Article

An instantaneous voice-synthesis neuroprosthesis

https://doi.org/10.1038/s41586-025-09127-3

Received: 19 September 2024

Accepted: 8 May 2025

Published online: 12 June 2025

Check for updates

Maitreyee Wairagkar[1]✉, Nicholas S. Card[1], Tyler Singer-Clark[1,2], Xianda Hou[1,3], Carrina Iacobacci[1], Lee M. Miller[4,5,6], Leigh R. Hochberg[7,8,9], David M. Brandman[1,10]✉ & Sergey D. Stavisky[1,10]✉

Brain–computer interfaces (BCIs) have the potential to restore communication for people who have lost the ability to speak owing to a neurological disease or injury. BCIs have been used to translate the neural correlates of attempted speech into text[1–3]. However, text communication fails to capture the nuances of human speech, such as prosody and immediately hearing one's own voice. Here we demonstrate a brain-to-voice neuroprosthesis that instantaneously synthesizes voice with closed-loop audio feedback by decoding neural activity from 256 microelectrodes implanted into the ventral precentral gyrus of a man with amyotrophic lateral sclerosis and severe dysarthria. We overcame the challenge of lacking ground-truth speech for training the neural decoder and were able to accurately synthesize his voice. Along with phonemic content, we were also able to decode paralinguistic features from intracortical activity, enabling the participant to modulate his BCI-synthesized voice in real time to change intonation and sing short melodies. These results demonstrate the feasibility of enabling people with paralysis to speak intelligibly and expressively through a BCI.

Speaking is an essential human ability and losing the ability to speak is devastating for people living with a neurological disease or injury. Brain–computer interfaces (BCIs) are a promising therapy to restore speech by bypassing the damaged parts of the nervous system by decoding neural activity[4]. Recent demonstrations of BCIs have focused on decoding neural activity into text on a screen[2,3] with high accuracy[1]. Although these approaches offer an intermediate solution to restore communication, communication with text alone falls short of providing a digital surrogate vocal apparatus with closed-loop audio feedback and fails to restore critical nuances of human speech, including prosody.

These additional capabilities can be restored with a brain-to-voice BCI that decodes neural activity into sounds in real time that the user can hear as they attempt to speak. Developing such a speech-synthesis BCI poses several unsolved challenges: the lack of ground-truth training data, that is, not knowing how and when a person with a speech impairment is trying to speak; causal (only using past neural signals) low-latency decoding for instantaneous voice synthesis that provides continuous closed-loop audio feedback; and a flexible decoder framework for producing unrestricted vocalizations and modulating paralinguistic features of the synthesized voice.

An increasing number of studies have reconstructed voice offline from able speakers (without a speech disability) using previously recorded neural signals measured with electrocorticography (ECoG)[5–12], stereoencephalography[13] and intracortical microelectrode arrays[14,15]. Decoders trained on overt speech of able speakers could synthesize unintelligible speech during miming, whispering or imagining speaking tasks online[8,16] and offline[17]. Recently, intermittently intelligible speech was synthesized seconds after a user with amyotrophic lateral sclerosis (ALS) spoke overtly (and intelligibly)[18] from a six-word vocabulary. Although the aforementioned studies were done with able speakers, a study[3] with a participant with anarthria adapted a text-decoding approach to decode discrete speech units acausally (that is, using neural data from both before and after a given utterance) at the end of the sentence to synthesize speech from a 1,024-word vocabulary. However, this is still different from healthy speech, during which people immediately hear what they are saying and can use this to accomplish communication goals such as interjecting in a conversation. Even lower-latency text-to-speech approaches are acausal and would need to wait for the word to be decoded, which would be substantially slower. In this research, we sought to causally synthesize voice continuously and with low latency from neural activity as the user attempted to speak. We refer to this method as instantaneous voice synthesis to contrast it with previous research that used acausal delayed synthesis.

Here we report an instantaneous brain-to-voice BCI using 256 microelectrodes chronically placed in the precentral gyrus of a man with severe dysarthria due to ALS. We did not have ground-truth voice data from this participant. To overcome this limitation, we generated synthetic target speech waveforms from text cued on-screen and time-aligned these with neural activity to estimate the intended speech. We were then able to train a deep-learning model that synthesized his intended voice in real time by decoding his neural activity causally within 10 ms. The resulting synthesized voice was often

[1]Department of Neurological Surgery, University of California, Davis, Davis, CA, USA. [2]Department of Biomedical Engineering, University of California, Davis, Davis, CA, USA. [3]Department of Computer Science, University of California, Davis, Davis, CA, USA. [4]Center for Mind and Brain, University of California, Davis, Davis, CA, USA. [5]Department of Neurobiology, Physiology and Behavior, University of California, Davis, Davis, CA, USA. [6]Department of Otolaryngology, Head and Neck Surgery, University of California, Davis, Davis, CA, USA. [7]School of Engineering and Carney Institute for Brain Sciences, Brown University, Providence, RI, USA. [8]VA Center for Neurorestoration and Neurotechnology, VA Providence Healthcare, Providence, RI, USA. [9]Center for Neurotechnology and Neurorecovery, Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [10]These authors jointly supervised this work: David M. Brandman, Sergey D. Stavisky. ✉e-mail: mwairagkar@ucdavis.edu; dmbrandman@ucdavis.edu; sstavisky@ucdavis.edu

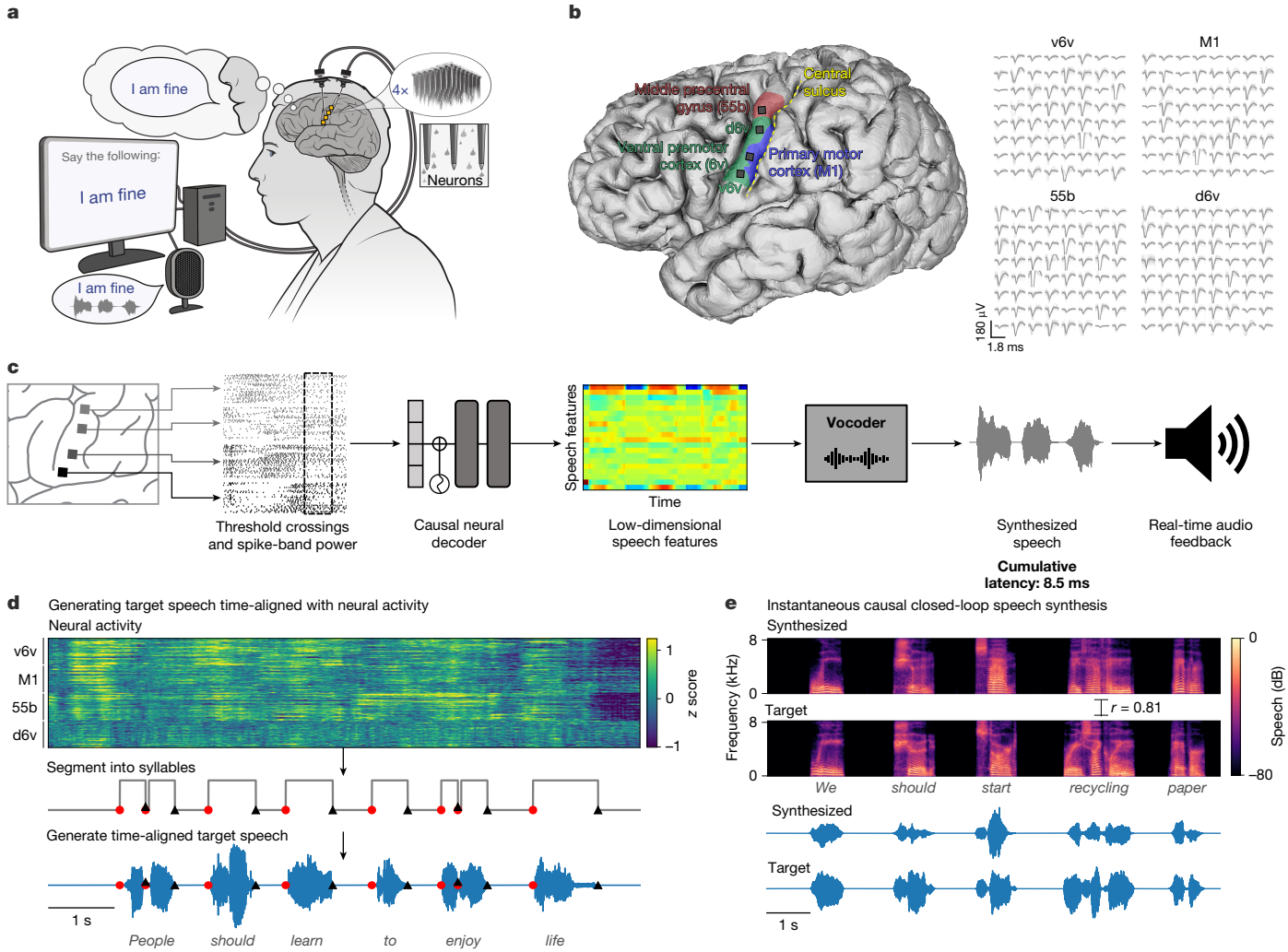


Fig. 1 | Closed-loop voice synthesis from intracortical neural activity in a participant with ALS. a, Brain-to-voice neuroprosthesis. Neural features extracted from four chronically implanted microelectrode arrays were decoded in real time and used to directly synthesize voice. **b**, Array locations on the left hemisphere of the brain and typical neuronal action potentials from each microelectrode. Colour overlays are estimated using a Human Connectome Project cortical parcellation. v6v, ventral 6v; d6v, dorsal 6v. **c**, Closed-loop causal voice-synthesis pipeline: voltages were sampled at 30 kHz; threshold crossings and spike-band power features were extracted from 1-ms segments; these features were binned into 10-ms non-overlapping bins, normalized and smoothed. The Transformer-based decoder mapped these neural features to a low-dimensional representation of speech involving Bark-frequency cepstral coefficients, pitch and voicing, which were used as input to a vocoder. The vocoder then generated speech samples, which were continuously played through a speaker. **d**, Lacking ground-truth speech from T15, we first generated synthetic speech from the known text cue in the training data using a text-to-speech algorithm and then used the neural activity itself to time-align the synthetic speech on a syllable level with the neural data time series to obtain a target speech waveform for training the decoder. Red circles and black triangles show onset and offset of each syllable, respectively. **e**, A representative example of causally synthesized speech from neural data, which matches the target speech with high fidelity.

(but not consistently) intelligible and human listeners were able to identify the words with high accuracy. This flexible brain-to-voice framework, which maps neural activity to acoustic features without an intermediary such as discrete speech tokens or limited vocabulary, could convert the neural activity of the participant to a realistic representation of his voice before ALS, demonstrating voice personalization, and it enabled the participant to speak out-of-dictionary pseudo-words and make interjections.

We also found that, in addition to previously documented phonemic information[1,2], there is substantial paralinguistic information in the intracortical signals recorded from the ventral precentral gyrus. These features were causally decoded to enable the participant to modulate his BCI voice to change intonation so he could ask a question, emphasize specific words in a sentence or sing melodies with different pitch targets. Furthermore, we investigated the dynamics of the neural ensemble activity, which revealed that putatively output-null neural dimensions are highly active well before each word is vocalized, with greater output-null activity present when there were more upcoming words planned and when the upcoming word needed to be modulated.

Continuous low-latency voice synthesis

We recorded neural activity from four microelectrode arrays with a total of 256 electrodes placed in the ventral premotor cortex (6v), primary motor cortex (M1) and middle precentral gyrus (55b) (Fig. 1a,b) as estimated using the Human Connectome Project pipeline[1,19] in BrainGate2 clinical-trial participant T15 (Extended Data Fig. 1a,b). T15 is a 45-year-old man with ALS and severe dysarthria. He retained some orofacial movement and an ability to vocalize but was unable to produce intelligible speech (Supplementary Video 1).

We developed a real-time neural-decoding pipeline (Fig. 1c) to synthesize the voice of T15 instantaneously from intracortical neural activity,

with continuous audio feedback, as he attempted to speak sentences cued on a screen at his own pace. Because the participant could not speak intelligibly, we did not have the ground truth for how and when he attempted to speak. Therefore, to generate aligned neural and voice data for training the decoder, we developed an algorithm to identify putative syllable boundaries directly from neural activity. This allowed us to generate target speech that was time-aligned to neural recordings as a proxy to T15's intended speech (Fig. 1d).

We trained a multilayer Transformer-based[20] model to causally predict spectral and pitch features of the target speech every 10 ms using the preceding binned threshold crossings and spike-band power. The base Transformer model architecture was augmented to compensate for session-to-session neural signal nonstationarities[21] and to lower the inference time for instantaneous voice synthesis. The entire neural processing, from signal acquisition to synthesis of speech samples, occurred within 10 ms, enabling near-instantaneous speech synthesis (Extended Data Fig. 1c shows end-to-end timings). The resulting audio was synthesized into voice samples by a vocoder[22] and continuously played back to T15 through a speaker (Fig. 1e).

Flexible and accurate brain-to-voice BCI

We first tested the ability of the brain-to-voice BCI to causally synthesize voice from neural activity while T15 attempted to speak cued sentences (Fig. 2a and Supplementary Video 2). Each trial consisted of a unique sentence that was never repeated in the training or evaluation trials. The synthesized voice was similar to the target speech (Fig. 2b), with a Pearson correlation coefficient of 0.83 ± 0.04 across 40 mel-frequency bands after removing silences between words (Extended Data Fig. 2a reports mel-cepstral distortion). We quantified intelligibility in two ways. First we asked human listeners to match each of the 956 evaluation sentences with the correct transcript (choosing from 6 possible sentences of the same length). The mean and median accuracies were 94.34% and 100%, respectively (Fig. 2l, left). Second, we conducted open-transcription tests with naive human listeners (Fig. 2l, right). The transcription median phoneme error rate was 34.00% and median word error rate was 43.75%. By contrast, the residual dysarthric speech of T15 was transcribed by the same human listeners with a median phoneme error rate of 83.87% and word error rate of 96.43%. This indicates that the brain-to-voice BCI vastly improved the intelligibility of T15. The instantaneous voice synthesis accurately tracked T15's pace of attempted speech (Extended Data Fig. 3a), which, owing to his ALS, meant slowly speaking one word at a time with pauses in between words. These results indicate that the real-time synthesized speech recapitulates the intended speech to a high degree and can be identified by non-expert listeners. We also demonstrated that this brain-to-voice speech neuroprosthesis could be paired with our previously reported high-accuracy brain-to-text decoder[1], which essentially acted as closed captioning (Supplementary Video 3).

All four arrays showed significant speech-related modulation and contributed to voice synthesis, with the most speech-related modulation shown on the v6v and 55b arrays and much less speech-related modulation found on the d6v array (Extended Data Fig. 3). Different electrodes showed a rich variety in speech tuning (Extended Data Fig. 3b–e). Thanks to this high neural information content, the brain-to-voice decoder could be trained even with limited data, as shown by an online demonstration using a limited 50-word vocabulary on the first day of neuroprosthesis use (Supplementary Video 14). Brain-to-voice synthesis was robust to non-speech vocalizations such as coughing, throat clearing, yawning or people talking in the background (Extended Data Fig. 4) and did not show spurious synthesis during these events. We verified that the neural data were not acoustically contaminated by the residual speech of T15 (Extended Data Fig. 5a) and that intelligible speech could not be synthesized from his vocalizations or vibrations from his residual movements (Extended Data Fig. 5b–d and Supplementary Video 15). The performance of the brain-to-voice decoder over time is shown in Extended Data Fig. 2b. We compared this instantaneous voice-synthesis method with an acausal method[3] that decoded a sequence of discrete speech units at the end of each sentence (Supplementary Audio 1). As expected, acausal synthesis, which benefits from integrating over the entire utterance, generated high-quality voice (mel-cepstral distortion: 2.4 ± 0.03 versus 2.9 ± 0.5 for causal synthesis). This result illustrates that instantaneous voice synthesis is a substantially more challenging problem.

People with neurodegenerative diseases may eventually lose their ability to vocalize all together or may find vocalizing tiring. We therefore tested the brain-to-voice BCI during silent mimed speech during which the participant was instructed to attempt to mouth the sentence without vocalizing. Although the decoder was trained only on attempted vocalized speech, it generalized well to mimed speech: the Pearson correlation coefficient was 0.82 ± 0.03, which was not statistically different from voice synthesis during vocalized attempted speech (Fig. 2c,d and Supplementary Video 4). Figure 2l shows the human perception accuracy of synthesized speech during miming. T15 reported that he found attempting mimed speech less tiring than vocalized speech.

The aforementioned demonstrations involved T15 attempting to speak cued sentences. Next, we tested whether the brain-to-voice BCI could synthesize unprompted self-initiated speech, more similar to how a neuroprosthesis would be used during a real-world conversation. We presented T15 with open-ended questions on the screen (including asking for his feedback about the voice synthesis), which he responded to using the brain-to-voice BCI (Fig. 2e). We also asked him to say whatever he wanted (Supplementary Video 5). The accuracy of his free-response synthesis was slightly lower than that of cued speech (Pearson correlation coefficient 0.79 ± 0.05, Wilcoxon rank-sum test, $P = 10^{-6}$, $n_1 = 57$ sentences, $n_2 = 956$ sentences; Fig. 2f). We speculate that this reflected T15 using a different attempted speech strategy (with less attention to enunciating each phoneme) that he commonly used for his personal use with the brain-to-text BCI[1]. We used a simultaneous text decoder[1] and confirmed all responses with T15 to obtain the ground-truth text of what he said, which was used to generate target speech and evaluate synthesized speech.

