Title
Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria.

Journal
The New England journal of medicine, 385(3)

ISSN
0028-4793

Authors
Moses, David A
Metzger, Sean L
Liu, Jessie R
et al.

Publication Date
2021-07-01

DOI
10.1056/nejmoa2027540

Peer reviewed

HHS Public Access

Author manuscript

N Engl J Med. Author manuscript; available in PMC 2022 April 01.

Published in final edited form as:

N Engl J Med. 2021 July 15; 385(3): 217–227. doi:10.1056/NEJMoa2027540.

Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria

David A. Moses, Ph.D.[#], **Sean L. Metzger, M.S.**[#], **Jessie R. Liu, B.S.**[#], **Gopala K. Anumanchipalli, Ph.D.**, **Joseph G. Makin, Ph.D.**, **Pengfei F. Sun, Ph.D.**, **Josh Chartier, Ph.D.**, **Maximilian E. Dougherty, B.A.**, **Patricia M. Liu, M.A.**, **Gary M. Abrams, M.D.**, **Adelyn Tu-Chan, D.O.**, **Karunesh Ganguly, M.D., Ph.D.**, **Edward F. Chang, M.D.**

From the Department of Neurological Surgery (D.A.M., S.L.M., J.R.L., G.K.A., J.G.M., P.F.S., J.C., M.E.D., E.F.C.), the Weill Institute for Neuroscience (D.A.M., S.L.M., J.R.L., G.K.A., J.G.M., P.F.S., J.C., K.G., E.F.C.), and the Departments of Rehabilitation Services (P.M.L.) and Neurology (G.M.A., A.T.-C., K.G.), University of California, San Francisco (UCSF), San Francisco, and the Graduate Program in Bioengineering, University of California, Berkeley–UCSF, Berkeley (S.L.M., J.R.L., E.F.C.).

[#] These authors contributed equally to this work.

Abstract

BACKGROUND—Technology to restore the ability to communicate in paralyzed persons who cannot speak has the potential to improve autonomy and quality of life. An approach that decodes words and sentences directly from the cerebral cortical activity of such patients may represent an advancement over existing methods for assisted communication.

METHODS—We implanted a subdural, high-density, multielectrode array over the area of the sensorimotor cortex that controls speech in a person with anarthria (the loss of the ability to articulate speech) and spastic quadriparesis caused by a brain-stem stroke. Over the course of 48 sessions, we recorded 22 hours of cortical activity while the participant attempted to say individual words from a vocabulary set of 50 words. We used deep-learning algorithms to create computational models for the detection and classification of words from patterns in the recorded cortical activity. We applied these computational models, as well as a natural-language model that yielded next-word probabilities given the preceding words in a sequence, to decode full sentences as the participant attempted to say them.

RESULTS—We decoded sentences from the participant's cortical activity in real time at a median rate of 15.2 words per minute, with a median word error rate of 25.6%. In post hoc analyses, we detected 98% of the attempts by the participant to produce individual words, and we classified words with 47.1% accuracy using cortical signals that were stable throughout the 81-week study period.

Address reprint requests to Dr. Chang at edward.chang@ucsf.edu.

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

A data sharing statement provided by the authors is available with the full text of this article at NEJM.org.

CONCLUSIONS—In a person with anarthria and spastic quadriparesis caused by a brain-stem stroke, words and sentences were decoded directly from cortical activity during attempted speech with the use of deep-learning models and a natural-language model. (Funded by Facebook and others; ClinicalTrials.gov number, NCT03698149.)

ANARTHRIA IS THE LOSS OF THE ABILITY to articulate speech. It can result from a variety of conditions, including stroke and amyotrophic lateral sclerosis.[1] Patients with anarthria may have intact language skills and cognition, and some are able to produce limited oral movements and undifferentiated vocalizations when attempting to speak.[2] However, paralyzed persons may be unable to operate assistive devices because of severe impairment of movement. Anarthria hinders communication with family, friends, and caregivers, thereby reducing patient-reported quality of life.[3] Advances have been made with typing-based brain–computer interfaces that allow speech-impaired persons to spell out messages by controlling a computer cursor.[4–8] However, letter-by-letter selection through interfaces driven by neural signal recordings is slow and effortful. A more efficient and natural approach may be to directly decode whole words from brain areas that control speech. Our understanding of how the area of the sensorimotor cortex that controls speech orchestrates the rapid articulatory movements of the vocal tract has expanded.[9–14] Engineering efforts have used these neurobiologic findings, together with advances in machine learning, to show that speech can be decoded from brain activity in persons without speech impairments.[15–19]

In paralyzed persons who cannot speak, recordings of neural activity cannot be precisely aligned with intended speech because of the absence of speech output, which poses an obstacle for training computational models.[20] In addition, it is unclear whether neural signals underlying speech control are still intact in persons who have not spoken for years or decades. In earlier work, a paralyzed person used an implanted, intracortical, two-channel microelectrode device and an audiovisual interface to generate vowel sounds and phonemes but not full words.[21,22] To determine whether speech can be directly decoded to produce language from the neural activity of a person who is unable to speak, we tested real-time decoding of words and sentences from the cortical activity of a person with limb paralysis and anarthria caused by a brain-stem stroke.

