https://doi.org/10.1038/s41586-019-1119-1

Speech synthesis from neural decoding of spoken sentences

Gopala K. Anumanchipalli[1,2,4], Josh Chartier[1,2,3,4] & Edward F. Chang[1,2,3]*

Technology that translates neural activity into speech would be transformative for people who are unable to communicate as a result of neurological impairments. Decoding speech from neural activity is challenging because speaking requires very precise and rapid multi-dimensional control of vocal tract articulators. Here we designed a neural decoder that explicitly leverages kinematic and sound representations encoded in human cortical activity to synthesize audible speech. Recurrent neural networks first decoded directly recorded cortical activity into representations of articulatory movement, and then transformed these representations into speech acoustics. In closed vocabulary tests, listeners could readily identify and transcribe speech synthesized from cortical activity. Intermediate articulatory dynamics enhanced performance even with limited data. Decoded articulatory representations were highly conserved across speakers, enabling a component of the decoder to be transferrable across participants. Furthermore, the decoder could synthesize speech when a participant silently mimed sentences. These findings advance the clinical viability of using speech neuroprosthetic technology to restore spoken communication.

Neurological conditions that result in the loss of communication are devastating. Many patients rely on alternative communication devices that measure residual nonverbal movements of the head or eyes[1], or on brain–computer interfaces (BCIs)[2,3] that control a cursor to select letters one-by-one to spell out words. Although these systems can enhance a patient's quality of life, most users struggle to transmit more than 10 words per min, a rate far slower than the average of 150 words per min of natural speech. A major hurdle is how to overcome the constraints of current spelling-based approaches to enable far higher or even natural communication rates.

A promising alternative is to directly synthesize speech from brain activity[4,5]. Spelling is a sequential concatenation of discrete letters, whereas speech is a highly efficient form of communication produced from a fluid stream of overlapping, multi-articulator vocal tract movements[6]. For this reason, a biomimetic approach that focuses on vocal tract movements and the sounds that they produce may be the only means to achieve the high communication rates of natural speech, and is also likely to be the most intuitive for users to learn[7,8]. In patients with paralysis—caused by for example, amyotrophic lateral sclerosis or brainstem stroke—high-fidelity speech-control signals may only be accessed by directly recording from intact cortical networks.

Our goal was to demonstrate the feasibility of a neural speech prosthetic by translating brain signals into intelligible synthesized speech at the rate of a fluent speaker. To accomplish this, we recorded high-density electrocorticography (ECoG) signals from five participants who underwent intracranial monitoring for epilepsy treatment as they spoke several hundreds of sentences aloud. We designed a recurrent neural network that decoded cortical signals with an explicit intermediate representation of the articulatory dynamics to synthesize audible speech.

Speech decoder design

The two-stage decoder approach is shown in Fig. 1a–d. Stage 1, a bidirectional long short-term memory (bLSTM) recurrent neural network[9], decodes articulatory kinematic features from continuous neural activity (high-gamma amplitude envelope[10] and low frequency component[11,12], see Methods) recorded from ventral sensorimotor cortex (vSMC)[13], superior temporal gyrus (STG)[14] and inferior frontal gyrus (IFG)[15] (Fig. 1a, b). Stage 2, a separate bLSTM, decodes acoustic features (pitch (F_0), mel-frequency cepstral coefficients (MFCCs), voicing and glottal excitation strengths) from the decoded articulatory features from stage 1 (Fig. 1c). The audio signal is then synthesized from the decoded acoustic features (Fig. 1d). To integrate the two stages of the decoder, stage 2 (articulation-to-acoustics) was trained directly on output of stage 1 (brain-to-articulation) so that it not only learns the transformation from kinematics to sound, but also corrects articulatory estimation errors made in stage 1.

A key component of our decoder is the intermediate articulatory representation between neural activity and acoustics (Fig. 1b). This step is crucial because the vSMC exhibits robust neural activations during speech production that predominantly encode articulatory kinematics[16,17]. Because articulatory tracking of continuous speech was not feasible in our clinical setting, we used a statistical approach to estimate vocal tract kinematic trajectories (movements of the lips, tongue and jaw) and other physiological features (for example, manner of articulation) from audio recordings. These features initialized the bottleneck layer within a speech encoder–decoder that was trained to reconstruct a participant's produced speech acoustics (see Methods). The encoder was then used to infer the intermediate articulatory representation used to train the neural decoder. With this decoding strategy, it was possible to accurately reconstruct the speech spectrogram.

Synthesis performance

Overall, we observed detailed reconstructions of speech synthesized from neural activity alone (see Supplementary Video 1). Figure 1e, f shows the audio spectrograms from two original spoken sentences plotted above those decoded from brain activity. The decoded spectrogram retained salient energy patterns that were present in the original spectrogram and correctly reconstructed the silence in between

[1]Department of Neurological Surgery, University of California San Francisco, San Francisco, CA, USA. [2]Weill Institute for Neurosciences, University of California San Francisco, San Francisco, CA, USA. [3]University of California Berkeley and University of California San Francisco Joint Program in Bioengineering, Berkeley, CA, USA. [4]These authors contributed equally: Gopala K. Anumanchipalli, Josh Chartier. *e-mail: Edward.Chang@ucsf.edu