This brain-to-voice decoder directly predicts acoustic speech features, which enables the user to produce a variety of expressive sounds, including non-word sounds and interjections, which are not possible with language- and vocabulary-dependent speech BCIs. To demonstrate this flexibility, we instructed T15 to use the brain-to-voice BCI to say made-up pseudo-words and interjections (for example, 'aah', 'eww', 'ooh', 'hmm' and 'shoo') (Fig. 2h,k and Supplementary Videos 7 and 8). The neuroprosthesis also enabled T15 to spell out words one letter at a time (Fig. 2j and Supplementary Video 9). The brain-to-voice decoder was not trained on pseudo-word, spelling or interjection tasks but was able to synthesize these sounds with a Pearson correlation coefficient of 0.79 ± 0.08 (Fig. 2i).

Voice is an important element of one's identity, and synthesizing the own voice of a user could further improve the restorative aspect of a speech neuroprosthesis. We therefore demonstrated that the instantaneous brain-to-voice framework was personalizable and could approximate voice of T15 before ALS (Fig. 2g and Supplementary Video 6). To achieve this, we trained the brain-to-voice decoder on target speech produced by a voice-cloning text-to-speech algorithm[23] that sounded like T15. The participant used the speech-synthesis BCI to report that listening to his own voice "made me feel happy and it felt like my real voice" (Fig. 2e). The accuracy of the own-voice synthesis was similar to the default voice synthesis (Pearson correlation coefficient of 0.77 ± 0.05) (Fig. 2i).

Through these varied speech tasks, we demonstrate that the brain-to-voice BCI framework is flexible and generalizable, enabling the participant to synthesize a wide variety of vocalizations.

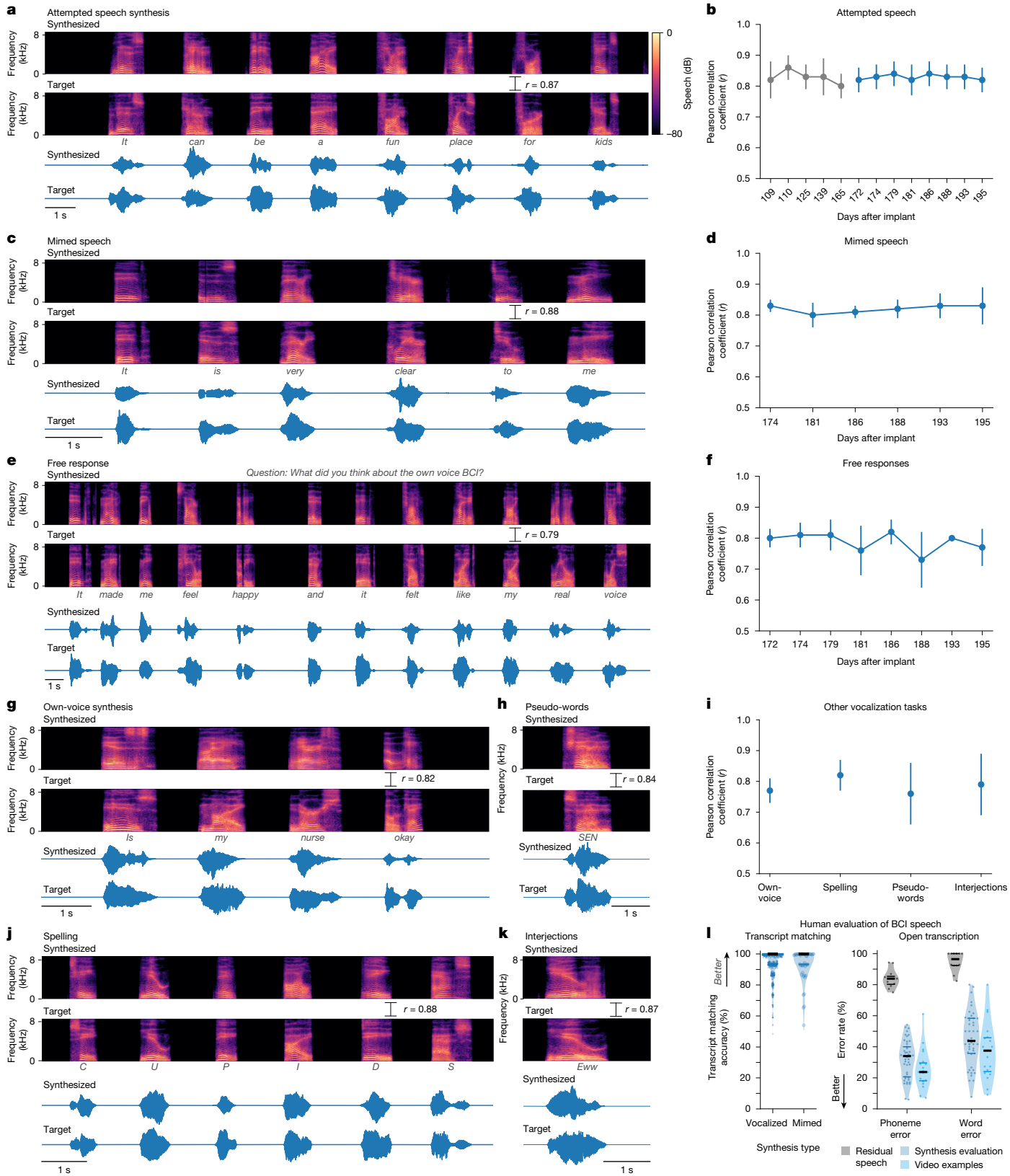


Fig. 2 | See next page for caption.

Closed-loop paralinguistic modulation

Paralinguistic features such as changes in pitch and cadence have an important role in human speech, enabling us to be more expressive. Changing the stress on different words can change the semantic meaning of a sentence; modulating intonations can convey a question, surprise or other emotions; and modulating pitch allows us to sing. Incorporating these paralinguistic features into a BCI-synthesized voice is an important step towards restoring naturalistic speech. We investigated whether these paralinguistic features are encoded

Fig. 2 | The voice neuroprosthesis enables a wide range of vocalizations. **a**, Spectrogram and waveform of an example closed-loop synthesis trial during attempted speech of a cued sentence (top) and the target speech (bottom). The Pearson correlation coefficient (r) is computed across 40 mel-frequencies between the synthesized and target speech excluding silences between words. **b**, Pearson correlations (mean ± s.d.) for vocalized attempted speech. Sessions in blue were predetermined as evaluation sessions (n_{blue} = 956 sentences, n_{grey} = 490 sentences); sessions in grey were earlier experimental sessions. **c**, An example mimed speech trial in which the participant attempted to speak without vocalizing. **d**, Mimed speech Pearson correlation coefficients (mean ± s.d., n = 58 sentences). **e**, An example trial of self-guided attempted speech in response to an open-ended question. **f**, Self-guided speech Pearson correlation coefficients (mean ± s.d., n = 57 sentences). **g**, An example personalized own-voice synthesis trial. **h**,**j**,**k**, Example trials in which the participant said pseudo-words (**h**), spelled out words (**j**) and said interjections (**k**). The decoder was not trained on these words or tasks. **i**, Pearson correlation coefficients of other vocalization tasks: own-voice synthesis, spelling, pseudo-words and interjection synthesis (mean ± s.d., $n_{own\ voice}$ = 100 trials, $n_{spelling}$ = 31 trials, $n_{pseudo\text{-}words}$ = 79 trials, $n_{interjections}$ = 61 trials). **l**, Left, human perception accuracy of synthesized speech. Human listeners selected, for each of the evaluation sentences (956 vocalized and 58 mimed), the correct transcript from 6 possible sentences of the same length. Individual points show the average matching accuracy of each sentence. The bold black line shows the median accuracy (100%) and the thin blue line shows the 25th percentile. Right, open transcription of BCI-synthesized speech by human listeners. Phoneme and word error rates of the BCI-synthesized speech (blue) and residual dysarthric speech of T15 (black) are shown. The bold black line shows the median error rate, and thin coloured lines show the 25th and 75th percentiles.

in the neural activity in the ventral precentral gyrus and developed algorithms to decode and modulate these speech features during closed-loop voice synthesis.

Because the brain-to-voice decoder causally and immediately synthesizes voice, it inherently captures the natural pace of speech of the user. To quantify this, T15 was asked to speak sentences at either a faster or slower speed. The voice synthesized by the neuroprosthesis reflected his intended speaking speed (Fig. 3a). The differing distributions of durations of synthesized words attempted at fast (average speed of 0.97 ± 0.19 s per word) and slow (average speed of 1.46 ± 0.31 s per word) speeds are shown in Fig. 3b.

Next, we decoded the intent to modulate intonation to ask a question or to emphasize a specific word. We recorded neural activity while T15 attempted to speak the same set of sentences as either statements (no extra modulation in pitch) or as questions (with increasing pitch at the end of the sentence). This revealed increased neural activity recorded on all four arrays towards the end of the questions (Extended Data Fig. 6a,b). To study the effect of attempted word emphasis on neural activity, in a different experiment we asked T15 to emphasize one of the seven words in the sentence "I never said she stole my money" by increasing that word's pitch. This sentence, modelled after ref. 24, changes its semantic meaning for each condition while keeping the phonemic content the same. Similar to the effect observed during the question-intonation task, we observed increased neural activity around the emphasized word (Fig. 3c) on all four arrays (Extended Data Fig. 6c), starting approximately 350 ms before the onset of the word.

As a proof of principle that these paralinguistic features could be captured by a speech neuroprosthesis, we trained two separate binary decoders to identify the change in intonation during these question-intonation and word-emphasis tasks. We then applied these intonation decoders in parallel to the brain-to-voice decoder to modulate the pitch and amplitude of the synthesized voice in a closed loop, enabling T15 to ask a question or emphasize a word (Extended Data Fig. 7). Two example closed-loop voice-synthesis trials, including their pitch contours, in which T15 spoke a sentence as a statement and as a question are shown in Fig. 3d. The synthesized speech pitch increased at the end of the sentence during question intonation (Supplementary Video 10). Two example synthesized trials of the same sentence in which different words were emphasized are shown in Fig. 3f (Supplementary Video 11). Across all closed-loop evaluation trials, we decoded and modulated question intonation with 90.5% accuracy (Fig. 3e) and word emphasis with 95.7% accuracy (Fig. 3g). The neural correlates of emphasized words, question words and statement words were distinguishable from each other, with 80.0% offline classification accuracy using a three-class decoder (Extended Data Fig. 6e).

After providing the aforementioned closed-loop binary intonation control for questions or word emphasis, we investigated decoding multiple pitch levels from neural activity. We designed a three-pitch melody task in which T15 attempted to sing different melodies consisting of 6–7 notes of low, medium and high pitch (for example, low-mid-high-high-mid-low). These data were used to train a two-stage Transformer-based pitch decoder. During closed-loop voice synthesis, this pitch decoder ran simultaneously with the brain-to-voice decoder to modulate its pitch output; visual feedback of the decoded pitch level was also provided on-screen (Supplementary Video 12). T15 was able to control the pitch levels of the synthesized melody (Fig. 3h). Three distinct distributions of pitch levels decoded from neural activity across all singing task evaluation trials are shown in Fig. 3i, top, demonstrating that the pitch and phonemic content of speech could be simultaneously decoded from neural activity in real time. To assess human perception of synthesized pitch levels, human listeners were asked to select the higher pitched note for 189 pairs (low × mid, low × high, mid × high). Average human classification accuracy of pitch was 73.02% (Fig. 3i, bottom).

In the preceding experiments, we used a data-efficient separate discrete decoder to modulate the synthesized voice because the vast majority of our training data consisted of neutral sentences without explicit instructions to modulate intonation or pitch. However, a more generalizable approach would be to develop a unified (single) brain-to-voice decoder that takes into account these paralinguistic features. We demonstrated the feasibility of such an approach by training our regular brain-to-voice decoder model architecture with the time-aligned target speech that consisted of different pitch levels for target notes in the three-pitch singing task. This enabled the decoder to implicitly learn the mapping between neural features and the desired pitch level in addition to learning the mapping from neural activity to phonemic content. During continuous closed-loop voice-synthesis evaluation, this unified pitch-enhanced brain-to-voice decoder was able to synthesize different pitch levels as T15 attempted to sing different melodies (Fig. 3j,k and Supplementary Video 13). This demonstrates that the brain-to-voice BCI framework has an inherent capability to synthesize paralinguistic features if provided with training data in which the participant attempts the desired range of vocal properties (in this case, pitch).

Output-null neural dynamics of speech

Instantaneous brain-to-voice synthesis provides a unique view into neural dynamics with high temporal precision. We noticed that neural activity increased before and during the utterance of each word in a cued sentence, but that the aggregate neural activity decreased over the course of the sentence (Fig. 1d and Extended Data Fig. 3a). Despite this broad activity decrease, the synthesis quality remained consistent throughout the sentence (Extended Data Fig. 8). This seeming mismatch between overall neural activity and voice output suggested that the extra activity—which preceded voice onset for each word and gradually diminished towards the end of a sentence—could be a form of output-null neural subspace activity previously implicated in movement preparation[25], feedback processing[26] and other computational support roles[27].

We estimated the output-null and output-potent neural dimensions by adopting established methods[25,26] for this speech BCI application. These consisted of linearly decomposing the population activity into

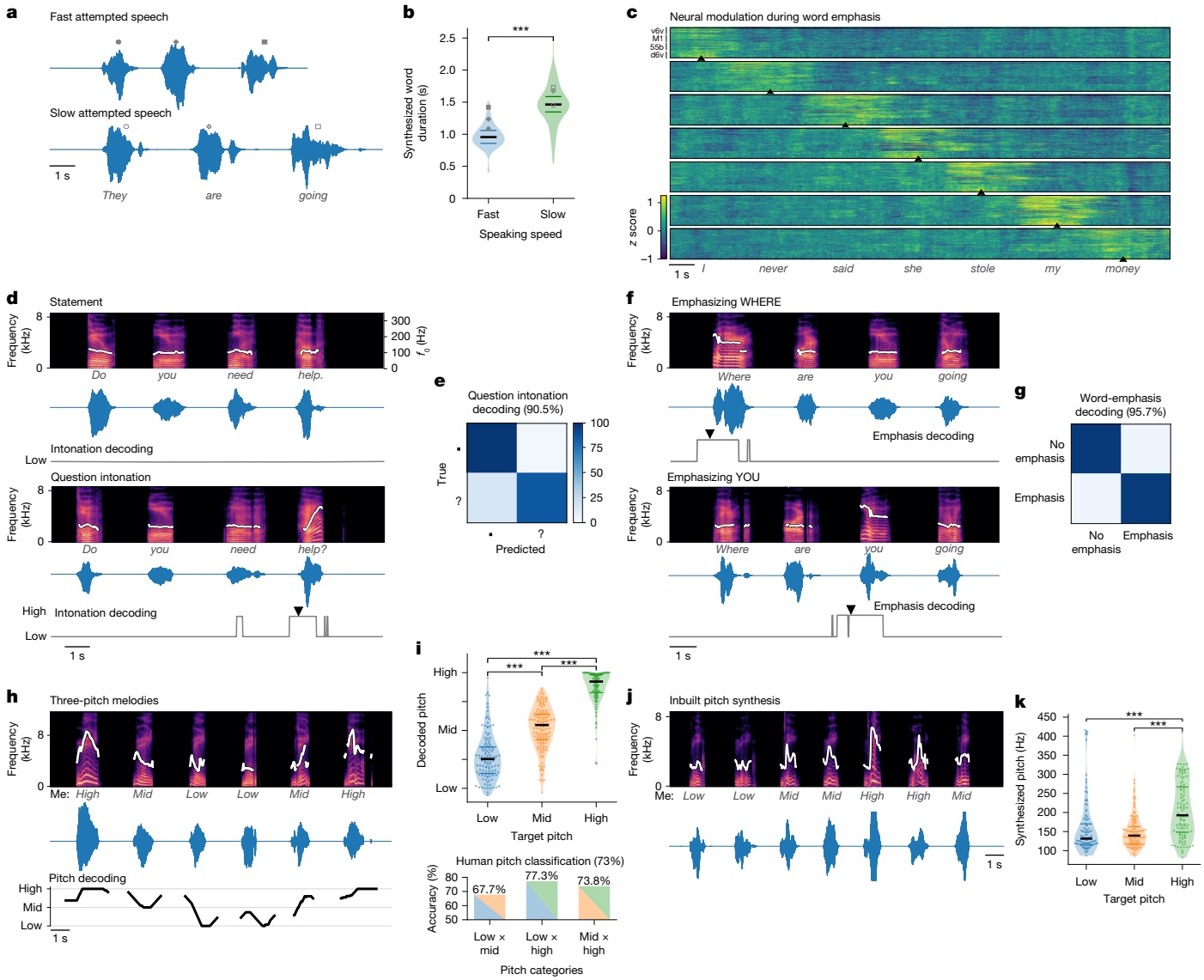


Fig. 3 | Modulating synthesized paralinguistic features. a, Example trials synthesized at faster and slower speeds. Symbols highlighted in **a** match those in **b**. **b**, Violin plots showing significantly different durations of words spoken fast and slowly (two-sided Wilcoxon rank-sum test, $^{***}P = 10^{-14}$, $n_1 = 72$ words spoken fast, $n_2 = 57$ words spoken slowly). **c**, Trial-averaged normalized spike-band power of all electrodes for trials in which the participant emphasized each word in the sentence "I never said she stole my money", grouped by emphasized word. Trials were aligned using dynamic time warping and the mean activity was subtracted. Arrowheads show the onset of each emphasized word. **d**, Spectrograms and waveforms of synthesized voice trials in which the participant said a statement or asked a question. Intonation decoder output is shown below each example. An arrowhead marks the onset of causal pitch modulation. The synthesized pitch contour (white overlay) on spectrograms is constant for a statement and increases during the question word. f_0, pitch (fundamental frequency) of the synthesized voice. **e**, Confusion matrix for closed-loop question intonation modulation during real-time voice synthesis. '?' indicates a question; filled square symbols indicate a statement. **f**, Spectrograms and waveforms of synthesized voice trials in which different words were emphasized. Emphasis decoder output is shown below the waveforms. Arrowheads show the onset of emphasis modulation. **g**, Confusion matrix for closed-loop word emphasis. **h**, Example trial of singing a melody with three pitch targets. The pitch decoder output (bottom) modulated pitch in a closed loop. The synthesized pitch contour accurately reflects the target melody. **i**, Top, violin plots showing decoded pitch levels for low-, medium- and high-pitch target notes (two-sided Wilcoxon rank-sum test, $^{***}P = 10^{-14}$ with Bonferroni correction, $n_1 = 122$ low-pitch notes, $n_2 = 132$ mid-pitch notes, $n_3 = 122$ high-pitch notes). Bottom, average accuracies of human listeners identifying the higher pitch in pairs of synthesized notes (low × mid, $n_1 = 62$; low × high, $n_2 = 66$; mid × high, $n_3 = 61$). **j**, Example three-pitch melody singing synthesized by a unified brain-to-voice model. **k**, Violin plot showing peak synthesized pitch frequencies achieved by the inbuilt pitch synthesis model. Synthesized high pitch was significantly different from low and medium pitch (two-sided Wilcoxon rank-sum test, $^{***}P = 10^{-3}$, $n_1 = 106$, $n_2 = 113$, $n_3 = 105$ (n_1, n_2, n_3 as defined in **h**)).