METHODS

STUDY OVERVIEW

This work was performed as part of the BCI Restoration of Arm and Voice (BRAVO) study, which is a single-institution clinical study to evaluate the potential of electrocorticography, a method for recording neural activity from the cerebral cortex with the use of electrodes placed on the surface of the cerebral hemisphere, and custom decoding techniques to enable communication and mobility. An investigational device exemption for the device used in this study was approved by the Food and Drug Administration. As of this writing, the device had been implanted only in the participant described here. Because of regulatory and clinical considerations regarding the proper handling of the percutaneous connector, the participant did not have the opportunity to use the system independently for daily activities but underwent testing at his home.

This work was approved by the Committee on Human Research at the University of California, San Francisco, and was supported in part by a research contract under Facebook's Sponsored Academic Research Agreement. All the authors were involved in the design and execution of the clinical study; the collection, storage, analysis, and interpretation of the data; and the writing of the manuscript. No study hardware or data were transferred to any sponsor, and we did not receive any hardware or software from a sponsor to use in this work. All the authors vouch for the accuracy and completeness of the data and for the fidelity of the study to the protocol (available with the full text of this article at NEJM.org) and confirm that the study was conducted ethically. Informed consent was obtained from the participant after the reason for and nature of implantation and the training procedures and risks were thoroughly explained to him.

PARTICIPANT

The participant was a right-handed man who was 36 years of age at the start of the study. At 20 years of age, he had had an extensive pontine stroke associated with a dissection of the right vertebral artery, which resulted in severe spastic quadriparesis and anarthria, as confirmed by a speech–language pathologist and neurologists (Video 1 and Fig. S1 in the Supplementary Appendix, both available at NEJM.org). His cognitive function was intact, and he had a score of 26 on the Mini–Mental State Examination (scores range from 0 to 30, with higher scores indicating better mental performance); because of his paralysis, it was not physically possible for his score to reach 30. He was able to vocalize grunts and moans but was unable to produce intelligible speech; eye movement was unaffected. He normally communicated using an assistive computer-based typing interface controlled by his residual head movements; his typing speed was approximately 5 correct words or 18 correct characters per minute (Section S1).

IMPLANT DEVICE

The neural implant used to acquire brain signals from the participant was a customized combination of a high-density electrocorticography electrode array (manufactured by PMT) and a percutaneous connector (manufactured by Blackrock Microsystems). The rectangular electrode array was 6.7 cm long, 3.5 cm wide, and 0.51 mm thick and consisted of 128 flat, disk-shaped electrodes arranged in a 16-by-8 lattice formation, with a center-to-center distance between adjacent electrodes of 4 mm. During surgical implantation, general anesthesia was used, and the sensorimotor cortex of the left hemisphere, as identified by anatomical landmarks of the central sulcus, was exposed through craniotomy. The electrode array was laid on the pial surface of the brain in the subdural space. The electrode coverage enabled sampling from multiple cortical regions that have been implicated in speech processing, including portions of the left precentral gyrus, postcentral gyrus, posterior middle frontal gyrus, and posterior inferior frontal gyrus.[9,11–13] The dura was closed with sutures, and the cranial bone flap was replaced. The percutaneous connector was placed extra-cranially on the contralateral skull convexity and anchored to the cranium. This percutaneous connector conducts cortical signals from the implanted electrode array through externally accessible contacts to a detachable digital link and cable, enabling transmission of the acquired brain activity to a computer (Fig. S2). The participant underwent surgical

implantation of the device in February 2019 and had no complications. The procedure lasted approximately 3 hours. We began to collect data for this study in April 2019.

REAL-TIME ACQUISITION AND PROCESSING OF NEURAL DATA

A digital-signal processing system (NeuroPort System, Blackrock Microsystems) was used to acquire signals from all 128 electrodes of the implant device and transmit them to a computer running custom software for real-time signal analysis (Section S2 and Figs. S2 and S3).[18,23] As informed by previous research that had correlated neural activity in the 70 to 150 Hz (high-gamma) frequency range with speech processing,[9,12–14,18] we measured activity in the high-gamma band for each channel to use in subsequent analyses and during real-time decoding.

WORD AND SENTENCE TASK DESIGN

The study consisted of 50 sessions over the course of 81 weeks and took place at the participant's residence or a nearby office. The participant engaged in two types of tasks: an isolated-word task and a sentence task (Section S3 and Fig. S4). On average, we collected approximately 27 minutes of neural activity during these tasks at each session. In each trial of each task, a target word or sentence was presented visually to the participant as text on a screen, and then the participant attempted to produce (say aloud) that target.

In the isolated-word task, the participant attempted to produce individual words from a set of 50 English words. This word set contained common English words that can be used to create a variety of sentences, including words that are relevant to caregiving and words requested by the participant. In each trial, the participant was presented with one of these 50 words, and, after a 2-second delay, he attempted to produce that word when the text of the word on the screen turned green. We collected 22 hours of data from 9800 trials of the isolated-word task performed by the participant in the first 48 of the 50 sessions.

In the sentence task, the participant attempted to produce word sequences from a set of 50 English sentences consisting of words from the 50-word set (Sections S4 and S5). In each trial, the participant was presented with a target sentence and attempted to produce the words in that sentence (in order) at the fastest speed he could perform comfortably. Throughout the trial, the word sequence decoded from neural activity was updated in real time and displayed as feedback to the participant. We collected data from 250 trials of the sentence task performed by the participant in 7 of the final 8 sessions. This task is shown in Video 2. A conversational variant of this task, in which the participant was presented with prompts and attempted to respond to them, is shown in Figure 1 and Video 1.