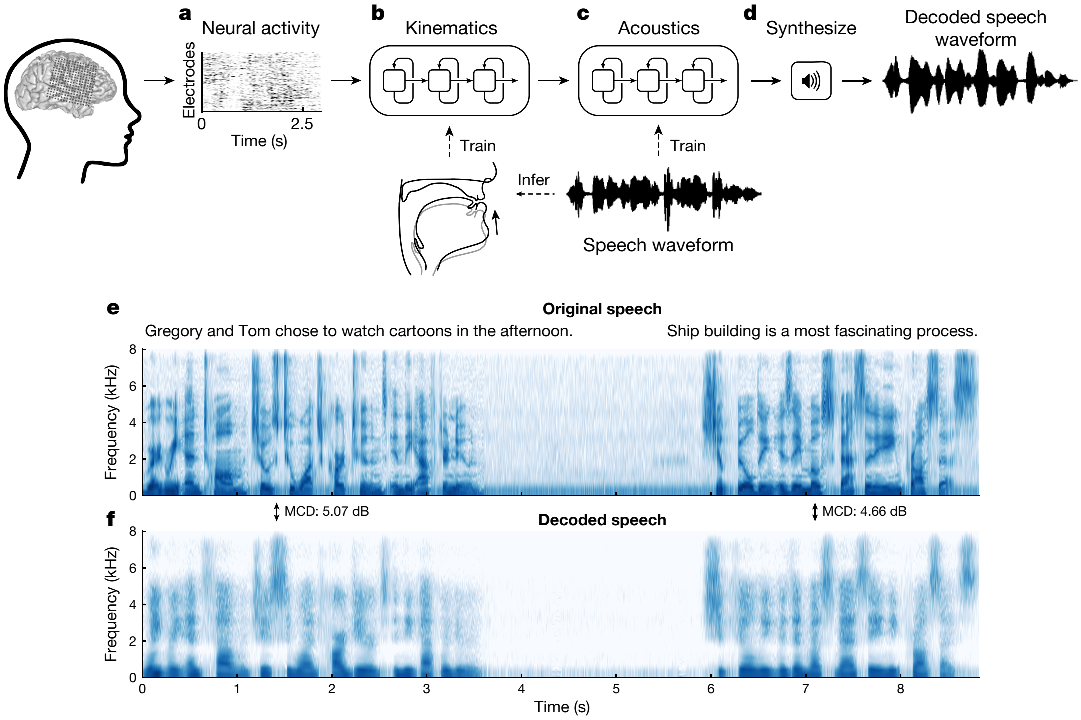

Fig. 1 | Speech synthesis from neurally decoded spoken sentences. **a**, The neural decoding process begins by extracting relevant signal features from high-density cortical activity. **b**, A bLSTM neural network decodes kinematic representations of articulation from ECoG signals. **c**, An additional bLSTM decodes acoustics from the previously decoded kinematics. Acoustics are spectral features (for example, MFCCs) extracted from the speech waveform. **d**, Decoded signals are synthesized into an acoustic waveform. **e**, Spectrogram shows the frequency content of two sentences spoken by a participant. **f**, Spectrogram of synthesized speech from brain signals recorded simultaneously with the speech in **e** (repeated five times with similar results). MCD was computed for each sentence between the original and decoded audio. Fivefold cross-validation was used to find consistent decoding.

the sentences when the participant was not speaking. Extended Data Figure 1a, b illustrates the quality of reconstruction at the phonetic level. Median spectrograms of original and synthesized phonemes—units of sound that distinguish one word from another—showed that the typical spectrotemporal patterns were preserved in the decoded examples (for example, resonant frequency bands in the spectrograms called formants F_1–F_3 in vowels /iː/ and /æ/; and key spectral patterns of mid-band energy and broadband burst for consonants /z/ and /p/, respectively).

To understand to what degree the synthesized speech was perceptually intelligible to naive listeners, we conducted two listening tasks that involved single-word identification and sentence-level transcription, respectively. The tasks were run on Amazon Mechanical Turk (see Methods), using all 101 sentences from the test set of participant 1.

For the single-word identification task, we evaluated 325 words that were spliced from the synthesized sentences. We quantified the effect of word length (number of syllables) and the number of choices (10, 25 and 50 words) on speech intelligibility, since these factors inform optimal design of speech interfaces[18]. Overall, we found that listeners were more successful at word identification as syllable length increased, and the number of word choices decreased (Fig. 2a), consistent with natural speech perception[19].

For sentence-level intelligibility, we designed a closed vocabulary, free transcription task. Listeners heard the entire synthesized sentence and transcribed what they heard by selecting words from a defined pool (of either 25 or 50 words) that included the target words and random words from the test set. The closed vocabulary setting was necessary because the test set was a subset of sentences from MOCHA-TIMIT[20], which was primarily designed to optimize articulatory coverage of English but contains highly unpredictable sentence constructions and low-frequency words.

Listeners were able to transcribe synthesized speech well. Of the 101 synthesized trials, at least one listener was able to provide a perfect transcription for 82 sentences with a 25-word pool and 60 sentences with a 50-word pool. Of all submitted responses, listeners transcribed 43% and 21% of the trials perfectly, respectively (Extended Data Fig. 2). Figure 2b shows the distributions of mean word error rates (WER) of each sentence. Transcribed sentences had a median 31% WER with a 25-word pool size and 53% WER with a 50-word pool size. Table 1 shows listener transcriptions for a range of WERs. Median level transcriptions still provided a fairly accurate, and in some cases legitimate, transcription (for example, ‘mum’ transcribed as ‘mom’). The errors suggest that the acoustic phonetic properties of the phonemes are still present in the synthesized speech, albeit to the lesser degree (for example, ‘rabbits’ transcribed as ‘rodents’). This level of intelligibility for neurally synthesized speech would already be immediately meaningful and practical for real world application.

We then quantified the decoding performance at a feature level for all participants. In speech synthesis, the spectral distortion of synthesized speech from ground-truth is commonly reported using the mean mel-cepstral distortion (MCD)[21]. Mel-frequency bands emphasize the distortion of perceptually relevant frequency bands of the audio spectrogram[22]. We compared the MCD of neurally synthesized speech to a reference synthesis from articulatory kinematics and chance-level decoding (a lower MCD is better; Fig. 2c). The reference synthesis simulates perfect neural decoding of the kinematics. For our five participants (participants 1–5), the median MCD scores of decoding speech ranged from 5.14 dB to 6.58 dB ($P < 1 \times 10^{-18}$, Wilcoxon signed-rank test, for each participant).

We also computed the correlations between original and decoded acoustic features. For each sentence and feature, the Pearson’s correlation coefficient was computed using every sample (at 200 Hz) for that feature. The sentence correlations between the mean decoded acoustic features (consisting of intensity, MFCCs, excitation strengths and voicing) and inferred kinematics across participants are plotted in Fig. 2d. Prosodic features such as pitch (F_0), speech envelope and voicing were