a subspace with activity that was time-aligned with speech features (output-potent dimensions, which putatively relate most directly to behavioural output) and its orthogonal complement (output-null dimensions, which putatively have a less direct effect on the behavioural output). Both subspaces contained substantial speech-related information: the Pearson correlation coefficients of decoding speech using only the output-potent dimensions (which captured 2.5% of total variance) or only the output-null dimensions (97.5% of total variance) were 0.82 ± 0.06 and 0.85 ± 0.07, respectively. The output-null and output-potent components of neural activity around the onset of each word in sentences of different lengths are shown in Fig. 4a. A clear decrease in output-null activity can be seen over the course of a sentence regardless of its length, whereas the output-potent activity remains consistent (Fig. 4c). An exception to this was the last word,

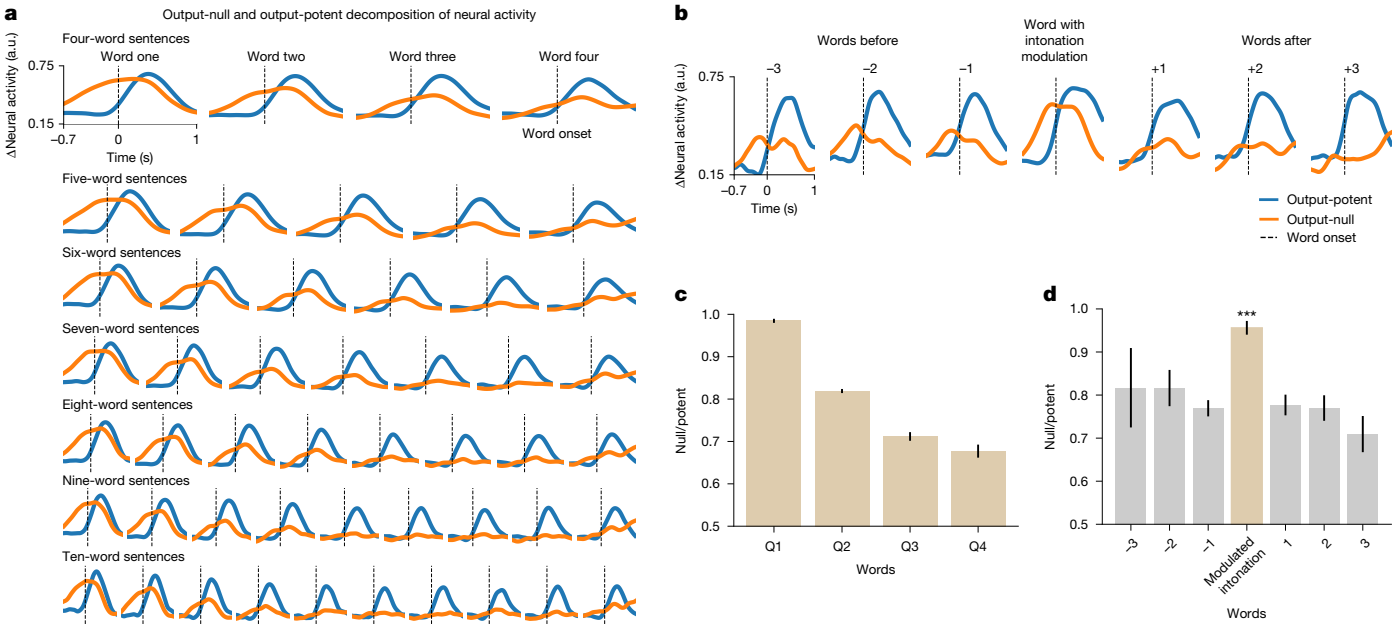


Fig. 4 | Output-null and output-potent neural dynamics during speech production. a, Average approximated output-null (orange) and output-potent (blue) components of neural activity during attempted speech of cued sentences of different lengths. Output-null activity gradually decayed over the course of the sentence, whereas the output-potent activity remained consistent irrespective of the length of the sentence. Each plot in **a** has vertical and horizontal axes as at top left. a.u., arbitrary units. **b**, Average output-null and output-potent activity during intonation modulation (question asking or word emphasis); data are trial-averaged and aligned to the modulated word (centre) and the words before and/or after that word in the sentence. The output-null activity increased during pitch modulation compared with the words before or after it. **c**, Data from **a** are summarized by taking the average null to potent activity ratios for words in the first (Q1), second (Q2), third (Q3) and fourth (Q4) quarter of each sentence (mean ± s.e.m., $n_{Q1} = 3{,}600$ words, $n_{Q2} = 4{,}181$ words, $n_{Q3} = 3{,}456$ words, $n_{Q4} = 3{,}134$ words). **d**, Data from **b** are summarized by calculating average null to potent activity ratios of the intonation-modulated word (beige) and the words before or after it (grey) (mean ± s.e.m.). The null to potent ratios of modulated words were significantly different from those of non-modulated words (two-sided Wilcoxon rank-sum test, $^{***}P = 10^{-21}$, $n_1 = 460$ modulated words, $n_2 = 922$ non-modulated words). Extended Data Fig. 9 shows these analyses for each array individually.

which tended to have an increase in output-null activity, especially as the last word was being finished. We do not know why the end of the sentence had this effect but speculate that it is related to an end-of-trial cognitive change (that is, the participant is assessing his performance).

We also examined the putative output-null and output-potent activity when the participant volitionally modulated his intonation. We found that the output-null activity increased significantly ($P = 10^{-21}$) for the word that was modulated (Fig. 4b,d) compared with the words that preceded or followed it, explaining the previously noted increase in overall neural activity preceding intonation-emphasized words (Fig. 3c and Extended Data Fig. 6a).

Discussion

This study demonstrated a brain-to-voice neuroprosthesis that directly mapped the neural activity recorded from four microelectrode arrays spanning the ventral precentral gyrus in the language-dominant hemisphere into acoustic features. A man with severe dysarthria due to ALS used the system to synthesize his voice in real time as he attempted to speak during both highly structured and open-ended conversations. The resulting voice was often intelligible. The decoding models were trained for a participant who could no longer speak intelligibly (and therefore could not provide a ground-truth speech target) and could be adjusted to emulate his voice before ALS. Unlike previous studies[3,18], this brain-to-voice neuroprosthesis output produces sound as soon as the participant tried to speak, without being restricted to a small number of words[18] and without a constrained intermediate representation of discrete speech units that were generated after completion of each sentence[3]. To demonstrate the flexibility conferred by this direct voice-synthesis BCI, the participant used it to synthesize various vocalizations, including unseen words, interjections and made-up words.

Furthermore, this study demonstrates that a brain-to-voice neuroprosthesis can restore additional communication capabilities compared with existing brain-to-text BCIs[1–3,28]. Neuronal activity in the precentral gyrus encoded both phonemic and paralinguistic features simultaneously. Beyond providing a more immediate way to say words, this system could decode the neural correlates of pitch and speech rate. In online demonstrations, the neuroprosthesis enabled the participant to control a variety of aspects of his instantaneous digital vocalization, such as adjusting the duration of words, emphasizing specific words in a sentence, ending a sentence as either a statement or a question and singing three-pitch melodies. This represents a step towards restoring the ability of people living with speech paralysis to regain the full range of expression provided by the human voice.

We note that participant T15, who had been severely dysarthric for several years at the time of this study, reported that he found it difficult to try to precisely modulate the tone, pitch and amplitude of his attempted speech. Therefore, we propose that using discrete classifiers to generate real-time modulated voice (which provides feedback to the participant that helps them to mentally hone in on how to modulate their voice) can provide an intermediate set of training data useful for training a single unified decoder capable of continuous control of phonemic and paralinguistic vocal features. We demonstrated a proof of concept of this unified approach by training a single core decoder to intrinsically synthesize voice with different pitch levels, which the participant used for singing melodies. T15's attempted speech was slow because of his disease. Ongoing work is investigating whether alternative strategies such as imagined speech can facilitate speech decoding[29].

A functional neuroanatomy result observed in this study that could not be predicted from previous ECoG[24,30–32] and microstimulation studies[33,34] is that the neural activity is correlated with paralinguistic features

across all four microelectrode arrays, from the ventral-most precentral gyrus to the middle precentral gyrus. We also observed that cortical activity across all four arrays increased well before attempted speech. We hypothesize that this reflects output-null preparatory activity[25,27] and note that its presence is particularly fortuitous for the goal of causally decoding voice features, because it gives the decoder a 'sneak peek' shortly before intended vocalization. Notably, this output-null activity seems to decrease over the course of a sentence. This may indicate that the speech motor cortex has a buffer for the whole sentence, which is gradually emptied as the sentence approaches completion. We also observed an increase in output-null activity preceding words that were emphasized or modulated, which we speculate may be a signature of the additional neural computations involved in changing how that word is said. These results hint at considerable richness in speech-related motor cortical ensemble activity, beyond just the activity that is directly linked to driving the articulators. These phenomena represent an opportunity for future study, including leveraging the computation through dynamics framework and neural network modelling that have helped to explain the complexity of motor cortical activity when preparing and producing arm and hand movements[27].

Limitations

This study included a single participant with ALS who retained limited articulatory movement and the ability to vocalize (unintelligibly), with accompanying sensory feedback. It remains to be seen whether similar brain-to-voice performance will be replicated in additional participants, including those with other aetiologies of speech loss or people with late-stage ALS with complete paralysis and who are in a locked-in state. The participant's ALS should also be considered when interpreting the scientific results of the study. Encouragingly, however, previous studies have found that neural coding observations related to hand movements have generalized across people with ALS and models using healthy animal models[35] and across a variety of aetiologies of BCI clinical-trial participants[36,37]. Furthermore, the phonemic and paralinguistic tuning reported here at action potential resolution has parallels to meso-scale ECoG measurements of the sensorimotor cortex in healthy speakers being treated for epilepsy[24,30].

Although the performance demonstrated here compares favourably with previous studies, the synthesized words were still not consistently intelligible. We also anecdotally observed that the energy levels and engagement of the participant during a given block, as well as whether he attempted to enunciate the words clearly and fully, influenced synthesis quality. Brain-to-voice evaluations performed during the research sessions provided limited opportunity for practice-based improvement (that is, sensorimotor learning). It remains an open question whether consistent long-term use will result in improved accuracy owing to additional training data and/or learning and whether performance will be sustained at faster speaking rates. Separately, we predict that accuracy improvement is possible with further algorithm refinement and by increasing the number of electrodes, which was previously shown to improve brain-to-text decoding accuracy[1,2].

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-025-09127-3.

1. Card, N. S. et al. An accurate and rapidly calibrating speech neuroprosthesis. *N. Engl. J. Med.* **391**, 609–618 (2024).
2. Willett, F. R. et al. A high-performance speech neuroprosthesis. *Nature* **620**, 1031–1036 (2023).
3. Metzger, S. L. et al. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature* **620**, 1037–1046 (2023).
4. Silva, A. B., Littlejohn, K. T., Liu, J. R., Moses, D. A. & Chang, E. F. The speech neuroprosthesis. *Nat. Rev. Neurosci.* **25**, 473–492 (2024).
5. Herff, C. et al. Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices. *Front. Neurosci.* **13**, 1267 (2019).
6. Angrick, M. et al. Speech synthesis from ECoG using densely connected 3D convolutional neural networks. *J. Neural Eng.* **16**, 036019 (2019).
7. Anumanchipalli, G. K., Chartier, J. & Chang, E. F. Speech synthesis from neural decoding of spoken sentences. *Nature* **568**, 493–498 (2019).
8. Meng, K. et al. Continuous synthesis of artificial speech sounds from human cortical surface recordings during silent speech production. *J. Neural Eng.* **20**, 046019 (2023).
9. Le Godais, G. et al. Overt speech decoding from cortical activity: a comparison of different linear methods. *Front. Hum. Neurosci.* **17**, 1124065 (2023).
10. Liu, Y. et al. Decoding and synthesizing tonal language speech from brain activity. *Sci. Adv.* **9**, eadh0478 (2023).
11. Berezutskaya, J. et al. Direct speech reconstruction from sensorimotor brain activity with optimized deep learning models. *J. Neural Eng.* **20**, 056010 (2023).
12. Shigemi, K. et al. Synthesizing speech from ECoG with a combination of transformer-based encoder and neural vocoder. In *ICASSP 2023 – 2023 IEEE Int. Conf. Acoust. Speech Signal Process.* 1–5 (IEEE, 2023).
13. Chen, X. et al. A neural speech decoding framework leveraging deep learning and speech synthesis. *Nat. Mach. Intell.* **6**, 467–480 (2024).
14. Wilson, G. H. et al. Decoding spoken English from intracortical electrode arrays in dorsal precentral gyrus. *J. Neural Eng.* **17**, 066007 (2020).
15. Wairagkar, M., Hochberg, L. R., Brandman, D. M. & Stavisky, S. D. Synthesizing speech by decoding intracortical neural activity from dorsal motor cortex. In *2023 11th Int. IEEE/EMBS Conf. on Neural Eng. (NER)* 1–4 (IEEE, 2023).
16. Angrick, M. et al. Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity. *Commun. Biol.* **4**, 1055 (2021).
17. Wu, X., Wellington, S., Fu, Z. & Zhang, D. Speech decoding from stereo-electroencephalography (sEEG) signals using advanced deep learning methods. *J. Neural Eng.* **21**, 036055 (2024).
18. Angrick, M. et al. Online speech synthesis using a chronically implanted brain–computer interface in an individual with ALS. *Sci. Rep.* **14**, 9617 (2024).
19. Glasser, M. F. et al. A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).
20. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems* **30** (NIPS, 2017).
21. Downey, J. E., Schwed, N., Chase, S. M., Schwartz, A. B. & Collinger, J. L. Intracortical recording stability in human brain–computer interface users. *J. Neural Eng.* **15**, 046016 (2018).
22. Valin, J.-M. & Skoglund, J. LPCNET: improving neural speech synthesis through linear prediction. In *ICASSP 2019 – 2019 IEEE Int. Conf. on Acoust. Speech Signal Process.* 5891–5895 (IEEE, 2019).
23. Li, Y. A., Han, C., Raghavan, V. S., Mischler, G. & Mesgarani, N. StyleTTS 2: towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *Adv. Neural Inf. Process. Syst.* **36**, 19594–19621 (2023).
24. Dichter, B. K., Breshears, J. D., Leonard, M. K. & Chang, E. F. The control of vocal pitch in human laryngeal motor cortex. *Cell* **174**, 21–31 (2018).
25. Kaufman, M. T., Churchland, M. M., Ryu, S. I. & Shenoy, K. V. Cortical activity in the null space: permitting preparation without movement. *Nat. Neurosci.* **17**, 440–448 (2014).
26. Stavisky, S. D., Kao, J. C., Ryu, S. I. & Shenoy, K. V. Motor cortical visuomotor feedback activity is initially isolated from downstream targets in output-null neural state space dimensions. *Neuron* **95**, 195–208 (2017).
27. Churchland, M. M. & Shenoy, K. V. Preparatory activity and the expansive null-space. *Nat. Rev. Neurosci.* **25**, 213–236 (2024).
28. Moses, D. A. et al. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *N. Engl. J. Med.* **385**, 217–227 (2021).
29. Kunz, E. M. et al. Representation of verbal thought in motor cortex and implications for speech neuroprostheses. Preprint at *bioRxiv* https://doi.org/10.1101/2024.10.04.616375 (2024).
30. Bouchard, K. E., Mesgarani, N., Johnson, K. & Chang, E. F. Functional organization of human sensorimotor cortex for speech articulation. *Nature* **495**, 327–332 (2013).
31. Chartier, J., Anumanchipalli, G. K., Johnson, K. & Chang, E. F. Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex. *Neuron* **98**, 1042–1054 (2018).
32. Lu, J. et al. Neural control of lexical tone production in human laryngeal motor cortex. *Nat. Commun.* **14**, 6917 (2023).
33. Breshears, J. D., Molinaro, A. M. & Chang, E. F. A probabilistic map of the human ventral sensorimotor cortex using electrical stimulation. *J. Neurosurg.* **123**, 340–349 (2015).
34. Ammanuel, S. G. et al. Intraoperative cortical stimulation mapping with laryngeal electromyography for the localization of human laryngeal motor cortex. *J. Neurosurg.* **141**, 268–277 (2024).
35. Pandarinath, C. et al. Neural population dynamics in human motor cortex during movements in people with ALS. *eLife* **4**, e07436 (2015).
36. Stavisky, S. D. et al. Neural ensemble dynamics in dorsal motor cortex during speech in people with paralysis. *eLife* **8**, e46015 (2019).
37. Willett, F. R. et al. Hand knob area of premotor cortex represents the whole body in a compositional way. *Cell* **181**, 396–409 (2020).