MODELING

We used neural activity data collected during the tasks to train, fine-tune, and evaluate custom models (Sections S6 and S7 and Table S1). Specifically, we created speech-detection and word-classification models that used deep-learning techniques to make predictions from the neural activity. To decode sentences from the participant's neural activity in real time during the sentence task, we also used a natural-language model and a Viterbi decoder (Fig. 1). The speech-detection model processed each time point of neural activity during a

task and detected onsets and offsets of word-production attempts in real time (Section S8 and Fig. S5). We fitted this model using neural activity data and task-timing information collected only during the isolated-word task.

For each attempt that was detected, the word-classification model predicted a set of word probabilities by processing the neural activity spanning from 1 second before to 3 seconds after the detected onset of attempted speech (Section S9 and Fig. S6). The predicted probability associated with each word in the 50-word set quantified how likely it was that the participant was attempting to say that word during the detected attempt. We fitted this model to neural data collected during the isolated-word task.

In English, certain sequences of words are more likely than others. To use this underlying linguistic structure, we created a natural-language model that yielded next-word probabilities given the previous words in a sequence (Section S10).[24,25] We trained this model on a collection of sentences that included only words from the 50-word set; the sentences were obtained with the use of a custom task on a crowd-sourcing platform (Section S4).

The final component in the decoding approach involved the use of a custom Viterbi decoder, which is a type of model that determines the most likely sequence of words given predicted word probabilities from the word classifier and word-sequence probabilities from the natural-language model (Section S11 and Fig. S7).[26] With the incorporation of the language model, the Viterbi decoder was capable of decoding more plausible sentences than what would result from simply stringing together the predicted words from the word classifier.

EVALUATIONS

To evaluate the performance of our decoding approach, we analyzed the sentences that were decoded in real time using two metrics: the word error rate and the number of words decoded per minute (Section S12). The word error rate of a decoded sentence was defined as the number of word errors made by the decoder divided by the number of words in the target sentence.

To further characterize the detection and classification of word-production attempts from the participant's neural activity, we processed the collected isolated-word data with the speech-detection and word-classification models in off line analyses performed after the recording sessions had been completed (Section S13). We measured classification accuracy as the percentage of trials in which the word classifier correctly predicted the target word that the participant attempted to produce. We also measured electrode contributions as the size of the effect that each individual electrode had on the predictions made by the detection and classification models.[19,27]

To investigate the viability of our approach for a long-term application, we evaluated the stability of the acquired cortical signals over time using the isolated-word data (Section S14). By sampling neural data from four different date ranges spanning the 81-week study period, we assessed whether classification accuracy on a subset of data collected in the

final sessions could be improved by including data from earlier subsets as part of the training set for the classification model; such improvement would indicate that training data accumulated across months or years of recording could be used to reduce the need for frequent model recalibration in practical applications of our approach.

STATISTICAL ANALYSES

Results for each experimental condition are presented with 95% confidence intervals when appropriate (Section S15). No adjustments were made for multiple comparisons. The evaluation metrics (word error rate, number of words decoded per minute, and classification accuracy) were specified before the start of data collection. Analyses to assess the long-term stability of speech-detection and word-classification performance with our implant device were designed post hoc.

RESULTS

SENTENCE DECODING

During real-time sentence decoding, the median word error rate across 15 sentence blocks (each block comprised 10 trials) was 60.5% (95% confidence interval [CI], 51.4 to 67.6) without language modeling and 25.6% (95% CI, 17.1 to 37.1) with language modeling (Fig. 2A, top). The lowest word error rate observed for a single sentence block was 7.0% (with language modeling). When chance performance was measured with the use of sentences that had been randomly generated by the natural-language model (Section S12), the median word error rate was 92.1% (95% CI, 85.7 to 97.2). Across all 150 trials, the median number of words decoded per minute was 15.2 with the inclusion of all decoded words and 12.5 with the inclusion of only correctly decoded words (with language modeling) (Fig. 2A, middle). In 92.0% of the trials, the number of detected words was equal to the number of words in the target sentence (Fig. 2A, bottom). Across all 15 sentence blocks, five speech attempts were erroneously detected before the first trial in the block and were excluded from real-time decoding and analysis (all other detected speech attempts were included). For almost all target sentences, the mean number of word errors decreased when the natural-language model was used (Fig. 2B), and in 80 of 150 trials with language modeling, sentences were decoded without error. Use of the natural-language model during decoding improved performance by correcting grammatically and semantically implausible word sequences in the predictions (Fig. 2C). Real-time demonstrations are shown in Videos 1 and 2.

WORD DETECTION AND CLASSIFICATION

In the offline analyses that included data from 9000 attempts to produce isolated words (and excluded the use of the natural-language model), the mean classification accuracy was 47.1% with the use of the speech detector and word classifier to predict the identity of the target word from cortical activity. The accuracy of chance performance (without the use of any models) was 2%. Additional results of the isolated-word analyses are provided in Figures S8 and S9. A total of 98% of these word-production attempts were successfully detected (191 attempts were not detected), and 968 detected attempts were spurious (not associated with a speech attempt) (Section S8). Electrodes in the most ventral aspect of the ventral sensorimotor cortex contributed to word-classification performance to a greater

extent than electrodes in the dorsal aspect of the ventral sensorimotor cortex, whereas the electrodes in the dorsal aspect contributed more to speech-detection performance (Fig. 3A). Classification accuracy was consistent across most of the target words (mean [±SD] classification accuracy across the 50 target words, 47.1±14.5%) (Fig. 3B).