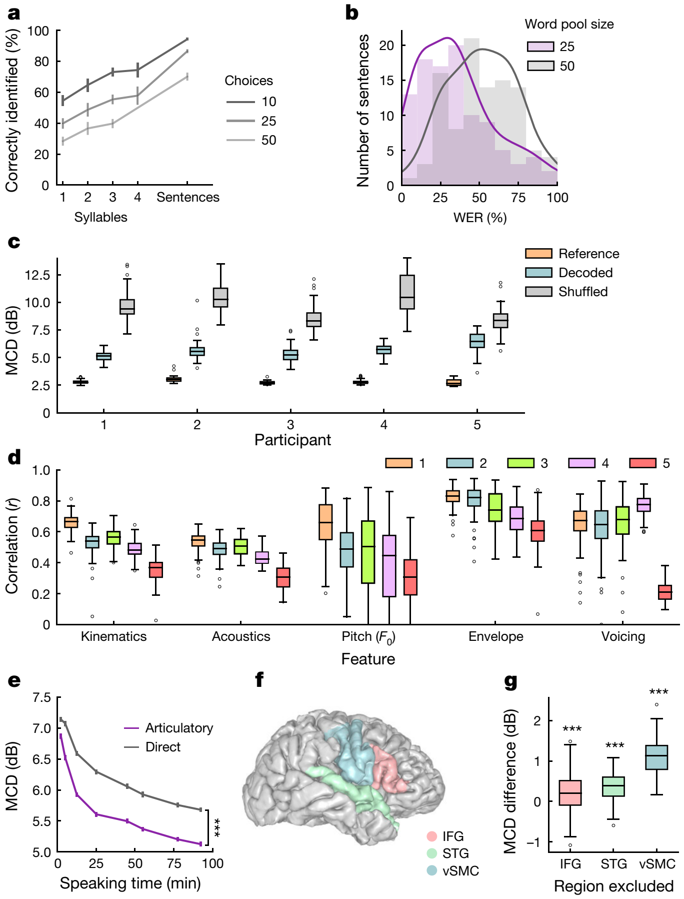

Fig. 2 | Synthesized speech intelligibility and feature-specific performance. a, Listening tests for identification of excerpted single words ($n = 325$) and full sentences ($n = 101$) for synthesized speech from participant 1. Points represent mean word identification rate. Words were grouped by syllable length ($n = 75$, 158, 68 and 24, respectively, for one, two, three and four syllables). Listeners identified speech by selecting from a set of choices (10, 25 or 50 words). Data are mean ± s.e.m. **b**, Listening tests for closed vocabulary transcription of synthesized sentences ($n = 101$). Responses were constrained in word choice (25 or 50), but not in sequence length. Outlines are kernel density estimates of the distributions. **c**, Spectral distortion, measured by MCD (lower values are better), between original spoken sentences and neurally decoded sentences ($n = 101$, 100, 93, 81 and 44, respectively, for participants 1–5). Reference MCD refers to the synthesis of original (inferred) kinematics without neural decoding. **d**, Correlation of original and decoded kinematic and acoustic features ($n = 101$, 100, 93, 81 and 44 sentences, respectively, for participants 1–5). Kinematic and acoustic values represent mean correlation of 33 and 32 features, respectively. **e**, Mean MCD of sentences ($n = 101$) decoded from models trained on varying amounts of training data. The neural decoder with an articulatory intermediate stage (purple) performed better than the direct ECoG to acoustics decoder (grey). All data sizes: $n = 101$ sentences; $P < 1 \times 10^{-5}$, Wilcoxon signed-rank test. **f**, Anatomical reconstruction of the brain of participant 1 with the following regions used for neural decoding: ventral sensorimotor cortex (vSMC), superior temporal gyrus (STG) and inferior frontal gyrus (IFG). **g**, Difference in median MCD of sentences ($n = 101$) between decoder trained on all regions and decoders trained on all-but-one region. Exclusion of any region resulted in decreased performance. $n = 101$ sentences; $P < 3 \times 10^{-4}$, Wilcoxon signed-rank test. All box plots depict median (horizontal line inside box), 25th and 75th percentiles (box), 25th or 75th percentiles ±1.5× interquartile range (whiskers) and outliers (circles). Distributions were compared with each as other as indicated or with chance-level distributions using two-tailed Wilcoxon signed-rank tests. ***$P < 0.001$.

Table 1 | Listener transcriptions of neurally synthesized speech

Word error rate	Original sentences and transcriptions of synthesized speech
0%	o: Is this seesaw safe t: Is this seesaw safe
~10%	o: Bob bandaged both wounds with the skill of a doctor t: Bob bandaged full wounds with the skill of a doctor
~20%	o: Those thieves stole thirty jewels t: Thirty thieves stole thirty jewels
	o: Help celebrate brother's success t: Help celebrate his brother's success
~30%	o: Get a calico cat to keep the rodents away t: The calico cat to keep the rabbits away
	o: Carl lives in a lively home t: Carl has a lively home
~50%	o: Mum strongly dislikes appetizers t: Mom often dislikes appetizers
	o: Etiquette mandates compliance with existing regulations t: Etiquette can be made with existing regulations
>70%	o: At twilight on the twelfth day we'll have Chablis t: I was walking through Chablis

Examples are shown for several word error rate levels. The original text is indicated by 'o' and the listener transcriptions are indicated by 't'.

decoded well above the level expected by chance ($r > 0.6$, except F_0 for participant 2: $r = 0.49$ and all features for participant 5; $P < 1 \times 10^{-10}$, Wilcoxon signed-rank test, for all participants and features in Fig. 2d). Correlation decoding performance for all other features is shown in Extended Data Fig. 4a, b.

Decoder characteristics

The following analyses were performed on data from participant 1. When designing a neural decoder for clinical applications, there are several key considerations that determine model performance. First, in patients with severe paralysis or limited speech ability, training data may be very difficult to obtain. Therefore, we assessed the amount of data that would be necessary to achieve a high level of performance. We found a clear advantage in explicitly modelling articulatory kinematics as an intermediate step over decoding acoustics directly from the ECoG signals. The 'direct' decoder was a bLSTM recurrent neural network that was optimized for decoding acoustics (MFCCs) directly from same ECoG signals as used in an articulatory decoder. We found robust performance could be achieved with as little as 25 min of speech, but performance continued to improve with the addition of data (Fig. 2e). Without the articulatory intermediate step, the direct ECoG to acoustic decoding MCD was offset by 0.54 dB (0.2 dB is perceptually noticeable[21]) using the full dataset (Fig. 3a; $n = 101$, $P = 1 \times 10^{-17}$, Wilcoxon signed-rank test).

This performance gap between the two approaches persisted with increasing data sizes. One interpretation is that aspects of kinematics are more preferentially represented by cortical activity than acoustics[16], and are thus learned more quickly with limited data. Another aspect that may underlie this difference is that articulatory kinematics lie on a low-dimensional manifold that constrains the potential high-dimensionality of acoustic signals[6,7,23] (Extended Data Fig. 5). Therefore, separating out the high-dimensional translation of articulation to speech, as done in stage 2 of our decoder may be critical for performance. It is possible that with sufficiently large datasets both decoding approaches would converge on one another.

Second, we wanted to understand the phonetic properties that were preserved in synthesized speech. We used Kullback–Leibler divergence to compare the distribution of spectral features of each decoded phoneme to those of each ground-truth phoneme to determine how similar they were (Extended Data Fig. 6). We expected that, in addition to the same decoded and ground-truth phoneme being similar to one another, phonemes with shared acoustic properties would also be characterized as similar to one another.