Methods

Participant

A participant with ALS and severe dysarthria (referred to as T15), who gave informed consent, was enrolled in the BrainGate2 clinical trial (ClinicalTrials.gov identifier: NCT00912041). This pilot clinical trial was approved under an Investigational Device Exemption (IDE) by the US Food and Drug Administration (IDE G090003). Permission was also granted by the institutional review boards at the University of California, Davis (protocol 1843264 and 1950310) and Mass General Brigham (2009P000505). Over the course of multiple visits and several hours of discussion with D.M.B., the potential participant was advised about the risks of the study, including the risks related to surgery and chronic device implant. He was also informed of the goals of the study (primary: safety assessment; secondary: assess the feasibility of establishing approaches for neural decoding) and how his participation could provide data that further these scientific goals. The participant used his gyroscopic head mouse system and the help of his care partners to ask and answer questions before, during and after the formal informed consent meeting. T15 consented to publication of photographs and videos containing his likeness.

T15 is a left-handed 45-year-old man. His ALS symptoms began 5 years before enrolment into this study. At the time of enrolment, he was non-ambulatory, had no functional use of his upper and lower extremities, and was dependent on others for activities of daily living (for example, moving his wheelchair, dressing, eating and maintaining hygiene). T15 had mixed upper- and lower-motor neuron dysarthria and an ALS Functional Rating Scale Revised (ALSFRS-R) score of 23 (range, 0–48; higher scores indicate better function). He retained some neck and eye movement but had limited orofacial movement. T15 could vocalize but was unable to produce intelligible speech (Supplementary Video 1). He could be interpreted by expert listeners in his care team; this was his primary mode of communication. Owing to ALS, T15's pace of speech was significantly slower than that of healthy speakers. Furthermore, he had previously undergone training with speech language pathologists who taught him to speak slower in an effort to make his residual speech more intelligible.

Four chronic 64-electrode, 1.5-mm-length silicon microelectrode arrays coated with sputtered iridium oxide (Utah array, Blackrock Neurotech) were surgically placed in the left precentral gyrus of T15 (putatively in the ventral premotor cortex, dorsal premotor cortex, primary motor cortex and middle precentral gyrus; Fig. 1b). The array placement locations were estimated on the basis of pre-operative scans using the Human Connectome Project pipeline[1,19] (functional magnetic resonance imaging confirmed that T15 was left-hemisphere dominant for language). Voltage measurements from the arrays were transmitted to a percutaneous connection pedestal. An external receiver (Neuroplex-E, Blackrock Neurotech) connected to the pedestal digitized and processed the measurements and sent information to a series of computers used for neural decoding. Data reported here are from days 25–489 after implant.

Real-time neural signal processing

Raw neural signals (voltage time series filtered between 0.3 and 7.5 kHz and sampled at 30 kHz with 250 nV resolution) were recorded from 256 electrodes and sent to the processing computers in 1 ms packets. We developed the real-time signal-processing and neural-decoding pipeline using the custom-made BRAND platform[38]. Each processing step was conducted in a separate node running asynchronously.

We extracted neural features of action-potential threshold crossings and spike-band power from each 1 ms incoming signal packet (30 samples) within 1 ms to minimize downstream delays. First, each packet was bandpass filtered between 250 and 5,000 Hz (fourth-order zero-phase non-causal Butterworth filter) by adding 1 ms padding on both sides (from previous samples on one side and a constant mean value of the current samples on the other side) to minimize discontinuities at edges and denoised using linear regression referencing[39]. Then, threshold crossings were detected when the voltage dropped below −4.5 times the root mean squared value for each channel (electrode). Spike-band power was computed by squaring and taking the mean of the samples in the filtered window for each channel and was clipped at 50,000 μV^2 to avoid outliers.

Neural features were binned into 10 ms non-overlapping bins (counting threshold crossings and taking the mean spike-band power across ten consecutive feature windows, such that each of the 256 electrodes contributed two features). Each bin was first log-transformed, then normalized using rolling means and s.d. from the past 10 s for that feature. Each feature was then causally smoothed using a sigmoid kernel of length 1.5 s of the past activity. Therefore, a vector of 1×512 binned neural features was sent to the brain-to-voice decoder every 10 ms. After each block of neural recording (a contiguous period of task performance), we re-computed root mean squared thresholds and linear regression referencing weights to be used in the next block. This helped to minimize nonstationarities in the neural signals.

Experimental paradigm

This study comprises multiple closed-loop speech tasks used to develop and evaluate a voice-synthesis neuroprosthesis. Research sessions were conducted in the home of the participant. They were structured as a series of blocks of approximately 50 trials of a specific task. Each trial began with a delay period of 1.5–4 s in which a red square and a text cue was shown on the screen. During this period, the participant was instructed to read the cue and prepare to speak. This was followed by a go period (indicated by a green square) during which the participant was instructed to attempt to speak the cued text at his own pace after which he ended the trial using an eye tracker by looking at a 'done' icon on the screen. Closed-loop instantaneous voice synthesis was active during the go period. There was then a short 1–1.5 s interval before the start of the next trial.

The participant performed the following speech tasks using the above trial structure: (1) attempting to speak cued sentences; (2) miming (without vocalizing) cued sentences; (3) responding to open-ended questions in his own words or saying anything he wanted; (4) spelling out words letter by letter; (5) attempting to speak made-up pseudo-words; (6) saying interjections; (7) speaking cued sentences at fast or slow speeds; (8) modulating the intonation of speech to say a sentence as a statement or as a question; (9) emphasizing certain words in a sentence; and (10) singing melodies with different pitch level targets (this task had a prompted reference audio cue for the melody that was played during the delay period). For all of the above tasks, the participant was instructed to attempt to speak as clearly as he could, that is, attempt to make articulatory movements (without vocalization for the miming condition). His attempted speech was slow.

After the initial eight research sessions with a mix of open-loop and closed-loop blocks, all cued sentence trials in the rest of the sessions were conducted with either closed-loop voice synthesis, closed-loop text decoding[1] or both to improve the participant's engagement in the task. All other types of tasks had closed-loop voice-synthesis feedback. In a typical research session, we recorded approximately 150–350 structured trials.

Closed-loop brain-to-voice synthesis

Generating decoder training target voice. Because T15 was unable to produce intelligible speech, we did not have a ground-truth reference of his speech to match with the neural activity required for training a decoder to causally synthesize voice. We therefore generated a target speech waveform aligned with neural activity as an approximation of the intended speech of T15.

We first generated synthetic speech waveforms from the known text cues in the training data using a text-to-speech (TTS) algorithm

(native TTS application on MacOS v.13.5.1). Next, we identified putative syllable boundaries of the attempted speech of T15 from the corresponding neural activity and aligned the synthetic speech by dynamically time stretching it to match these syllable boundaries to obtain the time-aligned target speech. The target speech was aligned at the syllable level because syllables are the fundamental units of prosody in human speech[40]. During our first research session, there was no prior neural data available, so we used coregistered microphone recordings of the attempted (unintelligible) speech of T15 to segment word boundaries and generate time-aligned target speech (that is, unlike for later sessions, the session 1 time alignment was done at a word rather than syllable level). This was done programmatically by identifying word boundaries on the basis of the voice amplitude envelope of the participant's attempted speech waveform, which was then adjusted manually if required. In subsequent sessions, we relied solely on neural data to estimate syllable boundaries: we used a brain-to-voice model trained on past neural data to synthesize speech and used its envelope to automatically segment syllable boundaries[41], which were manually reviewed and occasionally adjusted if required. We then used dynamic time warping to align the envelope of the synthetic TTS speech to the envelope of the neurally predicted speech to create a mapping between the syllable boundaries. Then, the TTS segment in each syllable boundary was time-stretched to have the same duration as the corresponding segment of neurally predicted speech. This produced the target speech, which was now time-aligned with neural activity and suitable for training the brain-to-voice decoder for the current session. This process was repeated iteratively for each session.

Brain-to-voice decoder architecture. The core brain-to-voice model was adapted from the Transformer architecture[20]. The model had two main components: an input-embedding network and a base Transformer. Separate input-embedding networks, consisting of two fully connected dense layers (512 and 128 units, respectively, rectified linear unit (ReLU) activation), were used for each week of neural data recording to compensate for week-to-week nonstationarities. The output from the input-embedding network was passed into the base Transformer model, which consisted of eight Transformer encoder blocks (head size 128, number of heads 4, a dropout of 0.5 after multi-head attention layer, each of the two feedforward layers with 256 and 128 units, respectively, and a normalization layer at the beginning and the end). Positional encoding was added before the first Transformer block. Furthermore, we included residual connections between each Transformer block (separate from the residual connections within each block). The output sequence from Transformer blocks was pooled by averaging, then passed through two dense layers (1,024 and 512 units, respectively, ReLU activation) and finally through a dense layer of size 20. The final output comprised 20-dimensional predicted speech features.

At each step, input to the brain-to-voice decoder comprised a 600-ms window of binned neural features (threshold crossings and spike-band power) of shape 60 × 512 (60 bins of 10 ms with 256 channels × 2 features). The first layer of the model averaged two adjacent bins of the input sequence to reduce the sequence length by half while preserving the temporal information. The output of the decoder was a vector of 20 predicted speech features (which were then sent to a vocoder to generate synthesized-speech samples in closed-loop blocks, described in the section 'Instantaneous voice synthesis' below). The decoder ran every 10 ms to produce a single 10-ms frame of voice samples. All of the model hyperparameters were tuned manually with special consideration given to minimize the inference time for instantaneous closed-loop voice synthesis.

Decoder training. We trained a separate decoder for each session using all cued sentence trials (which were unique) from all previous research sessions. To train the decoder robustly, we used 4–20 augmented copies of each trial. Neural features were augmented using three strategies: adding white noise (mean 0, s.d. 1.2) to all time points of all channels independently, a constant offset (mean 0, s.d. 0.6) to all spike-band channels independently and its scaled version (×0.67) to threshold crossings and the same cumulative (random walk) noise (mean 0, s.d. 0.02) to all channels along the time course of the trial. We extracted a 600-ms sliding window (shifted by 10 ms from step to step) from continuous neural features and its corresponding 10-ms frame of output target speech features (20-dimensional vector) as a single training sample. These 20-dimensional output speech features (18 Bark cepstral coefficients, pitch period and pitch strength) for every 10 ms of the target speech waveform were extracted from the time-aligned TTS target voice using the encoder for the pretrained LPCNet vocoder[22]. Each acoustic feature in the 20-dimensional vector was normalized independently on the basis of the minimum and maximum of that feature (obtained from the training TTS dataset) to bring different acoustic features to the same scale (having all output features have a similar numeric range helps with accurate prediction). The features were then temporally smoothed before sending this vector as output in a batch for training the decoder.

The model was trained for approximately 15–20 epochs with a batch size of 1,024 samples (each epoch had approximately 50,000 batches), a constant learning rate of 5×10^{-4}, Adam optimizer ($\beta1 = 0.9$, $\beta2 = 0.98$, $\varepsilon = 1 \times 10^{-9}$) and Hubert loss ($\delta = 1.35$). This method affords the advantage of both L1 and L2 losses and is less sensitive to outliers. The training took between 20 hours and 40 hours on three NVIDIA GeForce RTX 3090 graphic processing units, depending on the amount of data used for training.

During the first session of neural recording, we collected 180 open-loop trials of attempted speech using a 50-word vocabulary to train the decoder and were able to synthesize voice in a closed loop with audio feedback on the same day. Although the closed-loop synthesis of these data was less intelligible owing to the model not being optimized on the first day, we later demonstrated offline that with an optimized model, we could achieve intelligible synthesis with such a small amount of neural data and a limited vocabulary (Supplementary Video 14).

In subsequent sessions, we collected more attempted speech trials with a large vocabulary and iteratively optimized our brain-to-voice decoder architecture to improve the synthesis quality. Here we report the performance of closed-loop voice synthesis from neural activity using the final brain-to-voice decoder architecture for predetermined evaluation sessions. For each of these sessions, the decoder was trained on all of the data collected up to one week prior (total of approximately 4,100–8,300 trials). Further details of the data used for training the brain-to-voice decoders in each of the sessions shown in Fig. 2 are provided in Supplementary Table 1.

To train the personalized own-voice synthesis model, we first generated time-aligned target speech that sounded like the voice of T15 before ALS using the StyleTTS 2 text-to-speech model[23] fine-tuned on voice samples of T15 before he developed ALS. Supplementary Video 16 shows the details of the voice-cloning pipeline and comparison of the voice of T15 before ALS and the own-voice BCI synthesis. The rest of the process for decoder architecture and training was the same as above.

Instantaneous voice synthesis. During closed-loop real-time voice synthesis, we first extracted neural features every 1 ms. These were then binned, log-transformed, causally normalized and smoothed and aggregated into 600-ms causal sliding windows. This neural feature sequence was decoded by the brain-to-voice model into 20 acoustic speech features at each time step as described in the section 'Human evaluation of synthesized speech using open transcription' above. Inference was done on a single NVIDIA RTX A6000 graphic processing unit. The predicted speech features were rescaled back to their original range before normalization on the basis of the previously computed minimum and maximum of each feature (during training, each acoustic speech feature was normalized independently such that all features

were on the same scale and, hence, the brain-to-voice decoder predicted normalized speech features that therefore needed rescaling for synthesis by the vocoder). The LPCNet vocoder requires 16 additional linear predictive coding features in addition to the 20 acoustic features, but these are not independent and were derived from the 18 predicted cepstral features that were among the 20 decoded acoustic features. This resulted in a 36-feature vector (20 + 16), which was synthesized into a single 10-ms frame of speech waveform (sampled at 16 kHz) using the pretrained LPCNet vocoder every 10 ms. The entire pipeline from neural signal acquisition to reconstruction of speech samples of a single frame took less than 10 ms (Extended Data Fig. 1c). These samples were sent to the audio playback computer as they were generated, from which they were played through a speaker continuously, thereby providing closed-loop audio feedback to the participant. We focused our engineering efforts on reducing the inference latency, which fundamentally bounds speech-synthesis latency. However, pragmatically, we found that the largest latency occurred owing to the audio playback driver (Extended Data Fig. 1c). We were able to subsequently lower these playback latencies substantially and predict that further reductions are possible with additional optimization in interfacing with sound drivers. All results reported in this study are for closed-loop voice synthesis unless specified otherwise.

Evaluation of synthesized speech

Automated evaluation metrics of synthesized speech. We evaluated synthesized speech by measuring the Pearson correlation coefficient between the synthesized and TTS-derived target (fully intelligible) speech after excluding silences between words (because T15 took a pause after each word). Removing silences between words before computing correlation made the measure more sensitive to the quality of synthesis because it would not primarily benefit from just predicting speech versus silence. We computed average Pearson correlations across 40 mel-frequency bands of audio sampled at 16 kHz. First, the mel spectrogram with 40 mel-frequency bands was computed using sliding (Hanning) windows of 50 ms with 10 ms overlap and converted to decibel units. Then silences between words were removed from mel spectrograms of target and predicted speech independently to conjoin all words. After removing silences, mel spectrograms of the target speech and synthesized speech were of different lengths. We therefore time-aligned the two spectrograms using dynamic time warping and then computed the Pearson correlation coefficient between different mel-frequency bands. We also computed mel-cepstral distortion, another common metric used to evaluate speech-synthesis models, between the synthesized speech and the target speech using a previously described method[28], excluding silences between words.

Human evaluation of synthesized speech by transcript matching. To evaluate human perception of BCI-synthesized speech, we first asked 15 naive listeners to listen to each synthesized-speech trial and identify the transcript that matched the audio from six possible sentences of the same length. We used the crowd analytics platform Amazon Mechanical Turk to evaluate 1,014 synthesized sentences (vocalized and mimed trials) each by 15 individuals. To test whether the evaluators actually listened to the audio, we included a fully intelligible control audio clip with each synthesized audio. We rejected any trials with wrong answers for the control audio and resubmitted these trials for evaluation until we had 15 accepted answers per evaluation sentence.

Human evaluation of synthesized speech using open transcription. We also conducted an open transcription task to obtain a more sensitive metric of intelligibility of synthesized speech, especially considering that the transcript-matching accuracies were close to ceiling effects. We recruited 28 naive listeners from the community of the University of California, Davis (native English speakers, 23 female, 5 male, 18–42 years of age) to transcribe each sentence from a representative subset of 90 randomly selected voice-synthesis trials from the evaluation set. By open transcription, we mean that the evaluators were instructed that the sentences they would hear could use any English words (there was no limited vocabulary for them to choose from). For quality control, human evaluators were invited in person to perform the transcription task using headphones and a computer in a quiet space. The evaluators were instructed to type out what they heard, as best they could, after listening to each sentence (which they could repeat multiple times if needed). Participants were remunerated for performing the evaluation task for up to 1 hour. To familiarize these naive listeners with the BCI voice and the pace of speaking for T15 (which is much slower than conventional speech), a short 5-min video was shown that played several audio examples of synthesized voice and displayed their corresponding transcripts as part of their brief orientation before they started the task. The evaluators reported that this video helped them to understand the transcription task better and know what to expect, but that this brief orientation was not sufficient to learn to understand the voice. All 28 evaluators were divided into four groups of seven. Each group transcribed 30–38 different BCI-synthesized trials and five dysarthric speech trials. Therefore, each of the 90 evaluation set trials was transcribed by seven human evaluators. Furthermore, open transcription was also conducted for 38 trials from Supplementary Videos 2–4 to help readers to quantitatively assess these examples and contextualize what they hear with the overall metrics. To compare the intelligibility of the BCI-synthesized speech with the residual dysarthric speech of T15, we also asked the human evaluators to transcribe 20 trials of attempted speech by T15 recorded using a microphone during the research blocks.