LONG-TERM STABILITY OF ACQUIRED CORTICAL SIGNALS

The long-term stability of the speech-related cortical activity patterns recorded during attempts to produce isolated words showed that the speech-detection and word-classification models performed consistently throughout the 81-week study period without daily or weekly recalibration (Fig. S10). When the models were used to analyze cortical activity recorded at the end of the study period, classification accuracy increased when the data set used to train the classification models contained data recorded throughout the study period, including data recorded more than a year before the collection of the data used to test the models (Fig. 4).

DISCUSSION

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

Previous demonstrations of word and sentence decoding from cortical neural activity have been conducted with participants who could speak without the need for assistive technology to communicate.[15–19] Similar to the problem of decoding intended movements in someone who is paralyzed, the lack of precise time alignment between intended speech and neural activity poses a challenge during model training. We addressed this time-alignment problem with speech-detection approaches[18,28,29] and word classifiers that used machine-learning techniques, such as model ensembling and data augmentation (Section S9), to increase reliability of the model to minor temporal variabilities in recorded signals.[30,31] Decoding performance was largely driven by neural-activity patterns in the ventral sensorimotor cortex, a finding consistent with previous work implicating this area in speech production.[9,12,13] This finding may inform electrode placement in future studies. We were also able to show the preservation of functional cortical representations of speech in a person who had had anarthria for more than 15 years, a finding analogous to previous findings of limb-related cortical sensorimotor representations in tetraplegic persons years after the loss of limb movement.[32,33]

The incorporation of language-modeling techniques in this study reduced the median word error rate by 35 percentage points and enabled perfect decoding in more than half the sentence trials. This improvement was facilitated through the use of all of the probabilistic information provided by the word classifier during decoding and by allowing the decoder

to update previously predicted words each time a new word was decoded. These results show the benefit of integrating linguistic information when decoding speech from neural recordings. Speech-decoding approaches generally become usable at word error rates below 30%,[34] which suggests that our approach may be applicable in other clinical settings.

In previously reported brain–computer interface applications, decoding models often require daily recalibration before deployment with a user,[6,35] which can increase the variability of decoder performance across days and impede long-term adoption of the interface for real-world use.[35,36] Because of the relatively high signal stability of electrocorticographic recordings,[5,37–39] we could accumulate cortical activity acquired by the implanted electrodes across months of recording to train our decoding models. Overall, decoding performance was maintained or improved by the accumulation of large quantities of training data over time without daily recalibration, which suggests that high-density electrocorticography may be suitable for long-term direct-speech neuroprosthetic applications.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Supported by a research contract under Facebook's Sponsored Academic Research Agreement, the National Institutes of Health (grant NIH U01 DC018671-01A1), Joan and Sandy Weill and the Weill Family Foundation, the Bill and Susan Oberndorf Foundation, the William K. Bowes, Jr. Foundation, and the Shurl and Kay Curci Foundation.

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

REFERENCES

1. Beukelman DR, Fager S, Ball L, Dietz A. AAC for adults with acquired neurological conditions: a review. Augment Altern Commun 2007;23:230–42. [PubMed: 17701742]
2. Nip I, Roth CR. Anarthria. In: Kreutzer J, DeLuca J, Caplan B, eds. Encyclopedia of clinical neuropsychology. 2nd ed. New York: Springer International Publishing, 2017:1.
3. Felgoise SH, Zaccheo V, Duff J, Simmons Z. Verbal communication impacts quality of life in patients with amyotrophic lateral sclerosis. Amyotroph Lateral Scler Frontotemporal Degener 2016; 17:179–83. [PubMed: 27094742]
4. Sellers EW, Ryan DB, Hauser CK. Noninvasive brain–computer interface enables communication after brainstem stroke. Sci Transl Med 2014;6(257):257re7.
5. Vansteensel MJ, Pels EGM, Bleichner MG, et al. Fully implanted brain–computer interface in a locked-in patient with ALS. N Engl J Med 2016;375:2060–6. [PubMed: 27959736]
6. Pandarinath C, Nuyujukian P, Blabe CH, et al. High performance communication by people with paralysis using an intracortical brain–computer interface. Elife 2017;6:e18554.
7. Brumberg JS, Pitt KM, Mantie-Kozlowski A, Burnison JD. Brain–computer interfaces for augmentative and alternative communication: a tutorial. Am J Speech Lang Pathol 2018;27:1–12. [PubMed: 29318256]