Hierarchical clustering based on the Kullback–Leibler divergence of each phoneme pair demonstrated that phonemes were clustered into

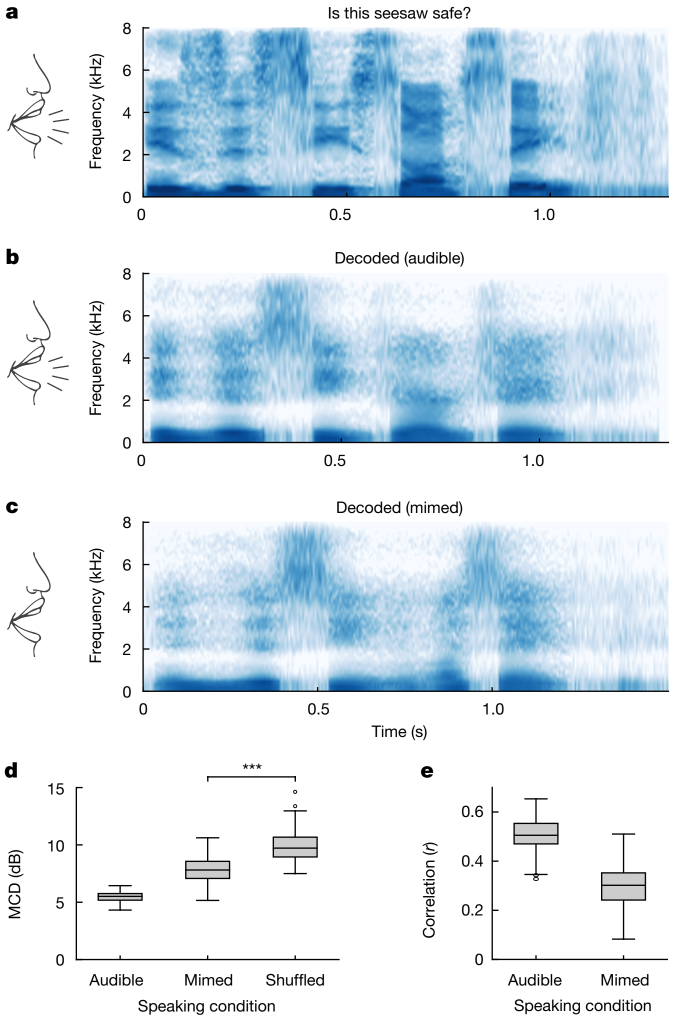

Fig. 3 | Speech synthesis from neural decoding of silently mimed speech. a–c, Spectrograms of original spoken sentence (**a**), neural decoding from audible production (**b**) and neural decoding from silently mimed production (**c**) (repeated five times with similar results). **d**, **e**, MCD (**d**) and correlation of original and decoded spectral features (**e**) for audibly and silently produced speech ($n = 58$ sentences). Decoded sentences were significantly better than chance-level decoding for both speaking conditions. $n = 58$; audible, $P = 3 \times 10^{-11}$; mimed, $P = 5 \times 10^{-11}$, Wilcoxon signed-rank test. Box plots as described in Fig. 2. ***P < 0.001.

four main groups. Group 1 contained consonants with an alveolar place of constriction (for example, /s/ and /t/). Group 2 contained almost all other consonants (for example, /f/ and /g/). Group 3 contained mostly high vowels (for example, /i/ and /u/). Group 4 contained mostly mid and low vowels (for example, /ɑ/ and /æ/). The difference between groups tended to correspond to variations along acoustically significant dimensions (frequency range of spectral energy for consonants and formants for vowels). Indeed, these groupings explain some of the confusions reflected in listener transcriptions of these stimuli. This hierarchical clustering was also consistent with the acoustic similarity matrix of only ground-truth phoneme pairs (Extended Data Fig. 7; cophenetic correlation[24] $= 0.71$, $P = 1 \times 10^{-10}$).

Third, because the success of the decoder depends on the initial electrode placement, we quantified the contribution of several anatomical regions (vSMC, STG and IFG) that are involved in continuous speech production[25]. Decoders were trained in a leave-one-region-out fashion, for which all electrodes from a particular region were held out (Fig. 2f). Removing any region led to some decrease in decoder performance (Fig. 2g; $n = 101$, $P = 3 \times 10^{-4}$, Wilcoxon signed-rank test). However, excluding the vSMC resulted in the largest decrease in performance (MCD increase of 1.13 dB).

Fourth, we investigated whether the decoder generalized to novel sentences that were never seen in the training data. Because participant 1 produced some sentences multiple times, we compared two decoders: one that was trained on all sentences (not the particular instances in the test set), and one that was trained excluding every instance of the sentences in the testing set. We found no significant difference in decoding performance of the sentences for both MCD and correlations of spectral features ($P = 0.36$ and $P = 0.75$, respectively, $n = 51$, Wilcoxon signed-rank test; Extended Data Fig. 8). Notably, this suggests that the decoder can generalize to arbitrary words and sentences that the decoder was never trained on.

Synthesizing mimed speech

To rule out the possibility that the decoder is relying on the auditory feedback of participants' vocalization, and to simulate a setting in which subjects do not overtly vocalize, we tested our decoder on silently mimed speech. We tested a held-out set of 58 sentences in which the participant 1 audibly produced each sentence and then mimed the same sentence, making the same articulatory movements but without making sound. Even though the decoder was not trained on mimed sentences, the spectrograms of synthesized silent speech demonstrated similar spectral patterns to synthesized audible speech of the same sentence (Fig. 3a–c). With no original audio to compare, we quantified performance of the synthesized mimed sentences with the audio from the trials with spoken sentences. We calculated the spectral distortion and correlation of the spectral features by first dynamically time-warping the spectrogram of the synthesized mimed speech to match the temporal profile of the audible sentence (Fig. 3d, e) and then comparing performance. Although synthesis performance for mimed speech was inferior to the performance for audible speech—which is probably due to absence of phonation signals during miming—this demonstrates that it is possible to decode important spectral features of speech that were never audibly uttered ($P < 1 \times 10^{-11}$ compared to chance, $n = 58$; Wilcoxon signed-rank test) and that the decoder did not rely on auditory feedback.

State–space of decoded speech articulation

Our findings suggest that modelling the underlying kinematics enhances the decoding performance, so we next wanted to better understand the nature of the decoded kinematics from population neural activity. We examined low-dimensional kinematic state–space trajectories, by computing the state–space projection using principal components analysis onto the articulatory kinematic features. The first ten principal components (of 33 components in total) captured 85% of the variance and the first two principal components captured 35% (Extended Data Fig. 5).