The resulting transcriptions were quantified by computing the phoneme error rate (PER) and word error rate (WER), that is, percentage of phonemes or words transcribed incorrectly. To avoid quantization and variability introduced by very short sentences (for example, all four-word sentences could only yield WERs of 0%, 25%, 50%, 75% or 100%), we grouped two sentences together into longer sentence pairs with a larger range of word lengths before computing error rates. For each sentence pair, the median PER and WER from all seven transcribers was computed.

Human evaluation of synthesized pitch. To assess whether pitch modulation in the BCI-synthesized voice was perceptible by human listeners, we assigned human listeners (recruited using Amazon Mechanical Turk) 189 pairs of different pitch notes synthesized by the BCI from randomly selected trials in the closed-loop three-pitch melody-singing task. For each pitch pair (low × mid, low × high and mid × high) drawn from the same evaluation set trial, listeners were asked to select the note with higher pitch. Each pitch pair was evaluated by seven human listeners and the median choice was selected as the final choice for that pair. Mean accuracy was computed for the three types of pitch pair.

Decoding paralinguistic voice features

Decoding intonation intent. We collected task blocks in which T15 was instructed to modulate his attempted speech intonation by saying cued sentences as statements (no change in pitch) or as questions (by changing the pitch from low to high towards the end of the sentence) or by emphasizing capitalized words (by increasing pitch and slightly increasing loudness for emphasis). We analysed question and word-emphasis tasks separately but followed the same decoding procedure. We did not have the ground truth of when exactly T15 modulated his intonation to train the intonation decoders. Therefore, trials were grouped by the cue sentence and their neural data aligned using dynamic time warping[42]. The average of these aligned trials was subtracted from each trial to reveal changes in neural activity (Fig. 3c). These trial-averaged data were used to manually label segments of neural

data from each (warped) single trial as intonation modulation (class 1) or no modulation (class 0) for subsequent use in training intonation decoders.

For intonation decoding, we used only the spike-band power feature owing to its higher signal-to-noise ratio (eliminating the threshold crossings helped to reduce the feature size, which was helpful for this more data-limited decoder training). Sliding windows of 600 ms shifted by 10 ms were derived from binned neural activity to generate samples for decoder training. For each window, we took the mean of two adjacent bins to reduce the sequence length in half (to 30 bins from 60 bins), while preserving the temporal information. This sequence of features was then flattened to obtain a single 7,680-dimensional feature vector (30 × 256) as input for the two-class decoder. A binary logistic regression decoder was trained to classify the neural feature vectors into no change in intonation (0) or change in intonation (1). During closed-loop trials, this intonation decoder ran in parallel to the brain-to-voice decoder and was used to predict intonation using features from the preceding 600 ms of neural activity, every 10 ms. Separate binary decoders were trained to detect intonation modulation for asking questions and for word emphasis. For a given task (questions or word emphasis), the intonation decoders ran simultaneously with the main brain-to-voice decoder.

In an offline analysis, we tested whether a single combined paralinguistics decoder could classify between no change in intonation (class 0), question intonation modulation (class 1) and word emphasis (class 2). We trained a three-class logistic regression classifier using the same training data as above. We generated training samples by using the neural spike-band power 600 ms before to the word onset up to word offset and by sampling it with sliding windows of 600 ms shifted by 10 ms to obtain the same sized feature vector as above. We applied this combined three-class decoder to individual words (600-ms sliding windows applied from a time range consisting of −400 ms to 400 ms from the onset) from the same evaluation data as those for the closed-loop paralinguistic-feature modulation analyses. The most frequently predicted class during this time range was selected as the final class for that word.

Closed-loop intonation modulation. One of the speech features predicted by the brain-to-voice decoder characterizes the pitch component, which is used by the LPCNet vocoder to synthesize a speech waveform. To adjust the intonation of the synthesized voice in real time, we artificially modified the value of this feature (relative to what the brain-to-voice decoder output) after detection of a change in intonation from neural activity by the parallel intonation decoder. For the question intonation task, when the binary intonation decoder detected the question intonation (defined as when more than 60% of bins in the previous 700-ms window were classified as positive), it sent a trigger to modulate the pitch feature predicted by the brain-to-voice decoder according to a predefined pitch profile for asking a question (gradually increasing the pitch of the word from low to high). Each intonation modulation trigger was followed by a refractory period of 1.5 s to avoid consecutive duplicate triggers. The (now-modified) speech features were then synthesized by the vocoder as described in the section 'Instantaneous voice synthesis' above (Extended Data Fig. 7).

Similarly, for emphasizing certain words in a sentence in a closed loop, the binary emphasis decoder sent a trigger (defined as when more than 60% of bins in the previous 800-ms window were classified as positive) to modulate pitch features predicted by the brain-to-voice decoder according to a predefined pitch profile for word emphasis—modulating the pitch from high to low—and increasing the volume of synthesized speech by 20%. We intentionally chose pitch profiles that were exaggerated so that the participant could clearly tell whether or not the intention modulation was working during the closed-loop tasks. However, these can be adjusted in the future to avoid the somewhat artificial-sounding exaggerated intonation.

We computed the accuracy of closed-loop intonation modulation by calculating the fraction of individual words in a sentence that were modulated appropriately (that is, matching the prompt).

Discrete pitch decoding for singing. We collected neural data while T15 attempted to sing three-pitch melodies comprising 6–7 notes of three pitch levels (for example, low-mid-high-high-mid-low). Each note consisted of a short word to be sung in the target pitch level. At the start of each trial, an audio cue of the melody was played during the delay period for reference only. However, we did not instruct the participant to match the frequencies of the three notes used in these cued melodies because it was difficult for T15 to precisely modulate his pitch owing to his severe dysarthria. Rather, he was instructed to try to make three distinct low, medium and high pitch notes of his own choice.

We used a two-stage pitch-decoding approach to decode pitch intent from spike-band power. A first Transformer-based decoder (same architecture as above, but with only two Transformer blocks and no input-embedding network) was used to identify the participant's intention to speak (that is, classify silence and intention to sing, on the basis of training labels in which the intent to sing was taken as the 600 ms before the start of a note to the end of the note). A second Transformer-based decoder then decoded his intended pitch level (1, low; 2, mid; 3, high) if (and only if) an intention to sing was detected. Both decoders were trained using categorical cross entropy loss. Because the results of the previous intonation-modulation tasks showed that changes in paralinguistic features can be detected in advance of speech onset, labels for each pitch level were assigned to the neural data from 600 ms before to the note onset to the end of the note attempted at that pitch.

During the closed-loop singing task, this two-staged pitch decoder ran simultaneously with the core brain-to-voice decoder. The output of the pitch decoder was smoothed with a moving average of a prior 800-ms window and then used to continuously modulate the predicted brain-to-voice pitch feature in real time, which was then vocoded as usual (Fig. 3h,i and Extended Data Fig. 7). Therefore, the participant was able to sing melodies consisting of both phonemic content and three different pitches (for example, '$_{la}$ la ^{la}') through his synthesized voice. Furthermore, we provided closed-loop visual feedback on the screen by showing the decoded pitch level and interactive target cues for the note in the melody that T15 was singing (Supplementary Video 13).

Inbuilt continuous pitch synthesis. In the previous intonation- and pitch-modulation tasks, we used a separate decoder to detect changes in the paralinguistic features and modulate the synthesized voice accordingly. Here we developed a unified brain-to-voice decoder that is inherently capable of synthesizing pitch in the melody-singing task. To achieve this, we trained the regular brain-to-voice decoder using neural data and target speech waveforms with varying pitch levels from the three-pitch singing task. These target waveforms were generated by first using the same time-aligned target speech generation algorithm described above from the TTS of the phonemic content of the note, and then adjusting the pitch to match the instructed pitch levels (low, mid or high) of the notes in the cued melody (Fig. 3j).

Control experiments

Acoustic contamination analysis. We performed an acoustic contamination analysis using a previously described method[43] to test for the possibility of contamination in neural recordings due to (unintelligible) vocalizations by T15 during attempted speech. This analysis statistically determined the likelihood of contamination by evaluating correlations between the spectrum of vocalizations by T15 recorded by a microphone and the spectrum of voltage time series of each intracortical electrode.

We analysed 627 representative trials sampled from all weeks of evaluation data from all three types of vocalized speech task used for

closed-loop voice synthesis: (1) regular attempted vocalized speech; (2) word emphasis; and (3) question intonation modulation. Some trials showed spurious acoustic correlations between microphone static and the background neural activity during silent periods, which were eliminated by removing background static from microphone recordings (by replacing the static only in non-speech silent sections of the microphone audio by zeros). The analysis resulted in 3.03% of trials identified as acoustically correlated (Extended Data Fig. 5a). This indicated that the majority of the speech-synthesis trials were not acoustically contaminated, and the small percentage of correlated trials is consistent with chance and extremely unlikely to meaningfully affect closed-loop voice synthesis.

Decoding non-invasive biosignals. We tested whether the residual speech of T15 or vibrations generated owing to his vocalization carried decodable information, which could possibly introduce artifacts in neural recordings that would inflate decoding accuracy. We performed control experiments by simultaneously recording intracortical signals and speech-related biosignals non-invasively using a microphone to capture the acoustics of T15; a stethoscopic microphone[44] (an acoustic stethoscope with a digital condenser microphone inserted in the stethoscope tube after removing the earpieces) placed on the left mastoid of T15; and an inertial measurement unit (IMU) sensor (MPU-6050, HiLetgo) placed on his right mastoid to capture vibrations as he attempted to speak 240 sentences using a limited 50-word vocabulary (Extended Data Fig. 5b).

We trained separate decoders (using the same decoder architecture as the 'brain-to-voice' decoder described above) for each modality of recordings (and a combined decoder for all biosignals together) using 10-fold cross-validation to synthesize voice offline (Supplementary Video 15). We extracted 40 mel-frequency cepstral coefficients (sufficient to reconstruct voice) as input features using signals from the microphone and stethoscope, and from each of the 6 axes of the IMU sensor for the respective decoders. Thus, the microphone and stethoscope decoders had 40 input features, the IMU decoder had 240 input features, the combined biosignals decoder had 320 input features and the neural decoder had 512 input features. A feature bin size of 10 ms, input sequence length of 600 ms and all other decoder parameters were kept the same as those for neural decoding, as described in the 'Brain-to-voice decoder architecture' section.

We computed the Pearson correlation coefficient of voice synthesized from each of the biosignals with the target speech (Extended Data Fig. 5c) and also measured intelligibility of speech synthesized from neural activity and from stethoscope vibrations by recruiting human listeners using Amazon Mechanical Turk to perform open transcription of (the same) 30 randomly selected trials synthesized by the two decoders: each trial was transcribed by seven human evaluators (using the same human listener evaluation method as described in the section 'Human evaluation of synthesized speech using open transcription' above). We used speech decoded from stethoscope vibrations as representative speech synthesis from biosignals because it showed the highest signal-to-noise ratio of all of the recorded non-invasive biosignals. The median PER for neural voice synthesis was 37.1% and the WER was 43.60% (consistent with human transcription results in Fig. 2l). By contrast, the median PER for biosignal decoding was 74.5% and the WER was 100% (Extended Data Fig. 5d). Therefore, intelligible speech could not be synthesized by decoding acoustic and muscle vibrations of the residual speech of T15.

Output-null and output-potent activity

To study the underlying neural dynamics of speech production, we decomposed the neural activity into two orthogonal output-null and output-potent components by adapting previously published methods[26]. To do this, we first adopted a simplified linear decoding approach. We fit a linear decoder $\mathbf{y} = W\mathbf{x}$, where $\mathbf{x}$ is a vector of neural features, $\mathbf{y}$ is a 20-dimensional vector of speech features, and W is the linear decoder, using ordinary least-squares regression. We trained a separate linear decoder for each session to account for session-to-session nonstationarities. The linear decoder matrix W was then decomposed into orthogonal null and row subspaces. The neural activity $\mathbf{x}$ was projected onto the null space (putative output-null dimensions of the neural activity) and row space (putative output-potent dimensions of the neural activity). Although the linear decoder did not have sufficient power to decode high-quality speech features, it served as a reliable and interpretable method for decomposing neural activity into orthogonal putatively output-null and output-potent subspaces. The change in neural activity for null- and row-space projections for each trial was obtained by computing the Euclidean distance of the projections from the baseline activity (first 500 ms of the trial) and normalizing it between 0 and 1 (minimum and maximum across that whole trial) to get output-null and output-potent components, respectively. This normalization was used to account for potential neural nonstationarities across a session and between datasets, because we were interested in the relative changes in these neural projections over the course of each sentence.

Trial-averaged output-null and output-potent neural-activity components were obtained for all sentences of the same length between −700 ms to +1 s from the onset of each word in a sentence (Fig. 4). This output-null and output-potent analysis was also performed on intonation modulation for questions and word emphasis tasks and the output was compared with that of the regular cued attempted speech task.

We verified that the decaying output-null activity was not merely tracking gross head motion during attempted speech, which could indicate a neural correlate of head movement. To do so, we tracked the head motion of T15 from videos simultaneously recorded with neural signals and compared changes in head movements with output-null and output-potent neural dynamics over time (Extended Data Fig. 10).

Statistical testing

We used a two-sided Wilcoxon rank-sum test to compare two groups of data. The P values were corrected for multiple comparisons using Bonferroni correction where necessary. We used a non-parametric test because datasets being compared were of different sizes and a normal distribution was not assumed because the actual underlying distribution was unknown.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Neural data and brain-to-voice models related to this study are publicly available on Dryad (https://doi.org/10.5061/dryad.2280gb64f)[45].

Code availability

Code to implement brain-to-voice synthesis described in this study is publicly available on GitHub (https://github.com/Neuroprosthetics-Lab/brain-to-voice-2025).

38. Ali, Y. H. et al. BRAND: a platform for closed-loop experiments with deep network models. *J. Neural Eng.* **21**, 026046 (2024).
39. Young, D. et al. Signal processing methods for reducing artifacts in microelectrode brain recordings caused by functional electrical stimulation. *J. Neural Eng.* **15**, 026014 (2018).
40. Levelt, W. J., Roelofs, A. & Meyer, A. S. A theory of lexical access in speech production. *Behav. Brain Sci.* **22**, 1–38 (1999).
41. Räsänen, O., Doyle, G. & Frank, M. C. Unsupervised word discovery from speech using automatic segmentation into syllable-like units. *Proc. Interspeech* **2015**, 3204–3208 (2015).
42. Williams, A. H. et al. Discovering precise temporal patterns in large-scale neural recordings through robust and interpretable time warping. *Neuron* **105**, 246–259 (2020).

43. Roussel, P. et al. Observation and assessment of acoustic contamination of electrophysiological brain signals during speech production and sound perception. *J. Neural Eng.* **17**, 056028 (2020).
44. Shah, N., Sahipjohn, N., Tambrahalli, V., Subramanian, R. & Gandhi, V. StethoSpeech: speech generation through a clinical stethoscope attached to the skin. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **8**, 123 (2024).
45. Wairagkar, M. et al. Data for an instantaneous voice synthesis neuroprosthesis. *Dryad* https://doi.org/10.5061/dryad.2280gb64f (2025).

Acknowledgements We thank participant T15 and his family and care partners for their contributions to this research. Support was provided by the Office of the Assistant Secretary of Defense for Health Affairs through the Amyotrophic Lateral Sclerosis Research Program under award number AL220043; a New Innovator Award (DP2) from the NIH Office of the Director and managed by NIDCD (1DP2DC021055); a Seed Grant from the ALS Association (23-SGP-652); A. P. Giannini Postdoctoral Fellowship (N.S.C.); Searle Scholars Program; a Pilot Award from the Simons Collaboration for the Global Brain (AN-NC-GB-Pilot Extension-00002343-01); NIH-NIDCD (U01DC017844) and VA RR&D (A2295-R). S.D.S. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, and a Cultivating Team Science Award from the University of California Davis School of Medicine.

Author contributions M.W., S.D.S. and D.M.B. conceived the study and experiment design. M.W. led the experiments and developed and implemented the target speech generation, decoder training algorithms and end-to-end pipeline for instantaneous voice synthesis: feature extraction, noise removal, preprocessing, real-time brain-to-voice decoders, pitch decoders, vocoder and output audio playback, experimental tasks, post-processing. M.W. also performed human listener evaluations, analysed all of the data and created figures. M.W. and N.S.C. developed and implemented the real-time neural signal processing, noise removal and feature-extraction pipelines. M.W., N.S.C., T.S.-C. and X.H. coded the real-time data-collection system and built the neuroprosthetic cart system. N.S.C. generated cloned voice samples for T15. M.W., N.S.C. and C.I. collected the primary data for this study. N.S.C. and C.I. interfaced with the participant and scheduled research sessions. L.M.M. contributed to the human listener evaluations. D.M.B. led planning and performed the surgical-implant-placement procedure. L.R.H. was the sponsor–investigator of the multisite clinical trial. D.M.B. was responsible for all clinical-trial-related activities at University of California Davis. S.D.S. and D.M.B. supervised all aspects of the project. M.W and S.D.S. wrote the paper. All authors reviewed and edited the paper.