8. Linse K, Aust E, Joos M, Hermann A. Communication matters — pitfalls and promise of hightech communication devices in palliative care of severely physically disabled patients with amyotrophic lateral sclerosis. Front Neurol 2018;9: 603. [PubMed: 30100896]
9. Bouchard KE, Mesgarani N, Johnson K, Chang EF. Functional organization of human sensorimotor cortex for speech articulation. Nature 2013;495:327–32. [PubMed: 23426266]
10. Lotte F, Brumberg JS, Brunner P, et al. Electrocorticographic representations of segmental features in continuous speech. Front Hum Neurosci 2015;9:97. [PubMed: 25759647]
11. Guenther FH, Hickok G. Neural models of motor speech control. In: Hickok G, Small S, eds. Neurobiology of language. Cambridge, MA: Academic Press, 2015: 725–40.
12. Mugler EM, Tate MC, Livescu K, Templer JW, Goldrick MA, Slutzky MW. Differential representation of articulatory gestures and phonemes in precentral and inferior frontal gyri. J Neurosci 2018;38: 9803–13. [PubMed: 30257858]
13. Chartier J, Anumanchipalli GK, Johnson K, Chang EF. Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex. Neuron 2018;98(5): 1042.e4–1054.e4.
14. Salari E, Freudenburg ZV, Branco MP, Aarnoutse EJ, Vansteensel MJ, Ramsey NF. Classification of articulator movements and movement direction from sensorimotor cortex activity. Sci Rep 2019;9: 14165.
15. Herff C, Heger D, de Pesters A, et al. Brain-to-text: decoding spoken phrases from phone representations in the brain. Front Neurosci 2015;9:217. [PubMed: 26124702]
16. Angrick M, Herff C, Mugler E, et al. Speech synthesis from ECoG using densely connected 3D convolutional neural networks. J Neural Eng 2019;16:036019.
17. Anumanchipalli GK, Chartier J, Chang EF. Speech synthesis from neural decoding of spoken sentences. Nature 2019;568:493–8. [PubMed: 31019317]
18. Moses DA, Leonard MK, Makin JG, Chang EF. Real-time decoding of question-and-answer speech dialogue using human cortical activity. Nat Commun 2019;10:3096. [PubMed: 31363096]
19. Makin JG, Moses DA, Chang EF. Machine translation of cortical activity to text with an encoder-decoder framework. Nat Neurosci 2020;23:575–82. [PubMed: 32231340]
20. Martin S, Iturrate I, Millán JDR, Knight RT, Pasley BN. Decoding inner speech using electrocorticography: progress and challenges toward a speech prosthesis. Front Neurosci 2018;12:422. [PubMed: 29977189]
21. Guenther FH, Brumberg JS, Wright EJ, et al. A wireless brain–machine interface for real-time speech synthesis. PLoS One 2009;4(12):e8218.
22. Brumberg JS, Wright EJ, Andreasen DS, Guenther FH, Kennedy PR. Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech-motor cortex. Front Neurosci 2011;5:65. [PubMed: 21629876]
23. Moses DA, Leonard MK, Chang EF. Real-time classification of auditory sentences using evoked cortical activity in humans. J Neural Eng 2018;15:036005.
24. Kneser R, Ney H. Improved backing-off for M-gram language modeling. In: Conference proceedings: 1995 International Conference on Acoustics, Speech, and Signal Processing. Vol. 1. New York: Institute of Electrical and Electronics Engineers, 1995:181–4.
25. Chen SF, Goodman J. An empirical study of smoothing techniques for language modeling. Comput Speech Lang 1999;13:359–94.
26. Viterbi AJ. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans Inf Theory 1967;13:260–9.
27. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. In: Bengio Y, LeCun Y, eds. Workshop at the International Conference on Learning Representations. Banff, AB, Canada: ICLR Workshop, 2014.
28. Kanas VG, Mporas I, Benz HL, Sgarbas KN, Bezerianos A, Crone NE. Real-time voice activity detection for ECoG-based speech brain machine interfaces. In: 19th International Conference on Digital Signal Processing: proceedings. New York: Institute of Electrical and Electronics Engineers, 2014:862–5.
29. Dash D, Ferrari P, Dutta S, Wang J. NeuroVAD: real-time voice activity detection from non-invasive neuromagnetic signals. Sensors (Basel) 2020;20:2248.

30. Sollich P, Krogh A. Learning with ensembles: how overfitting can be useful. In: Touretzky DS, Mozer MC, Hasselmo ME, eds. Advances in neural information processing systems 8. Cambridge, MA: MIT Press, 1996:190–6.
31. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Bartlett P, Pereira FCN, Burges CJC, Bottou L, Weinberger KQ, eds. Advances in neural information processing systems 25. Red Hook, NY: Curran Associates, 2012:1097–105.
32. Shoham S, Halgren E, Maynard EM, Normann RA. Motor-cortical activity in tetraplegics. Nature 2001;413:793. [PubMed: 11677592]
33. Hochberg LR, Serruya MD, Friehs GM, et al. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. Nature 2006;442:164–71. [PubMed: 16838014]
34. Watanabe S, Delcroix M, Metze F, Hershey JR, eds. New era for robust speech recognition: exploiting deep learning. Berlin: Springer-Verlag, 2017.
35. Wolpaw JR, Bedlack RS, Reda DJ, et al. Independent home use of a brain–computer interface by people with amyotrophic lateral sclerosis. Neurology 2018; 91(3):e258–e267. [PubMed: 29950436]
36. Silversmith DB, Abiri R, Hardy NF, et al. Plug-and-play control of a brain–computer interface through neural map stabilization. Nat Biotechnol 2021;39:326–35. [PubMed: 32895549]
37. Chao ZC, Nagasaka Y, Fujii N. Long-term asynchronous decoding of arm motion using electrocorticographic signals in monkeys. Front Neuroeng 2010;3:3. [PubMed: 20407639]
38. Rao VR, Leonard MK, Kleen JK, Lucas BA, Mirro EA, Chang EF. Chronic ambulatory electrocorticography from human speech cortex. Neuroimage 2017;153:273–82. [PubMed: 28396294]
39. Pels EGM, Aarnoutse EJ, Leinders S, et al. Stability of a chronic implanted brain–computer interface in late-stage amyotrophic lateral sclerosis. Clin Neurophysiol 2019;130:1798–803. [PubMed: 31401488]