We projected the kinematic trajectory of an example sentence onto the first two principal components (Fig. 4a, b). These trajectories were well decoded, as shown in the example (Pearson's correlation: $r = 0.91$ and $r = 0.91$, principal components 1 and 2, respectively; Fig. 4a, b), and summarized across all test sentences and participants (median $r > 0.72$ for all participants except participant 5, where r represents the mean r of first two principal components, Fig. 4e). Furthermore, state–space trajectories of mimed speech were well decoded (median $r = 0.6$, $P = 1 \times 10^{-5}$, $n = 38$, Wilcoxon signed-rank test; Fig. 4e).

The state–space trajectories appeared to manifest the dynamics of syllabic patterns in continuous speech. The time courses of consonants and vowels were plotted on the state–space trajectories and tended to correspond with the troughs and peaks of the trajectories, respectively (Fig. 4a, b). Next, we sampled from every vowel-to-consonant transition ($n = 22{,}453$) and consonant-to-vowel transition ($n = 22{,}453$), and plotted 500-ms traces of the average trajectories for principal components 1 and 2 centred at the time of transition (Fig. 4c, d). Both types of trajectories were biphasic in nature, transitioning from the 'high' state during the vowel to the 'low' state during the consonant and vice versa. When examining transitions of specific phonemes, we found that principal components 1 and 2 retained their biphasic trajectories of vowel or consonant states, but showed specificity towards particular

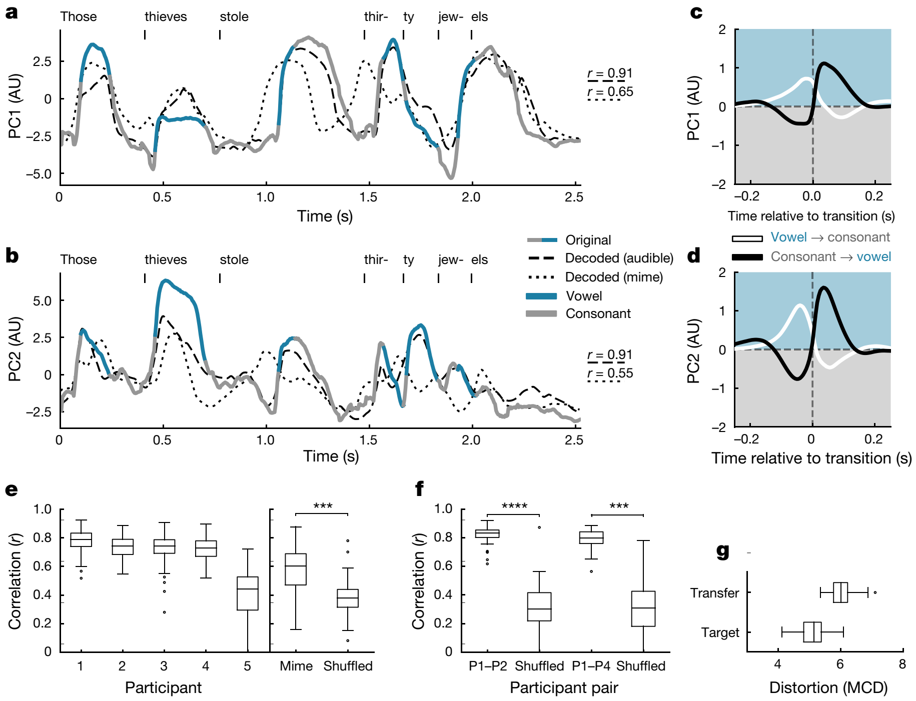

Fig. 4 | Kinematic state–space representation of speech production. **a**, **b**, A kinematic trajectory (grey–blue) from a single trial (participant 1) projected onto the first two principal components—principal components (PC)1 (**a**) and 2 (**b**)—of the kinematic state–space. Decoded audible (dashed) and mimed (dotted) kinematic trajectories are also plotted. Pearson's r, $n = 510$ time samples. The trajectory for mimed speech was uniformly stretched to align with the audible speech trajectory for visualization as it occurred at a faster time scale. **c**, **d**, Average trajectories for principal components 1 (**c**) and 2 (**d**) from **a** and **b**, respectively, for transitions from a vowel to a consonant (black, $n = 22{,}453$) and from a consonant to a vowel (white, $n = 22{,}453$). Time courses are 500 ms. **e**, Distributions of correlations between original and decoded kinematic state–space trajectories (averaged across principal components 1 and 2) ($n = 101, 100, 93, 81, 44$ sentences, respectively, for participants 1–5). Pearson's correlations for mimed trajectories were calculated by dynamically time-warping to the audible production of the same sentence and then compared to correlations of the dynamically time-warping of a randomly selected sentence trajectory. $n = 58$ sentences; $***P = 1 \times 10^{-5}$, Wilcoxon signed-rank test. **f**, Distributions of correlations for state–space trajectories of the same sentence across participants. Alignment between participants was done by dynamically time-warping and compared to correlations of dynamically time-warping of unmatched sentence pairs. $n = 92$; $****P = 1 \times 10^{-16}$ and $n = 44$; $***P = 1 \times 10^{-8}$, respectively, Wilcoxon signed-rank test. **g**, Comparison between acoustic decoders (stage 2) ($n = 101$ sentences). 'Target' refers to an acoustic decoder trained on data from the same participant as the kinematic decoder (stage 1) is trained on (participant 1). 'Transfer' refers to an acoustic decoder that was trained on kinematics and acoustics from a different participant (participant 2). Box plots as described in Fig. 2.

phonemes, indicating that principal components 1 and 2 do not necessarily describe only jaw opening and closing, but rather describe global opening and closing configurations of the vocal tract (Extended Data Fig. 9). These findings are consistent with theoretical accounts of human speaking behaviour, which postulate that high-dimensional speech acoustics lie on a low-dimensional articulatory state–space[6].

To evaluate the similarity of the decoded state–space trajectories, we correlated productions of the same sentence across participants that were projected onto their respective kinematic state–spaces (only participants 1, 2 and 4 had comparable sentences). The state–space trajectories were highly similar ($r > 0.8$; Fig. 4f), suggesting that the decoder is probably relying on a shared representation across speakers, a critical basis for generalization.

A shared kinematic representation across speakers could be very advantageous for someone who cannot speak as it may be more intuitive and faster to learn to use the kinematics decoder (stage 1), while using an existing kinematics-to-acoustics decoder (stage 2) trained on speech data collected independently. We show synthesis performance when transferring stage 2 from a source participant (participant 1) to a target participant (participant 2) (Fig. 4g). The acoustic transfer performed well, although less than when both stage 1 and stage 2 were trained on the target (participant 2), probably because the MCD metric is sensitive to speaker identity.

