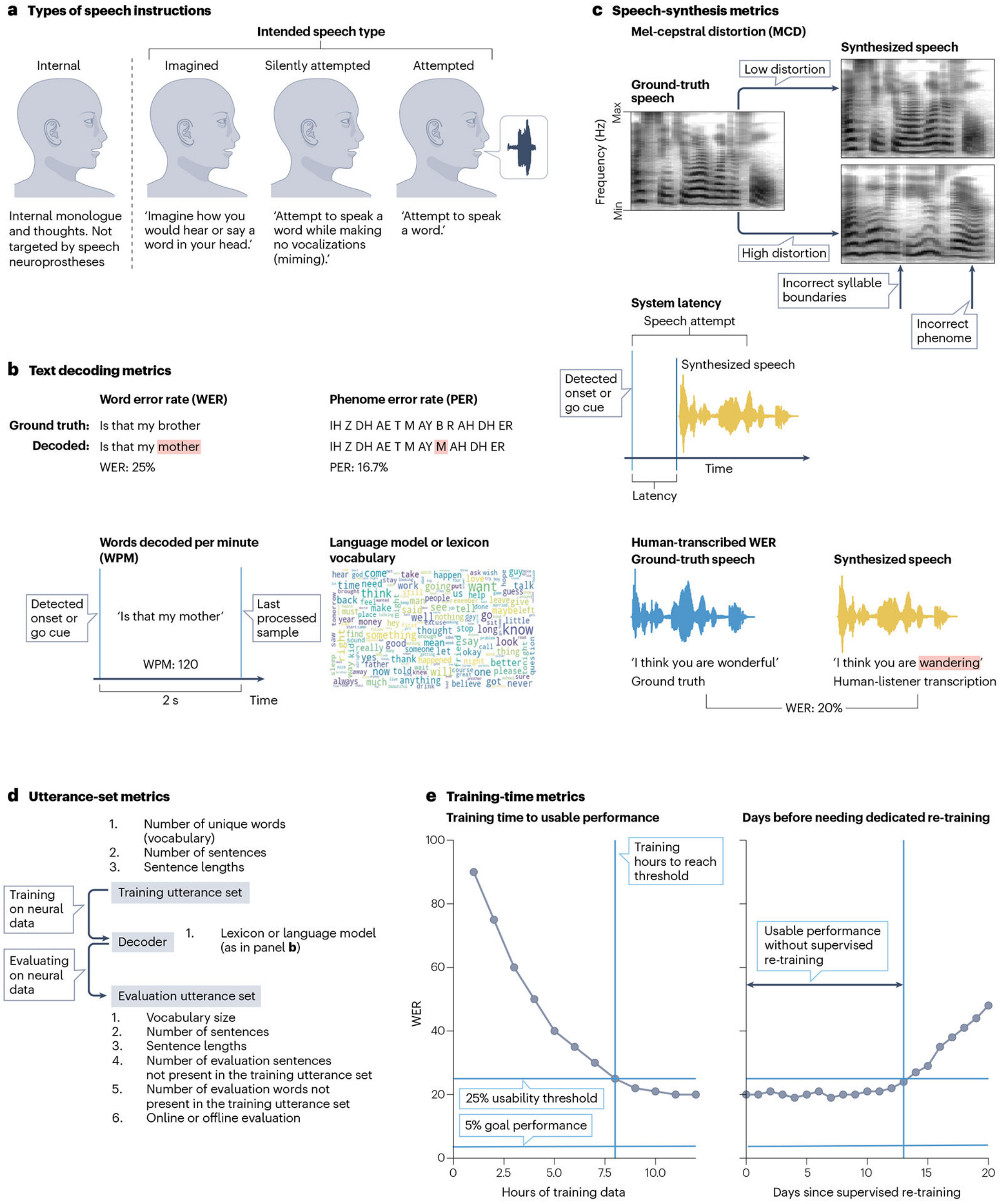
a Types of speech instructions
Intended speech type
Internal
Imagined
Silently attempted
Attempted
Internal monologue and thoughts. Not targeted by speech neuroprostheses
'Imagine how you would hear or say a word in your head.'
'Attempt to speak a word while making no vocalizations (miming).'