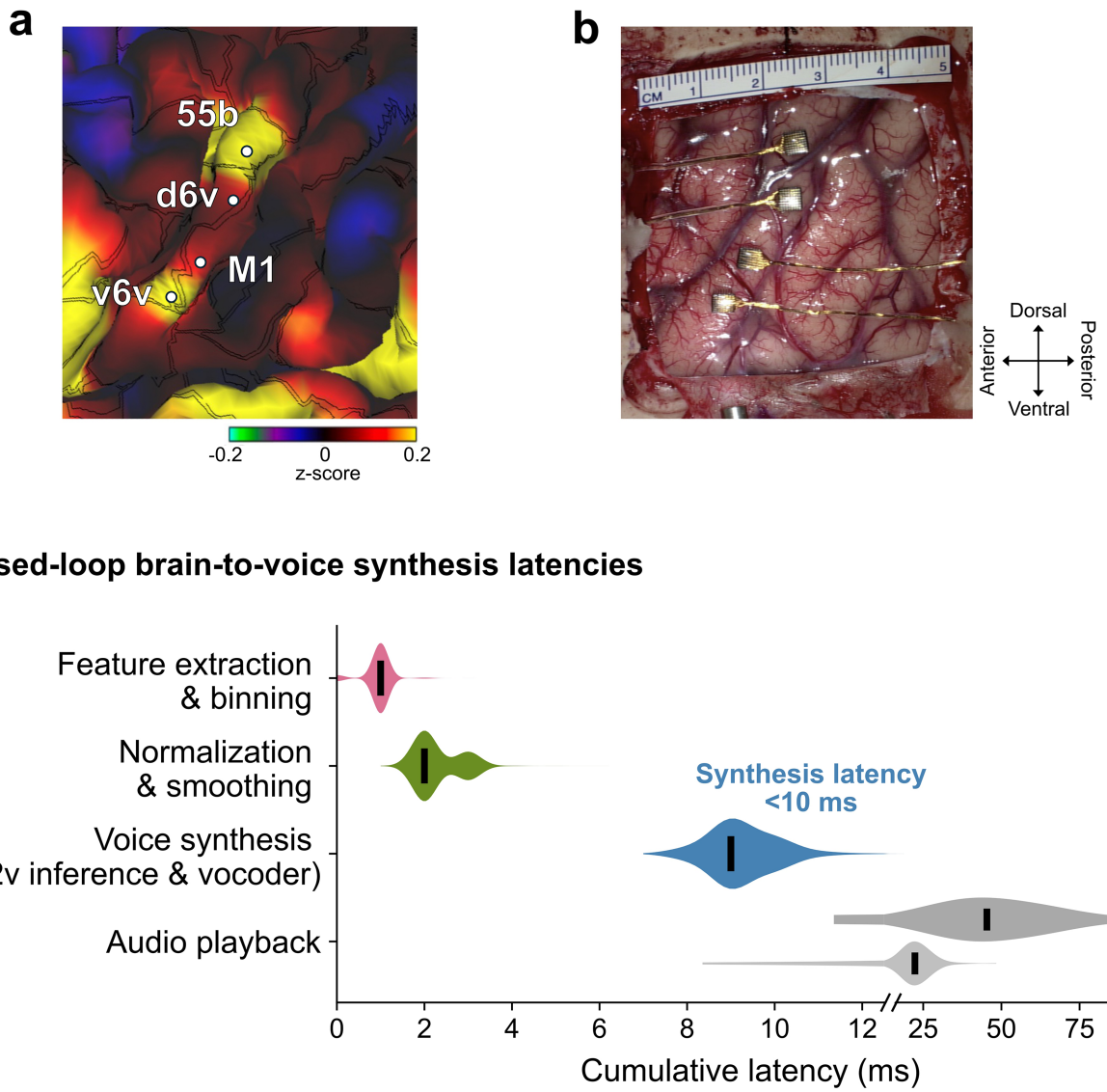
Competing interests S.D.S. is an inventor on intellectual property related to speech decoding submitted and owned by Stanford University (US patent no. 12008987) that has been licensed to Blackrock Neurotech and Neuralink. M.W., S.D.S. and D.M.B. have patent applications related to speech BCI submitted and owned by the Regents of the University of California (US patent application no. 63/461,507 and 63/450,317), including intellectual property licensed by Paradromics. D.M.B. was a surgical consultant with Paradromics, completing his consultation during the revision period of the paper. He is a consultant for Globus Medical. S.D.S. is a scientific adviser to Sonera. The MGH Translational Research Center has a clinical research support agreement with Ability Neuro, Axoft, Neuralink, Neurobionics, Paradromics, Precision Neuro, Synchron and Reach Neuro, for which L.R.H. provides consultative input. Mass General Brigham is convening the Implantable Brain-Computer Interface Collaborative Community (iBCI-CC); charitable gift agreements to Mass General Brigham, including those received to date from Paradromics, Synchron, Precision Neuro, Neuralink and Blackrock Neurotech, support the iBCI-CC, for which L.R.H. provides effort. The other authors declare no competing interests.

Additional information

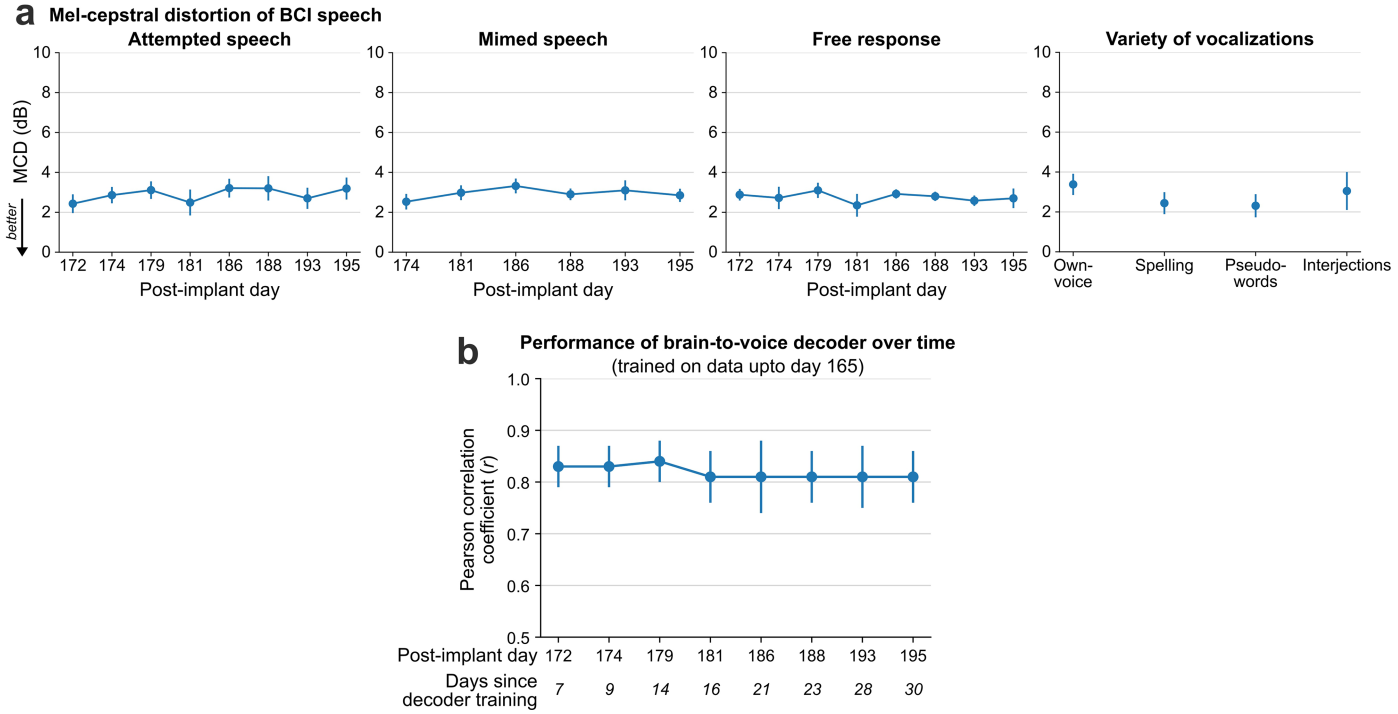
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-025-09127-3.

Correspondence and requests for materials should be addressed to Maitreyee Wairagkar, David M. Brandman or Sergey D. Stavisky.

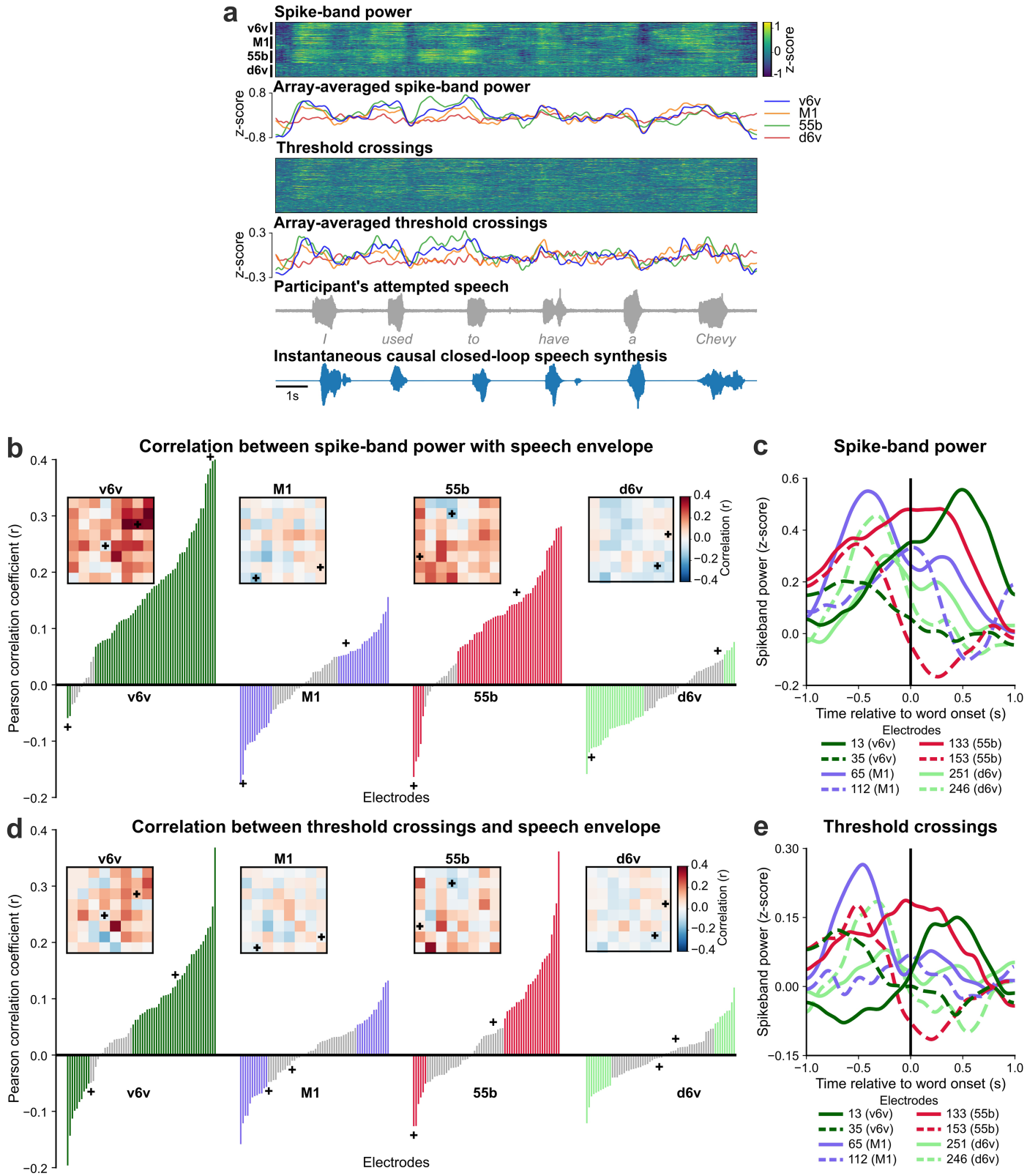
Peer review information *Nature* thanks Nai Ding, Nick Ramsey and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



Extended Data Fig. 1 | Microelectrode array placement and brain-to-voice synthesis latencies. a. The estimated resting state language network from Human Connectome Project data overlaid on T15's brain anatomy. **b**. Intraoperative photograph showing the four microelectrode arrays placed on T15's precentral gyrus. Images in **a** and **b** are adapted from ref. 1 (Copyright © 2024 Massachusetts Medical Society, reprinted with permission from Massachusetts Medical Society). **c**. Closed-loop cumulative latencies across different stages in the voice synthesis and audio playback pipeline are shown. Voice samples were synthesized from raw neural activity measurements within 10 ms and the resulting audio was played out loud continuously to provide closed-loop feedback. Note the linear horizontal axis is split to expand the visual dynamic range. We focused our engineering primarily on reducing the brain-to-voice inference latency, which fundamentally bounds the speech synthesis latency. As a result, the largest remaining contribution to the latency occurred after voice synthesis decoding during the (comparably more mundane) step of audio playback through a sound driver. The cumulative latencies with the audio driver settings used for T15 closed-loop synthesis in earlier experiments are shown in dark grey. Audio playback latencies were subsequently substantially lowered through software optimizations (light grey) in latter sessions and we predict that further reductions will be possible with additional computer engineering.

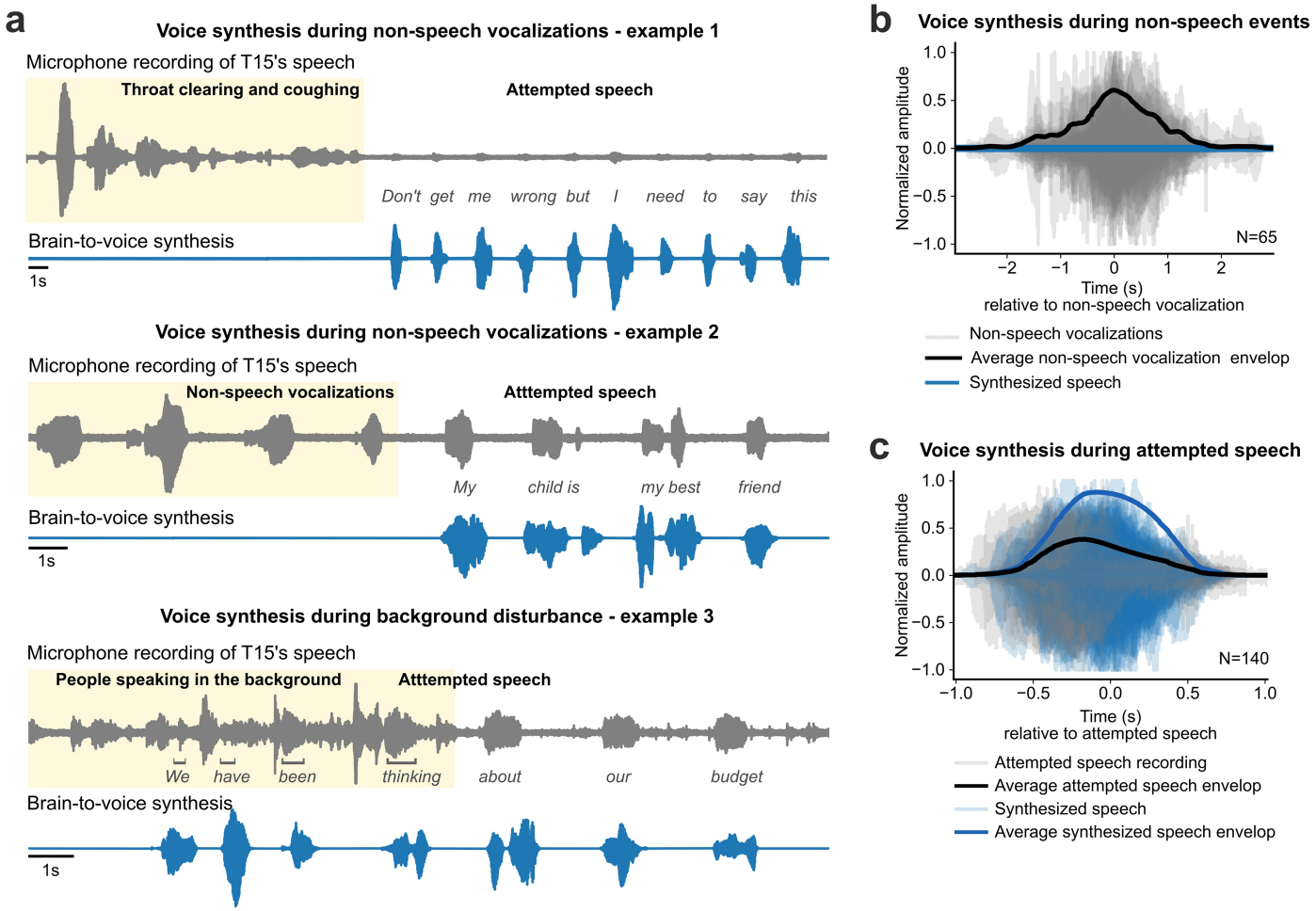


Extended Data Fig. 2 | Additional BCI speech synthesis performance metrics. **a**. Mel-cepstral distortion (MCD) is computed across 25 Mel-frequency bands between the closed-loop synthesized speech and the target speech after removing silences between words. The four subpanels show MCDs (mean ± s.d) between the synthesized and target speech for different speech tasks in evaluation research sessions. **b**. Performance of brain-to-voice decoder measured over time by evaluating neural trials from different sessions offline with a fixed decoder. Decoder trained on post-implant day 165 was fixed and used to synthesize voice offline from neural trials collected in sessions over the next month. Performance was measured by computing Pearson correlation coefficient between the target speech and the synthesized speech across 40 Mel-frequencies after removing silences between words (mean ± s.d., n = 956 sentences). A noticeable decline in brain-to-voice performance was observed after approximately 15 days.