Discussion

Here we demonstrate speech synthesis using high-density, direct cortical recordings from the human speech cortex. Previous strategies for neural decoding of speech production focused on direct classification of speech segments such as phonemes or words[26,27]; however, these approaches are generally limited in their ability to scale to larger vocabulary sizes and communication rates. Meanwhile, sensory decoding of auditory cortex has been promising for speech sounds[28–30] or for auditory imagery[31] in part because of the direct relationship between the auditory encoding of spectrotemporal information and the reconstructed spectrogram. An outstanding question has been whether motor decoding of vocal tract movements during speech production could be used for generating high-fidelity acoustic speech output.

Previous work focused on understanding movement that was encoded at single electrodes[16]; however, a fundamentally different challenge for speech synthesis is decoding the population activity that addresses the complex mapping between vocal tract movements and sounds. Natural speech production involves over 100 muscles and the mapping from movement to sounds is not one-to-one. Our decoder explicitly incorporated this knowledge to simplify the translation of neural activity to sound by first decoding the primary physiological correlate of neural activity and then transforming to speech acoustics. This statistical mapping permits generalization with limited amounts of training.

Direct speech synthesis has several major advantages over spelling-based approaches. In addition to the capability to communicate unconstrained vocabularies at a natural speaking rate, it captures prosodic elements of speech that are not available with text output, such as pitch intonation[32]. Furthermore, a practical limitation for current alternative communication devices is the cognitive effort required to learn and use them. For patients in whom the cortical processing of articulation is still intact, a speech-based BCI decoder may be far more intuitive and easier to learn to use[7,8].

BCIs are rapidly becoming a clinically viable means to restore lost function. Neural prosthetic control was first demonstrated in participants without disabilities[33–35] before translating the technology to participants with tetraplegia[36–39]. Our findings represent one step forward for addressing a major challenge posed by patients who are paralysed and cannot speak. The generalization results presented here demonstrate that speakers share a similar kinematic state–space representation that is speaker-independent and that it is possible to transfer model knowledge about the mapping of kinematics to sound across subjects. Tapping into this emergent, low-dimensional representation of neural activity from a coordinated population in the intact cortex may be a critical for bootstrapping a decoder[23], as well facilitating BCI learning[7]. Our results may be an important next step in realizing speech restoration for patients with paralysis.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-019-1119-1.

Received: 29 October 2018; Accepted: 21 March 2019;
Published online 24 April 2019.

1. Fager, S. K., Fried-Oken, M., Jakobs, T. & Beukelman, D. R. New and emerging access technologies for adults with complex communication needs and severe motor impairments: state of the science. *Augment. Altern. Commun.* https://doi.org/10.1080/07434618.2018.1556730 (2019).
2. Brumberg, J. S., Pitt, K. M., Mantie-Kozlowski, A. & Burnison, J. D. Brain–computer interfaces for augmentative and alternative communication: a tutorial. *Am. J. Speech Lang. Pathol.* **27**, 1–12 (2018).
3. Pandarinath, C. et al. High performance communication by people with paralysis using an intracortical brain–computer interface. *eLife* **6**, e18554 (2017).
4. Guenther, F. H. et al. A wireless brain–machine interface for real-time speech synthesis. *PLoS ONE* **4**, e8218 (2009).
5. Bocquelet, F., Hueber, T., Girin, L., Savariaux, C. & Yvert, B. Real-time control of an articulatory-based speech synthesizer for brain computer interfaces. *PLOS Comput. Biol.* **12**, e1005119 (2016).
6. Browman, C. P. & Goldstein, L. Articulatory phonology: an overview. *Phonetica* **49**, 155–180 (1992).
7. Sadtler, P. T. et al. Neural constraints on learning. *Nature* **512**, 423–426 (2014).
8. Golub, M. D. et al. Learning by neural reassociation. *Nat. Neurosci.* **21**, 607–616 (2018).
9. Graves, A. & Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **18**, 602–610 (2005).
10. Crone, N. E. et al. Electrocorticographic gamma activity during word production in spoken and sign language. *Neurology* **57**, 2045–2053 (2001).
11. Nourski, K. V. et al. Sound identification in human auditory cortex: differential contribution of local field potentials and high gamma power as revealed by direct intracranial recordings. *Brain Lang.* **148**, 37–50 (2015).
12. Pesaran, B. et al. Investigating large-scale brain dynamics using field potential recordings: analysis and interpretation. *Nat. Neurosci.* **21**, 903–919 (2018).
13. Bouchard, K. E., Mesgarani, N., Johnson, K. & Chang, E. F. Functional organization of human sensorimotor cortex for speech articulation. *Nature* **495**, 327–332 (2013).
14. Mesgarani, N., Cheung, C., Johnson, K. & Chang, E. F. Phonetic feature encoding in human superior temporal gyrus. *Science* **343**, 1006–1010 (2014).
15. Flinker, A. et al. Redefining the role of Broca's area in speech. *Proc. Natl Acad. Sci. USA* **112**, 2871–2875 (2015).
16. 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).
17. Mugler, E. M. et al. Differential representation of articulatory gestures and phonemes in precentral and inferior frontal gyri. *J. Neurosci.* **38**, 9803–9813 (2018).
18. Huggins, J. E., Wren, P. A. & Gruis, K. L. What would brain–computer interface users want? Opinions and priorities of potential users with amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler.* **12**, 318–324 (2011).
19. Luce, P. A. & Pisoni, D. B. Recognizing spoken words: the neighborhood activation model. *Ear Hear.* **19**, 1–36 (1998).
20. Wrench, A. MOCHA: multichannel articulatory database. http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html (1999).
21. Kominek, J., Schultz, T. & Black, A. Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In *Proc. The first workshop on Spoken Language Technologies for Under-resourced languages (SLTU-2008)* 63–68 (2008).
22. Davis, S. B. & Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In Readings in speech recognition. *IEEE Trans. Acoust.* **28**, 357–366 (1980).
23. Gallego, J. A., Perich, M. G., Miller, L. E. & Solla, S. A. Neural manifolds for the control of movement. *Neuron* **94**, 978–984 (2017).
24. Sokal, R. R. & Rohlf, F. J. The comparison of dendrograms by objective methods. *Taxon* **11**, 33–40 (1962).
25. Brumberg, J. S. et al. Spatio-temporal progression of cortical activity related to continuous overt and covert speech production in a reading task. *PLoS ONE* **11**, e0166872 (2016).
26. Mugler, E. M. et al. Direct classification of all American English phonemes using signals from functional speech motor cortex. *J. Neural Eng.* **11**, 035015 (2014).
27. Herff, C. et al. Brain-to-text: decoding spoken phrases from phone representations in the brain. *Front. Neurosci.* **9**, 217 (2015).
28. Moses, D. A., Mesgarani, N., Leonard, M. K. & Chang, E. F. Neural speech recognition: continuous phoneme decoding using spatiotemporal representations of human cortical activity. *J. Neural Eng.* **13**, 056004 (2016).
29. Pasley, B. N. et al. Reconstructing speech from human auditory cortex. *PLoS Biol.* **10**, e1001251 (2012).
30. Akbari, H., Khalighinejad, B., Herrero, J. L., Mehta, A. D. & Mesgarani, N. Towards reconstructing intelligible speech from the human auditory cortex. *Sci. Rep.* **9**, 874 (2019).
31. Martin, S. et al. Decoding spectrotemporal features of overt and covert speech from the human cortex. *Front. Neuroeng.* **7**, 14 (2014).
32. 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).
33. Wessberg, J. et al. Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. *Nature* **408**, 361–365 (2000).
34. Serruya, M. D., Hatsopoulos, N. G., Paninski, L., Fellows, M. R. & Donoghue, J. P. Instant neural control of a movement signal. *Nature* **416**, 141–142 (2002).
35. Taylor, D. M., Tillery, S. I. & Schwartz, A. B. Direct cortical control of 3D neuroprosthetic devices. *Science* **296**, 1829–1832 (2002).
36. Hochberg, L. R. et al. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature* **442**, 164–171 (2006).
37. Collinger, J. L. et al. High-performance neuroprosthetic control by an individual with tetraplegia. *Lancet* **381**, 557–564 (2013).
38. Aflalo, T. et al. Decoding motor imagery from the posterior parietal cortex of a tetraplegic human. *Science* **348**, 906–910 (2015).
39. Ajiboye, A. B. et al. Restoration of reaching and grasping movements through brain-controlled muscle stimulation in a person with tetraplegia: a proof-of-concept demonstration. *Lancet* **389**, 1821–1830 (2017).