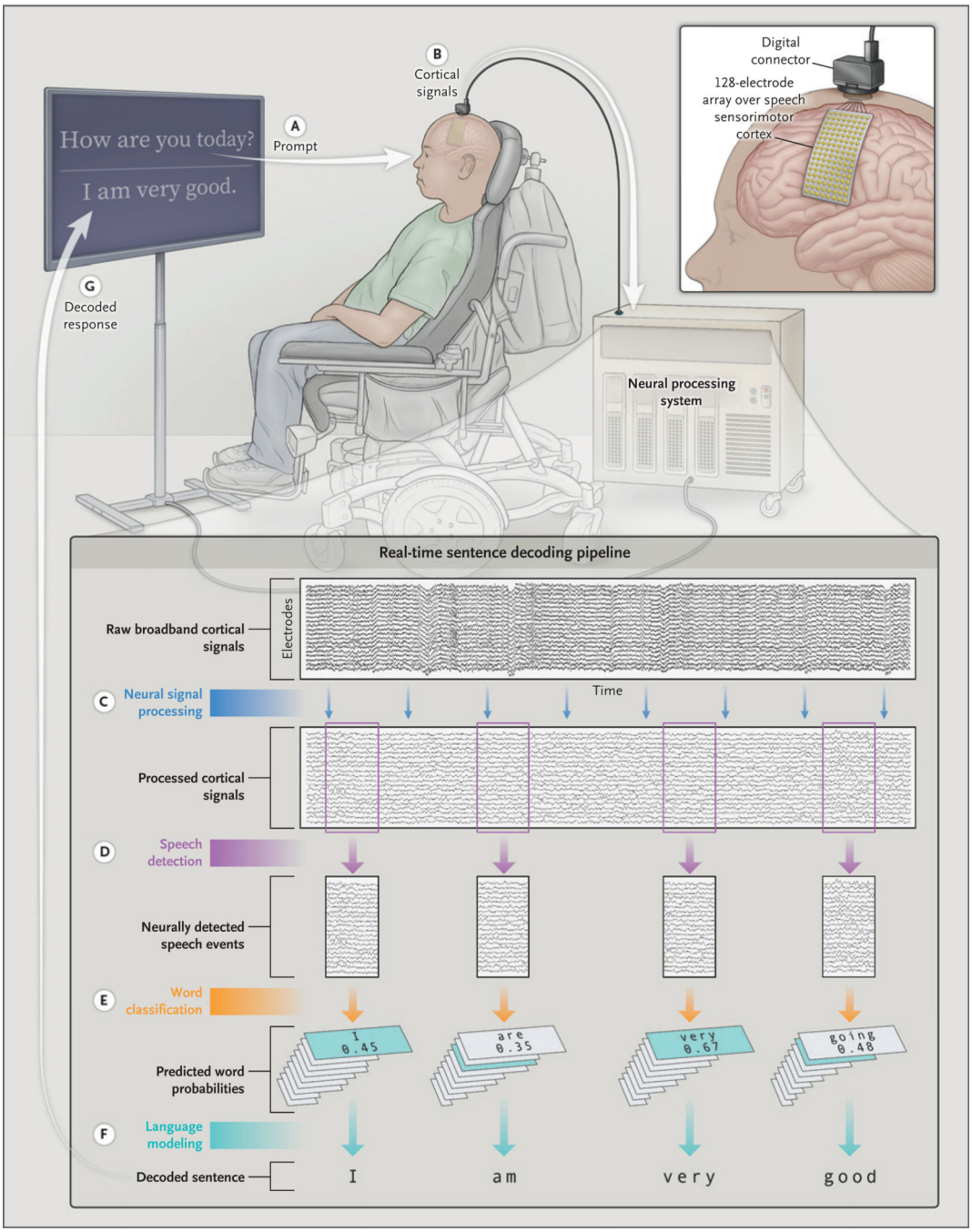
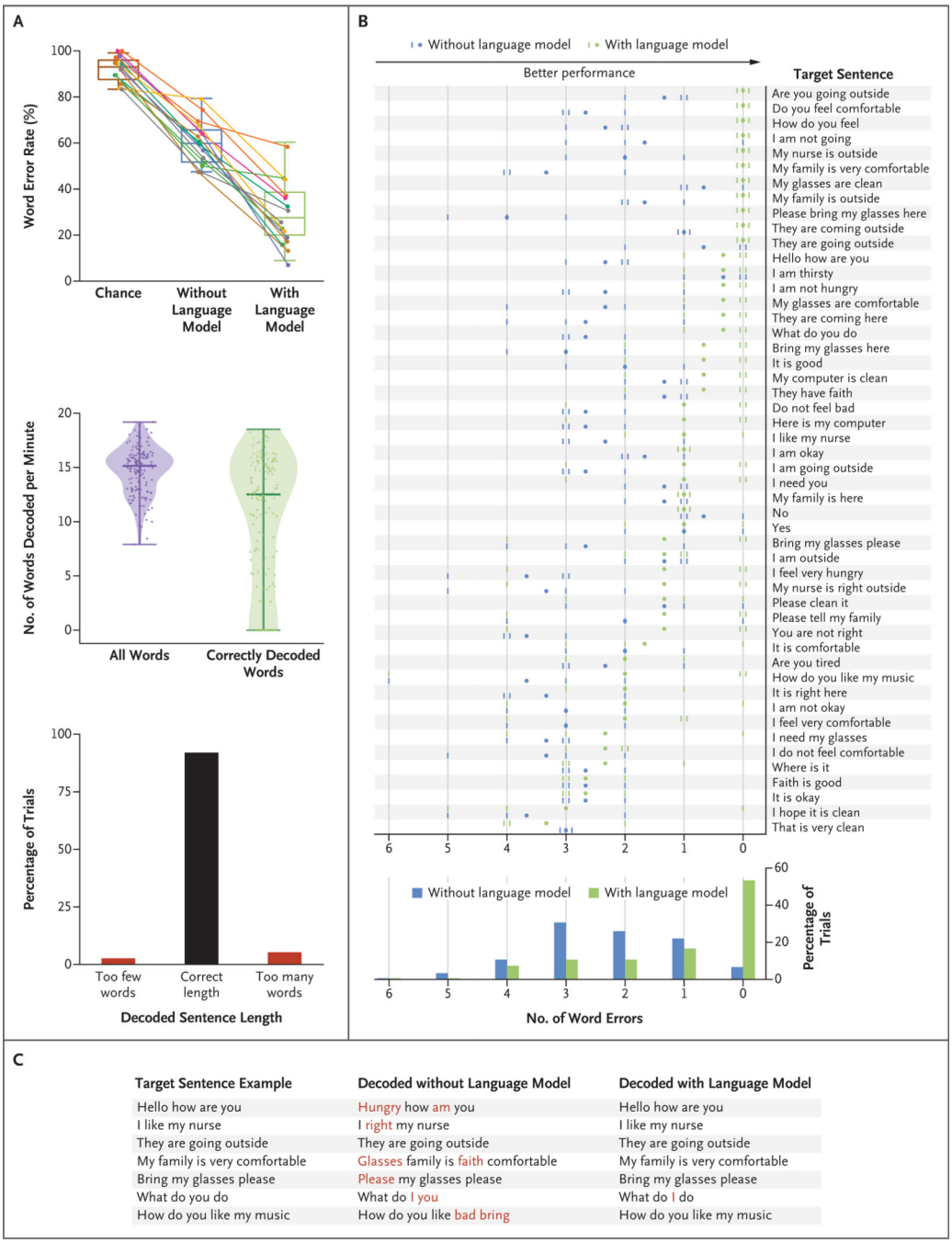