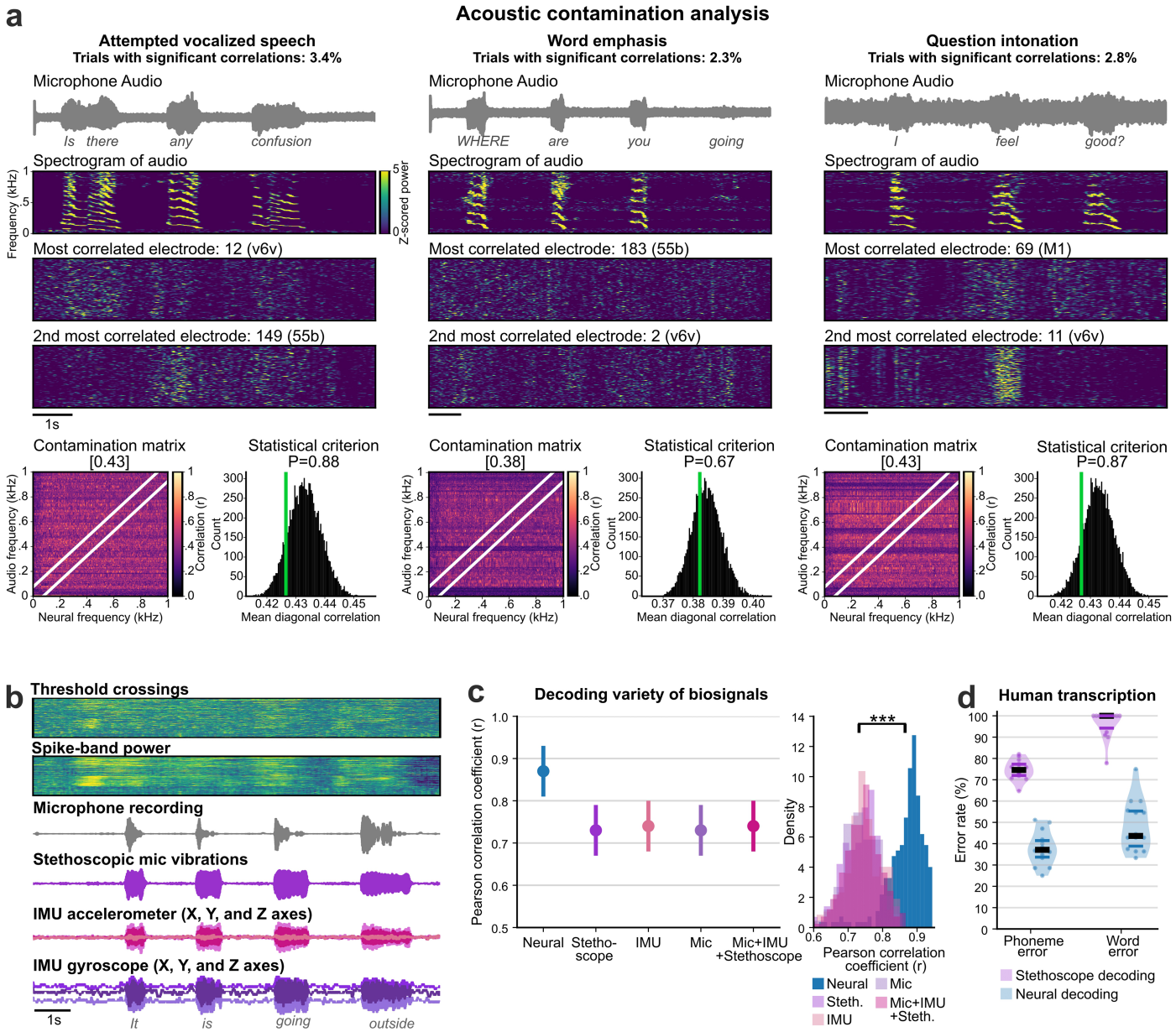


Extended Data Fig. 3 | See next page for caption.

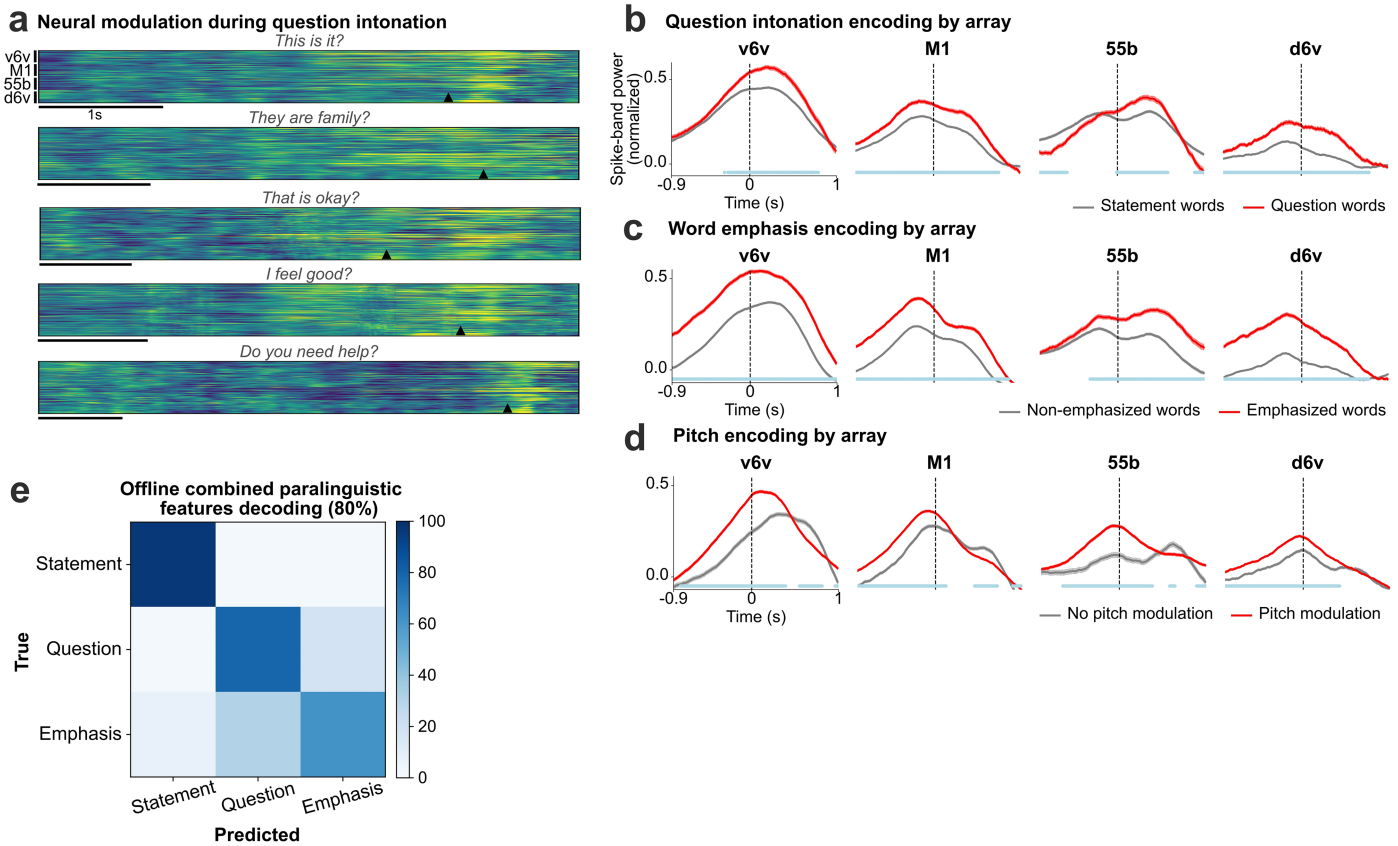
Extended Data Fig. 3 | Electrodes show variability in speech tuning. **a**. Example closed-loop speech synthesis trial. Spike-band power and threshold crossing spikes from each electrode are shown for one example sentence. These neural features were binned and causally normalized and smoothed on a rolling basis before being decoded to synthesize speech. The mean spike-band power and threshold crossing activity for each individual array are also shown. Speech-related modulation was observed on all arrays, with the highest modulation recorded in v6v and 55b. The synthesized speech is shown in the bottom-most row. The grey trace above it shows the participant's attempted (unintelligible) speech as recorded with a microphone. **b, d**. Pearson correlation coefficients of spike-band power and threshold crossings, respectively, between each electrode and the speech envelope (first LPCNet feature predicted by the brain-to-voice decoder). Electrodes are grouped by array and sorted in ascending order to show that different electrodes have different tuning (both positive and negative) with speech. Arrays v6v and 55b have higher correlations with speech and the majority of electrodes show positive tuning. Arrays M1 and d6v have lower correlations with more electrodes tuned negatively. Electrodes with non-significant correlation ($p > 0.05$) are shown in grey. Insets show the same correlations for each electrode arranged spatially in the array. **c, e**. The time course of average spike-band power and threshold crossings across trials of two example electrodes with positive (solid line) and negative (dashed line) correlations from each array (example electrodes are marked by the '+' symbol in (**b**, **d**)). Different electrodes show complex neural dynamics with respect to speech onset and a rich variety in speech tuning. For example, some electrodes have higher activity before speech onset (possibly contributing to speech preparation) while others have higher activity during or after speech onset.



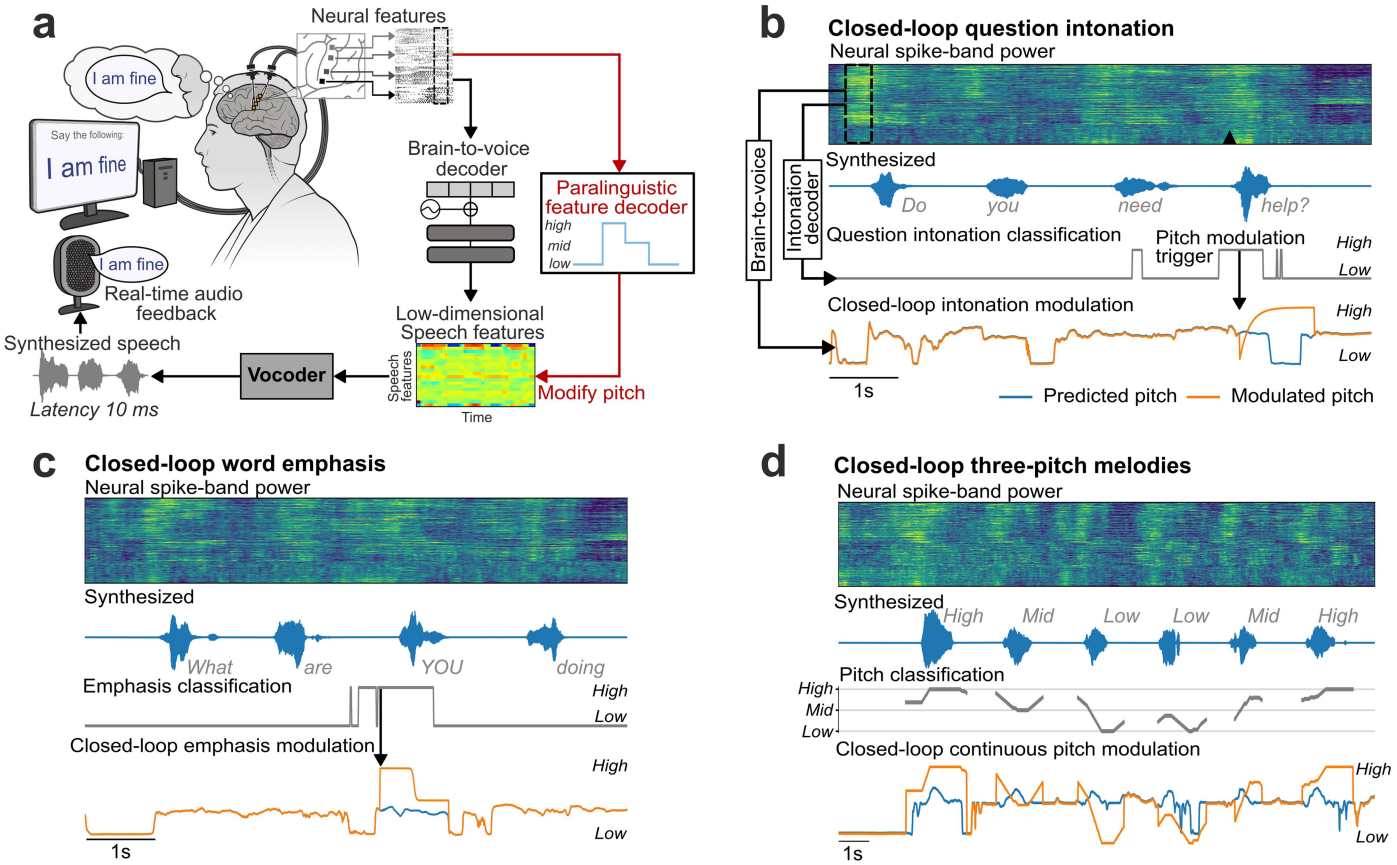
Extended Data Fig. 4 | Speech is not synthesized during non-speech vocalizations or orofacial movements. a. Three example trials show microphone recording (grey) of T15 during attempted speech trials with coughing, throat clearing, non-speech vocalizations or people speaking in the background and the corresponding brain-to-voice synthesis output (blue). The brain-to-voice decoder did not synthesize audible speech and instead output silence during these non-speech vocalizations or when other people were speaking simultaneously (note that T15 starts speaking via the neuroprosthesis midway through the background conversation in example 3). In contrast, it did synthesize voice when T15 voluntarily attempted to speak. Speech was synthesized instantaneously at the exact pace with which T15 attempted to speak and with very low latency. **b.** Samples of non-speech vocalization events (grey) and the corresponding speech synthesis output, which was zero (silence) throughout all these events (blue). **c.** Examples of speech synthesis during attempted speech of each word. Here, the speech is synthesized (blue) appropriately as expected during attempted speech (grey).



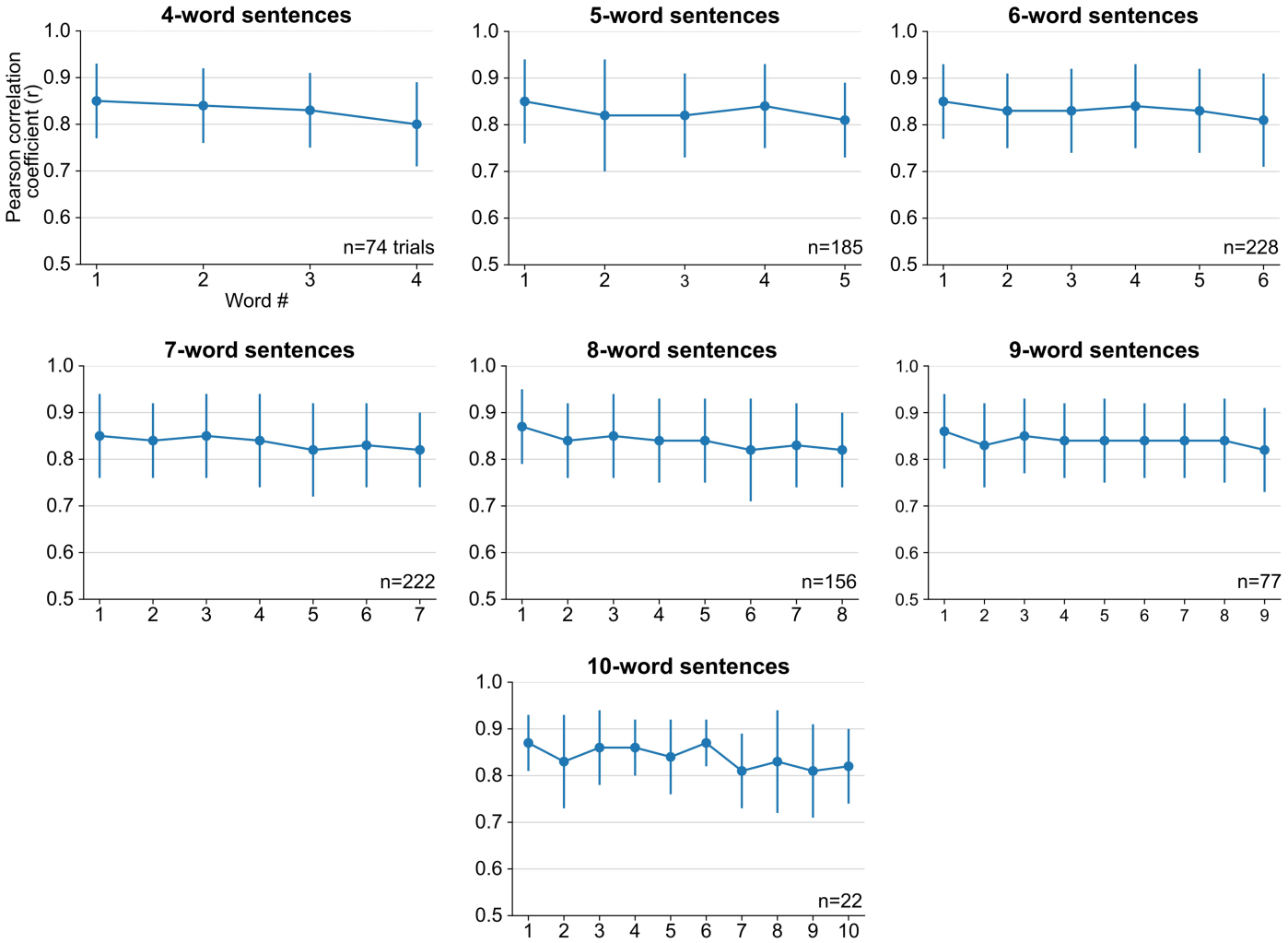
Extended Data Fig. 5 | Neural activity is not contaminated by acoustic artifacts and residual vocalization and movement cannot synthesize intelligible speech. a. Three example trials' audio recording, audio spectrogram, and the spectrograms of the two most acoustic-correlated neural electrodes. Examples are shown for the three types of speech tasks. The prominent spectral structures in the audio spectrogram cannot be observed even in the top two most correlated neural electrodes. An increase in neural activity can be observed before speech onset for each word, reflecting speech preparatory activity and further arguing against acoustic contamination. Note that in the word emphasis example, the last word 'going' is not vocalized fully (there is minimal activity in its audio spectrum), yet an increase in neural activity can be observed that is similar to other words. Contamination matrices and statistical criteria are shown in the bottom row, where *P*-value indicates whether the trial is significantly acoustically contaminated or not. **b.** An example trial of attempted speech with simultaneous recording of intracortical neural signals and various biosignals measured using a microphone, stethoscopic microphone and IMU sensors (accelerometer and gyroscope). Separate independent decoders were trained to synthesize speech using each of the biosignals (or all three together). **c.** Intelligible speech could not be synthesized from biosignals measuring sound, movement, and vibrations during attempted speech. (Left) Cross-validated Pearson correlation coefficients (mean ± s.d.) (compared to target speech) of speech synthesized using neural signals, each of the biosignals, and all biosignals together. Reconstruction accuracy is significantly lower for decoding speech from biosignals as compared to neural activity (two-sided Wilcoxon rank-sum, $P = 10^{-59}$, n = 240 sentences). (Right) Distribution of Pearson correlation coefficients of speech decoding from biosignals and neural signals are mostly non-overlapping, indicating that synthesis quality from biosignals is much lower than that of neural signals. **d.** To assess the intelligibility of voice synthesis from neural activity and biosignals (stethoscopic mic decoder), naive human listeners performed open transcription of (the same) 30 synthesized trials using both the decoders. Median phoneme error rates and word error rates for neural decoding were significantly lower (43.60%) than decoding stethoscope recordings, which had word error rate of 100%. This indicates that intelligible speech cannot be decoded from these non-neural biosignals.