Acknowledgements We thank M. Leonard, N. Fox and D. Moses for comments on the manuscript and B. Speidel for his help reconstructing MRI images. This work was supported by grants from the NIH (DP2 OD008627 and U01 NS098971-01). E.F.C. is a New York Stem Cell Foundation-Robertson Investigator. This research was also supported by The William K. Bowes Foundation, the Howard Hughes Medical Institute, The New York Stem Cell Foundation and The Shurl and Kay Curci Foundation.

Reviewer information *Nature* thanks David Poeppel and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions G.K.A., J.C. and E.F.C. conceived the study; G.K.A. inferred articulatory kinematics; G.K.A. and J.C. designed the decoder; J.C. performed decoder analyses; G.K.A., E.F.C. and J.C. collected data and prepared the manuscript; E.F.C. supervised the project.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-019-1119-1.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1119-1.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to E.F.C.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Participants and experimental task. Five participants (a 30-year-old female, 31-year-old female, 34-year-old male, 49-year-old female and 29-year-old female) underwent chronic implantation of a high-density, subdural electrode array over the lateral surface of the brain as part of their clinical treatment for epilepsy (Extended Data Fig. 3). Participants gave their written informed consent before the day of the surgery. All participants were fluent in English. All protocols were approved by the Committee on Human Research at UCSF and experiments and data in this study complied with all relevant ethical regulations. Each participant read and/or freely spoke a variety of sentences. Participant 1 read aloud two complete sets of 460 sentences from the MOCHA-TIMIT[20] database. Additionally, participant 1 also read aloud passages from the following stories: Sleeping Beauty, Frog Prince, Hare and the Tortoise, The Princess and the Pea, and Alice in Wonderland. Participant 2 read aloud one full set of 460 sentences from the MOCHA-TIMIT database and further read a subset of 50 sentences an additional nine times each. Participant 3 read 596 sentences that described 3 picture scenes and then freely described the scene, which resulted in another 254 sentences. Participant 3 also spoke 743 sentences during free-response interviews. Participant 4 read two complete sets of MOCHA-TIMIT sentences, 465 sentences related to scene descriptions and 399 sentences during free-response interviews. Participant 5 read one set of MOCHA-TIMIT sentences and 360 sentences related to scene descriptions. In addition to audible speech, participant 1 also read 10 sentences 12 times each alternating between audible and silently mimed (that is, making the necessary mouth movements) speech. Microphone recordings were obtained synchronously with the ECoG recordings.

Data acquisition and signal processing. Electrocorticography was recorded with a multi-channel amplifier optically connected to a digital signal processor (Tucker-Davis Technologies). Speech was amplified digitally and recorded with a microphone simultaneously with the cortical recordings. The grid placements were decided upon purely by clinical considerations. ECoG signals were recorded at a sampling rate of 3,052 Hz. Each channel was visually and quantitatively inspected for artefacts or excessive noise (typically 60 Hz line noise). The analytic amplitude of the high-gamma frequency component of the local field potentials (70–200 Hz) was extracted with the Hilbert transform and downsampled to 200 Hz. The low frequency component (1–30 Hz) was also extracted with a fifth order Butterworth bandpass filter, downsampled to 200 Hz and parallelly aligned with the high-gamma amplitude. Finally, the signals were z-scored relative to a 30-s window of running mean and standard deviation to normalize the data across different recording sessions. We studied the high-gamma amplitude, because it has been shown to correlate well with multi-unit firing rates and it has the temporal resolution to resolve fine articulatory movements[10]. We also included a low-frequency signal component owing to the decoding performance improvements noted for reconstructing perceived speech from auditory cortex[11,12]. Decoding models were constructed using all electrodes from vSMC, STG and IFG except for electrodes with bad signal quality as determined by visual inspection. We removed 8 electrodes for participant 1, 7 electrodes for participant 2 and 16 electrodes for participant 3. No electrodes were removed for participants 4 and 5. The decoder uses both high-gamma amplitude and raw low-frequency signals together as input to the model. For instance, n electrodes will result in $n \times 2$ input features.

Phonetic and phonological transcription. For the collected speech acoustic recordings, transcriptions were corrected manually at the word level so that the transcript reflected the vocalization that the participant actually produced. Given sentence level transcriptions and acoustic utterances chunked at the sentence level, hidden Markov model-based acoustic models were built for each participant so as to perform sub-phonetic alignment[40] within the Festvox[41] framework. Phonological context features were also generated from the phonetic labels, given their phonetic, syllabic and word contexts.

Cortical surface extraction and electrode visualization. We localized electrodes on each individual's brain by co-registering the preoperative T1 MRI with a postoperative computed tomography scan containing the electrode locations, using a normalized mutual information routine in SPM12. Pial surface reconstructions were created using Freesurfer. Final anatomical labelling and plotting was performed using the img_pipe Python package[42].