'Attempt to speak a word.'
b Text decoding metrics
Word error rate (WER)
Phenome error rate (PER)
Ground truth: Is that my brother
Decoded: Is that my mother
WER: 25%
IH Z DH AE T M AY B R AH DH ER
IH Z DH AE T M AY M AH DH ER
PER: 16.7%
Words decoded per minute (WPM)
Detected onset or go cue
'Is that my mother'
Last processed sample
WPM: 120
2 s
Time
Language model or lexicon vocabulary
c Speech-synthesis metrics
Mel-cepstral distortion (MCD)
Synthesized speech
Low distortion
Ground-truth speech
Frequency (Hz)
Max
Min
High distortion
Incorrect syllable boundaries
Incorrect phenome
System latency
Speech attempt
Detected onset or go cue
Synthesized speech
Time
Latency
Human-transcribed WER
Ground-truth speech
Synthesized speech
'I think you are wonderful'
Ground truth
'I think you are wandering'
Human-listener transcription
WER: 20%
d Utterance-set metrics
1. Number of unique words (vocabulary)
2. Number of sentences
3. Sentence lengths
Training on neural data
Training utterance set
Decoder
1. Lexicon or language model (as in panel b)
Evaluating on neural data
Evaluation utterance set
1. Vocabulary size
2. Number of sentences
3. Sentence lengths
4. Number of evaluation sentences not present in the training utterance set
5. Number of evaluation words not present in the training utterance set
6. Online or offline evaluation
e Training-time metrics
Training time to usable performance
Days before needing dedicated re-training
WER
100
80
60
40
20
0
Training hours to reach threshold
25% usability threshold
5% goal performance
0
2.5
5.0
7.5
10.0
Hours of training data
Usable performance without supervised re-training
0
5
10
15
20
Days since supervised re-training

Fig. 4 |. Evaluating and standardizing speech neuroprostheses.

As the field of speech neuroprosthetics continues to accelerate, it is important to define standardized methods of reporting and evaluating speech-decoding results. Here, we propose several metrics to become standardized, specifically relating to speech instructions, decoded outputs, utterance sets used to train and evaluate models and finally user training times. **a**, Common types of speech instructions given to individuals in speech-decoding studies include imagined, silently attempted and attempted speech (right). These intended speech types are distinct from internal thoughts or monologue of an individual (left). **b**, For text decoding, metrics related to the accuracy, speed and lexicon and language model of the system should be reported and standardized. To measure accuracy, the WER and PER (alternatively, the character error rate can be reported if the model is trained to decode sequences of characters) should be used and are defined as the edit distance between ground truth and decoded word and phoneme sequences. To measure speed, the WPM can be reported as the time between the onset of the speech attempt and the final processed sample of neural activity. Finally, the vocabulary size used to define a language model and lexicon should be reported. **c**, For speech synthesis, metrics related to accuracy and latency of the system should be reported and standardized. To measure accuracy, the MCD should be computed between ground truth and synthesized speech, capturing distortion in perceptually salient frequency bands. For a more interpretable measure of performance, the human-transcribed WER should be reported by having volunteers transcribe synthesized speech into text, which can then be compared with ground truth. Finally, the latency of the speech-synthesis system should be reported and can be defined as the time between the onset of a speech attempt and the first sample of synthesized audio that is played back to the participant. Visualization of acoustic waveforms and mel-spectrograms were created using data and algorithms from ref. 14. **d**, Utterance-set metrics, common to text-decoding and speech-synthesis systems, should also be reported. First, qualities of the training-utterance set, such as number of unique words (vocabulary size) and sentences, should be reported. Next, the lexicon and language model vocabulary size (as in panel **b**) should be reported. Finally, key characteristics of the evaluation-utterance set should also be reported. This includes the vocabulary size of, the number of sentences in and the length of sentences in the evaluation-utterance set. Importantly, the overlap between the training and evaluation sets should also be quantified and reported as the number of overlapping words and sentences between the two sets. **e**, Training-time metrics, also common to text-decoding and speech-synthesis systems, should be reported. The total amount of training data quantified as number of hours to reach usable and/or goal performance should be reported. The number of days the system can maintain usable performance, without supervised re-training (either by not re-training or using self-supervised recalibration techniques), should also be reported. Panel **e** adapted from ref. 14, Springer Nature Limited.