Figure 1. Schematic Overview of the Direct Speech Brain–Computer Interface.

Shown is how neural activity acquired from an investigational electrocorticography electrode array implanted in a clinical study participant with severe paralysis is used to directly decode words and sentences in real time. In a conversational demonstration, the participant is visually prompted with a statement or question (A) and is instructed to attempt to respond using words from a predefined vocabulary set of 50 words. Simultaneously, cortical signals are acquired from the surface of the brain through the electrode array (B) and processed in real time (C). The processed neural signals are analyzed sample by sample with the use

of a speech-detection model to detect the participant's attempts to speak (D). A classifier computes word probabilities (across the 50 possible words) from each detected window of relevant neural activity (E). A Viterbi decoding algorithm uses these probabilities in conjunction with word-sequence probabilities from a separately trained natural-language model to decode the most likely sentence given the neural activity data (F). The predicted sentence, which is updated each time a word is decoded, is displayed as feedback to the participant (G). Before real-time decoding, the models were trained with data collected as the participant attempted to say individual words from the 50-word set as part of a separate task (not depicted). This conversational demonstration is a variant of the standard sentence task used in this work, in that it allows the participant to compose his own unique responses to the prompts.



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

Figure 2. Decoding a Variety of Sentences in Real Time through Neural Signal Processing and Language Modeling.

Panel A shows the word error rates, the numbers of words decoded per minute, and the decoded sentence lengths. The top plot shows the median word error rate (defined as the number of word errors made by the decoder divided by the number of words in the target sentence, with a lower rate indicating better performance) derived from the word sequences decoded from the participant's cortical activity during the performance of the sentence task. Data points represent sentence blocks (each block comprises 10 trials); the median rate, as indicated by the horizontal line within a box, is shown across 15 sentence blocks. The

upper and lower sides of the box represent the interquartile range, and the I bars 1.5 times the interquartile range. Chance performance was measured by computing the word error rate on sentences randomly generated from the natural-language model. The middle plot shows the median number of words decoded per minute, as derived across all 150 trials (each data point represents a trial). The rates are shown for the analysis that included all words that were correctly or incorrectly decoded with the natural-language model and for the analysis that included only correctly decoded words. Each violin distribution was created with the use of kernel density estimation based on Scott's rule for computing the estimator band-width; the thick horizontal lines represent the median number of words decoded per minute, and the thinner horizontal lines the range (with the exclusion of outliers that were more than 4 standard deviations below or above the mean, which was the case for one trial). In the bottom chart, the decoded sentence lengths show whether the number of detected words was equal to the number of words in the target sentence in each of the 150 trials. Panel B shows the number of word errors in the sentences decoded with or without the natural-language model across all trials and all 50 sentence targets. Each small vertical dash represents the number of word errors in a single trial (there are 3 trials per target sentence; marks for identical error counts are staggered horizontally for visualization purposes). Each dot represents the mean number of errors for that target sentence across the 3 trials. The histogram at the bottom shows the error counts across all 150 trials. Panel C shows seven target sentence examples along with the corresponding sentences decoded with and without the natural-language model. Correctly decoded words are shown in black and incorrect words in red.