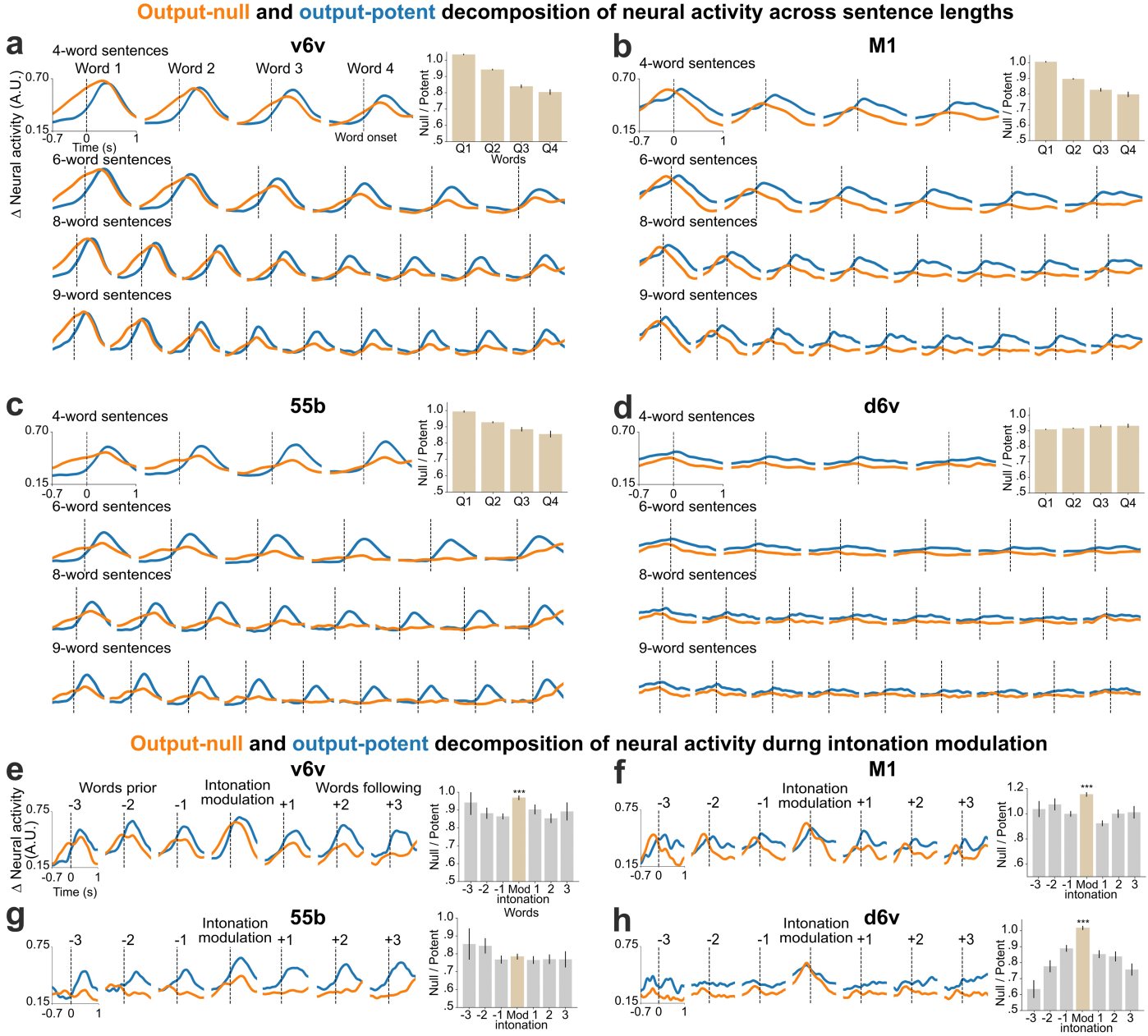
Extended Data Fig. 6 | Encoding of paralinguistic features in neural activity. **a**. Neural modulation during question intonation. Trial-averaged normalized spike-band power (each row in a group is one electrode) during trials where the participant modulated his intonation to say the cued sentence as a question. Trials with the same cue sentence (n = 16) were aligned using dynamic time warping and the mean activity across trials spoken as statements was subtracted to better show the increased neural activity around the intonation-modulation at the end of the sentence. The onset of the word that was pitch-modulated in closed-loop is indicated by the arrowhead at the bottom of each example. **b**. Paralinguistic features encoding recorded from individual arrays. Trial-averaged spike-band power (mean ± s.e.m.), averaged across all electrodes within each array, for words spoken as statements and as questions. At every time point, the spike-band power for statement words and question words were compared using the Wilcoxon rank-sum test. The blue line at the bottom indicates the time points where the spike-band power in statement words and question words were significantly different ($P < 0.001$, $n_1 = 970$ words, $n_2 = 184$ words). **c**. Trial averaged spike-band power across each array for non-emphasized and emphasized words. The spike-band power was significantly different between non-emphasized words and emphasized words at time points shown in blue ($P < 0.001$, $n_1 = 1269$ words, $n_2 = 333$ words). **d**. Trial-averaged spike-band power across each array for words without pitch modulation and words with pitch modulation (from the three-pitch melodies singing task). Words with low and high pitch targets are grouped together as the 'pitch modulation' category (we excluded medium pitch target words where the participant used his normal pitch). The spike-band power was significantly different between no pitch modulation and pitch modulation at time points shown in blue ($P < 0.001$, $n_1 = 486$ words, $n_2 = 916$ words). **e**. Confusion matrix showing offline accuracies for decoding question intonation and word emphasis paralinguistic features together using a single combined 3-class classifier.



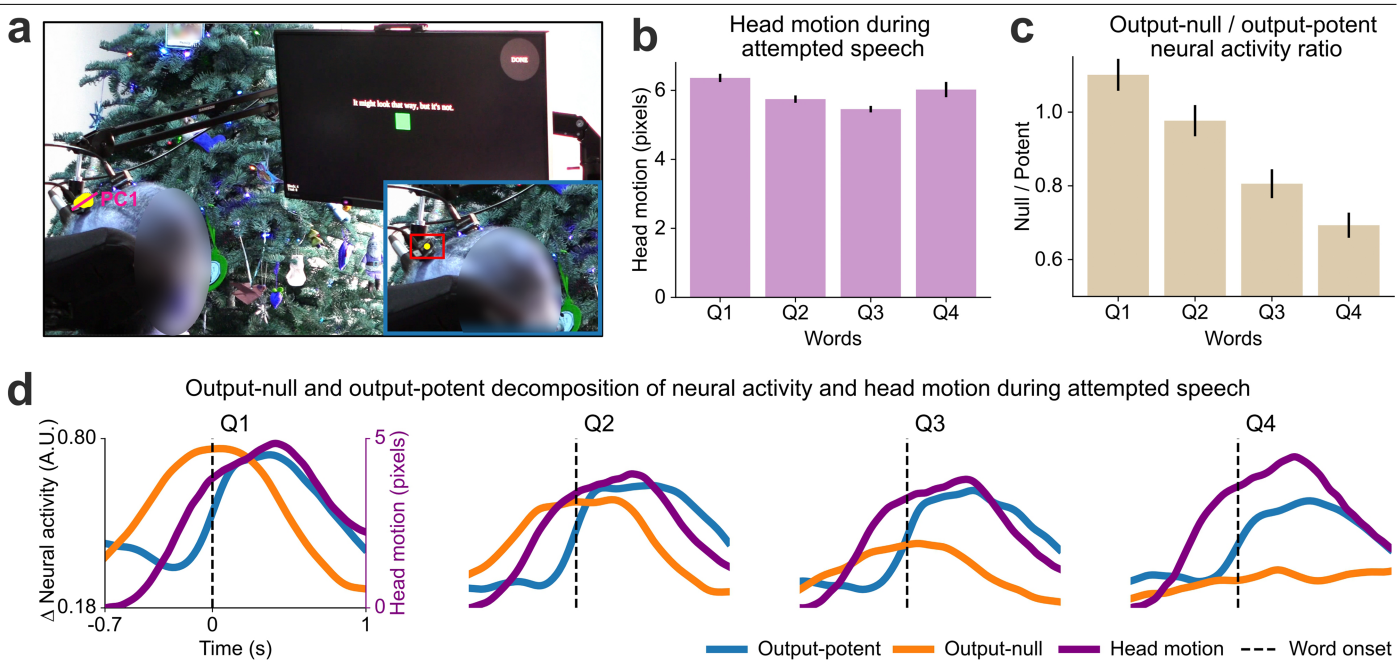
Extended Data Fig. 7 | Closed-loop paralinguistic features modulation. **a**. An overview of the paralinguistic feature decoder and pitch modulation pipeline. An independent paralinguistic feature decoder ran in parallel to the regular brain-to-voice decoder. Its output causally modulated the pitch feature predicted by brain-to-voice, resulting in a pitch-modulated voice. **b**. An example trial of closed-loop intonation modulation for speaking a sentence as a question. A separate binary decoder identified the change in intonation and sent a trigger (downward arrow) to modulate the pitch feature output of the regular brain-to-voice decoder according to a predefined pitch profile for asking a question (low pitch to high pitch). Neural activity of an example trial with its synthesized voice output is shown along with the intonation decoder output, time of modulation trigger (downward arrow), originally predicted pitch feature and the modulated pitch feature used for voice synthesis. **c**. An example trial of closed-loop word emphasis where the word "*YOU*" from "*What are YOU doing*" was emphasized. To emphasize a word, we applied a predefined pitch profile (high pitch to low pitch) along with a 20% increase in the loudness of the predicted speech samples. **d**. An example trial of closed-loop pitch modulation for singing a melody with three pitch levels. The three-pitch classifier output was used to continuously modulate the predicted pitch feature output from the brain-to-voice decoder.



Extended Data Fig. 8 | Pearson correlation coefficients over the course of a sentence. Pearson correlation coefficient (*r*) of individual words in sentences of different lengths (mean ± s.d.). The correlation between target and synthesized speech remained consistent throughout the length of sentence, indicating that the quality of synthesized voice was consistent throughout the sentence. Note that there were fewer longer evaluation sentences.



Extended Data Fig. 9 | Output-null and output-potent neural dynamics during speech production in individual arrays. a-d. Average approximated output-null (orange) and output-potent (blue) components of neural activity during attempted speech of cued sentences of different lengths. Here the neural components are computed for each array independently by training separate linear decoders (i.e., repeating the analyses of Fig. 4 for individual arrays independently). A subset of sentence lengths are shown in the interest of space. Note that the d6v array had much less speech-related modulation. Bar plots within each panel show a summary of all the data (including the not-shown sentence lengths) by taking the average null/potent activity ratios for words in the first-quarter, second-quarter, third-quarter, and fourth-quarter of each sentence (mean ± s.e.m., $n_{Q1} = 3{,}600$, $n_{Q2} = 4{,}181$, $n_{Q3} = 3{,}456$, $n_{Q4} = 3{,}134$ words). **e-h.** Average output-null and output-potent activity during intonation modulation (question-asking or word emphasis) computed separately for each array. Output-null activity shows an increase during intonation modulated word in all arrays. Null/potent activity ratios are summarized in bar plots of intonation-modulated word (red) and the words preceding or following it (grey) (mean ± s.e.m.). The null/potent ratios of modulated words were significantly different from that of non-modulated words for the v6v, M1 and d6v arrays (two-sided Wilcoxon rank-sum, v6v: $p = 10^{-11}$, M1: $p = 10^{-16}$, 55b: $p = 0.3$, d6v: $p = 10^{-26}$, $n_1 = 460$ modulated words, $n_2 = 922$ non-modulated words).



Extended Data Fig. 10 | Head motion during speech and its relationship with the neural dynamics. a. Head motion was tracked from videos by mapping the *x* and *y* positions of the NeuroPort pedestal in each frame (yellow points) using OpenCV. Overall head motion was summarized by the first principal component (pink axis) of *x*-*y* motion. The inset shows a single frame of head motion tracking. **b**. Small head motion was observed during uttering each word. Head motion remained consistent throughout the attempted speech sentence as measured by the motion from baseline during utterance of words in each of the four word-quartiles, regardless of the length of the sentence. **c**. The ratio of output-null and output-potent components of simultaneously recorded neural activity decayed over the course of the sentence, in contrast to the head motion in (**b**). **d**. Time course and amplitude of the output-null and output-potent components of the neural activity and simultaneous head motion in different quartiles of the sentence. The head motion (purple) follows the output-null activity (orange) but precedes the output-potent activity (blue). The output-null activity decayed over the course of the sentence, whereas the head motion during each word in a sentence remained constant. Taken together, this shows that the neural dynamics do not closely match the head motion time course.

nature portfolio

Corresponding author(s): Maitreyee Wairagkar, David Brandman, Sergey Stavisky

Last updated by author(s): Apr 25, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Custom software system was developed in C and Python 3.9.11 for running the experimental tasks, recording the neural data and instantaneous voice synthesis. Software packages used included Tensorflow 2.10.0, Keras 2, LPCNet, Mac OS 13.5 AppKit TTS, Redis 7.4, StyleTTS2, Librosa, Numpy 1.22.3, Numba 0.56.4, Scipy 1.9.1, Sounddevice 0.4.7, Parselmouth 0.4.5, DTW 1.3.0.
Data analysis	Data was analyzed using custom Python code. Code is publicly available on GitHub at https://github.com/Neuroprosthetics-Lab/brain-to-voice-2025

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Neural data related to this study is publicly available on Dryad at https://doi.org/10.5061/dryad.2280gb64f

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	This study included data from one participant, T15, who is a biological male and identifies as a man. This information was self-reported. No sex or gender based analyses were performed given there was only a single participant and the study was developing a brain-computer interface application.
Reporting on race, ethnicity, or other socially relevant groupings	This study assessed brain-computer interface performance for a single participant. No variables relating to race, ethnicity or other socially relevant groupings were reported or analyzed.
Population characteristics	This study includes data from one participant (identified as T15) who gave informed consent and was enrolled in the BrainGate2 clinical trial (ClinicalTrials.gov Identifier: NCT00912041, registered June 3, 2009) but this study did not report clinical trial results. T15 is a left-handed man, 45 years old at the time of data collection with ALS. His ALS symptoms began five years before enrolment into this study.
Recruitment	Participant T15 was enrolled in the BrainGate2 clinical trial after meeting inclusion criteria based in part on disease characteristics. Inclusion and exclusion criteria are available online (ClinicalTrials.gov).
Ethics oversight	The BrainGate2 Neural Interface System clinical trial was approved under an Investigational Device Exemption (IDE) by the US Food and Drug Administration (IDE #G09003). Permission was also granted by the Institutional Review Boards at the University of California, Davis (protocol #1843264 and #1950310) and Mass General Brigham (#2009P000505). All research was performed in accordance with relevant guidelines/regulations.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample-size calculation was performed. Data were collected in a single participant to characterize the performance of a brain-computer interface. Uncertainty in performance estimates were quantified with standard deviation and standard error, and show a robust result.
Data exclusions	This study is based on brain-computer interface performance evaluation data collected over a series of days. All days are reported in the study and all relevant data is included.
Replication	This study assessed brain-computer interface performance with a single participant. Results were replicated across ten independent days of performance evaluation.
Randomization	Randomization into groups is not relevant for this study as only one participant is included in the study.
Blinding	Blinding is not relevant to this study as only one participant was included to asses the performance of a brain-computer interface.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.