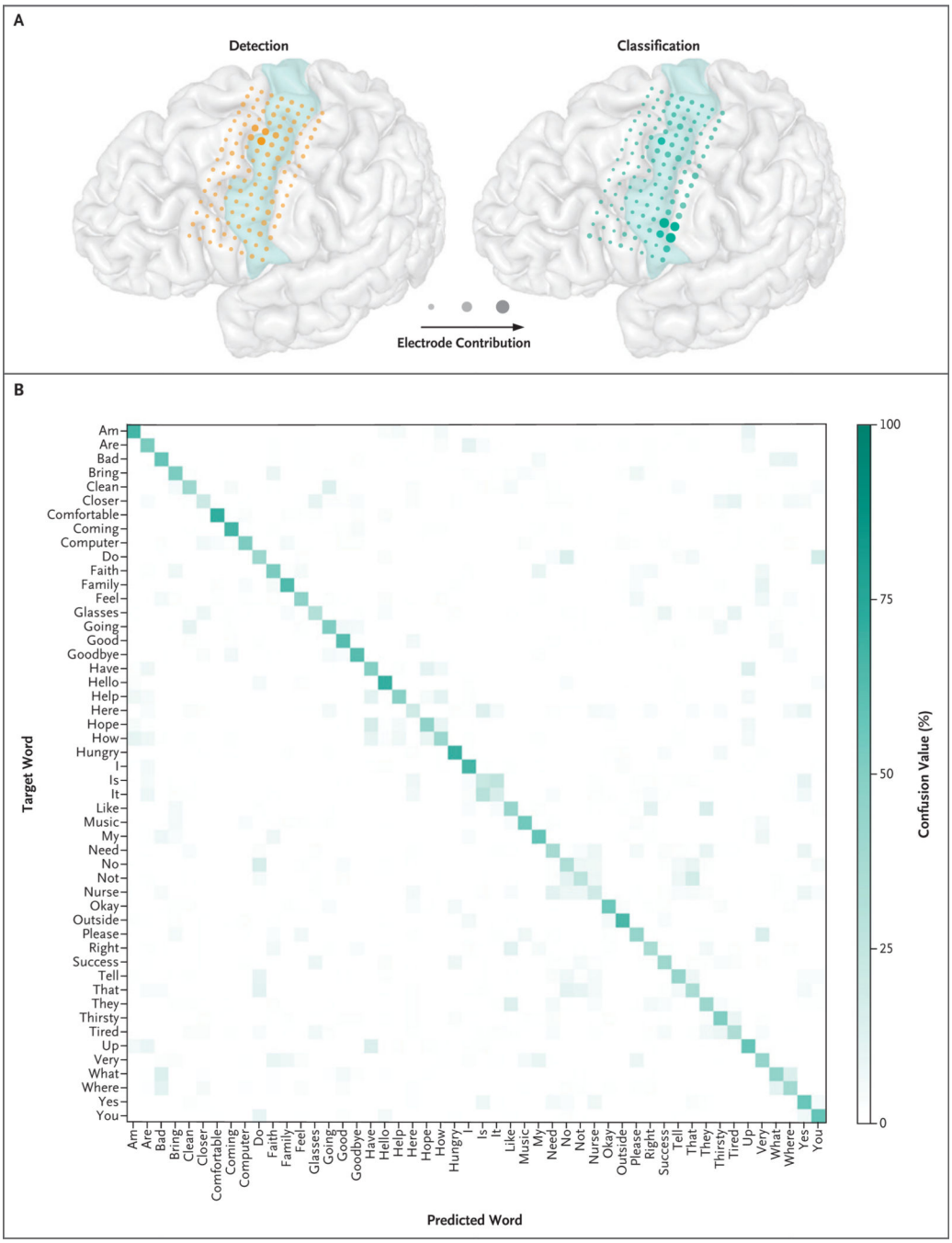


Figure 3. Distinct Neural Activity Patterns during Word-Production Attempts.

Panel A shows the participant's brain reconstruction overlaid with the locations of the implanted electrodes and their contributions to the speech-detection and word-classification models. Plotted electrode size (area) and opacity are scaled by relative contribution (important electrodes appear larger and more opaque than other electrodes). Each set of contributions is normalized to sum to 1. For anatomical reference, the precentral gyrus is highlighted in light green. Panel B shows word confusion values computed with the use of the isolated-word data. For each target word (each row), the confusion value measures

how often the word classifier predicted (regardless of whether the prediction was correct) each of the 50 possible words (each column) while the participant was attempting to say that target word. The confusion value is computed as a percentage relative to the total number of isolated-word trials for each target word, with the values in each row summing to 100%. Values along the diagonal correspond to correct classifications, and off-diagonal values correspond to incorrect classifications. The natural-language model was not used in this analysis.

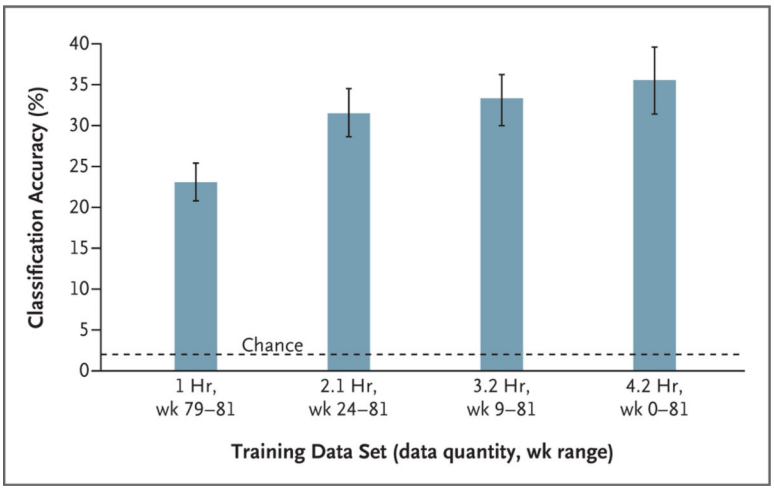


Figure 4. Signal Stability and Long-Term Accumulation of Training Data to Improve Decoder Performance.

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