Inference of articulatory kinematics. One of the most accurate methods to record vocal tract kinematics is called electromagnetic midsagittal articulography (EMA). The process involves gluing small sensors to the articulators, generally three sensors on the tongue, one on each lip and one on each incisor. A magnetic field is projected at the participant's head and as the participant speaks, each sensor can be precisely tracked as it moves through the magnetic field. Each sensor has a wire leading out of the participant's mouth and connected to a receiver to record measurements.

Because of the above requirements, we did not pursue using EMA in the setting of our ECoG recordings, because the potential disruption of medical instruments by the magnetic field and long set-up time conflicted with limited recording session time with patients and the set-up procedure was too uncomfortable. Instead, we developed a model to infer articulatory kinematics from audio recordings. The articulatory data used to build the articulatory inference models was from MOCHA-TIMIT[20] and MNGU0 corpora[43].

The articulatory kinematics inference model comprises a stacked deep encoder–decoder, in which the encoder combines phonological (linguistic and contextual features, resulting from the phonetic segmentation process) and acoustic representations (25-dimensional MFCC vectors sampled at 200 Hz) into a latent articulatory representation (also sampled at 200 Hz) that is then decoded to reconstruct the original acoustic signal. The latent representation is initialized with inferred articulatory movement and appropriate manner features.

We performed statistical subject-independent acoustic-to-articulatory inversion[16] to estimate 12-dimensional articulatory kinematic trajectories (x and y displacements of tongue dorsum, tongue blade, tongue tip, jaw, upper lip and lower lip, as would be measured by EMA) using only the produced acoustics and phonetic transcriptions. Because EMA features do not describe all acoustically consequential movements of the vocal tract, we append complementary speech features that improve reconstruction of original speech. First, to approximate laryngeal function, we add pitch, voicing (binary value indicating if a frame is voiced or not) and speech envelope, that is, the frame level intensity computed as the sum total power within all the Mel scale frequencies within a 25-ms analysis window, computed at a shift of 5 ms. Next, we added place–manner tuples (represented as continuous 0–1 valued features) to bootstrap the EMA with what we determined were missing physiological aspects in EMA. There were 18 additional values to capture the following place–manner feature tuples (such as palatal approximant and labial stop; see Supplementary Information for the complete list). We used an existing annotated speech database (Wall Street Journal Corpus[44]) and trained speaker-independent deep recurrent network regression models to predict continuous valued place–manner vectors only from the acoustics features, the phonetic labels were used to determine the ground-truth values for these labels (for example, the dimension labial stop would be 1 for all frames of speech that belong to the phonemes /p/, /b/ and so forth). However, with a regression output layer, predicted values were not constrained to the binary nature of the input features. The network architecture was three feedforward layers followed by one bLSTM layer to predict each time point of these manner descriptors from a 100-ms window of acoustic features. Combined with the EMA trajectories, these 33 feature vectors form the initial articulatory feature estimates.

To ensure that the articulatory representation has the potential to reliably reconstruct speech for the target subject, we designed a stacked encoder–decoder network to optimize these initial estimates for these values. Specifically, a recurrent neural network encoder is trained to convert phonological and acoustic features to the articulatory representation and then a decoder that converts the articulatory representation back to the acoustic features (original MFCC). The encoder is implemented as two feedforward layers followed by two bLSTM layers. The decoder is implemented as three feedforward layers. Software implementation was done using Keras Functional API within Tensorflow[45]. The stacked network is retrained optimizing the joint mean-squared error loss on acoustic and EMA parameters using the ADAM optimizer, with an initial learning rate set at 0.001. For regularization 40% dropout was allowed in all feedforward layers. After convergence, the trained encoder is used to estimate the final articulatory kinematic features that act as the articulatory intermediate to decode acoustic features from ECoG.

Neural decoder. The decoder maps ECoG recordings to MFCCs through a two-stage process by learning intermediate mappings between ECoG recordings and articulatory kinematic features, and between articulatory kinematic features and acoustic features. All data (ECoG, kinematics and acoustics) are sampled and processed by the model at 200 Hz. We implemented this model using TensorFlow in Python. In the first stage, a stacked three-layer bLSTM[9] learns the mapping between 300-ms (60 time points) sequences of high-gamma and local frequency component signals and a corresponding single time point (sampled at 200 Hz) of the 33 articulatory features. In the second stage, an additional stacked three-layer bLSTM learns the mapping between the output of the first stage (decoded articulatory features) and 32 acoustic parameters (200 Hz) for sequences of full sentences. These parameters are 25-dimensional MFCCs, 5 sub-band voicing strengths for glottal excitation modelling, log(F_0) and voicing.

During testing, a full sentence sequence of neural activity (high-gamma and low-frequency components) is processed by the decoder. The first stage processes 300 ms of data at a time, sliding over the sequence sample by sample, until it has returned a sequence of kinematics that is equal in length to the neural data. The neural data are padded with an additional 150 ms of data before and after the sequence to ensure the result is the correct length. The second stage processes the entire sequence at once, returning an equal length sequence of acoustic features. These features are then synthesized into an audio signal.

At each stage, the model is trained using the ADAM optimizer to minimize mean-squared error. The optimizer was initialized with learning rate = 0.001, beta1 = 0.9, beta2 = 0.999, epsilon = 1e-8. Training of the models was stopped after the validation loss no longer decreased. Dropout rate is set to 50% in stage 1 and 25% in stage 2 to suppress overfitting tendencies of the models. There are 100 hidden units for each LSTM cell. Each model used three stacked bLSTMs with an additional linear layer for regression. We use a bLSTM because of the ability of this model to retain temporally distant dependencies when decoding a sequence[46].

In the first stage, the batch size for training was 256 and in the second stage the batch size was 25. Training and testing data were randomly split based on recording sessions, meaning that the test set was collected during separate recording sessions from the training set. The training and testing sets were split in terms of total speaking time in minutes:seconds (n = number of sentences in test set) as follows: participant 1: training, 92:15 and testing, 4:46 ($n = 101$); participant 2: training, 36:57 and testing, 3:50 ($n = 100$); participant 3: training, 107:42 and testing, 4:44 ($n = 98$); participant 4: training, 27:39 and testing, 3:12 ($n = 82$); participant 5: training, 44:31 and testing, 2:51 ($n = 44$).

For shuffling the data to test for significance, we shuffled the order of the electrodes that were fed into the decoder. This method of shuffling preserved the temporal structure of the neural activity.

The direct ECoG to acoustics decoder described in Fig. 2e has a similar architecture as the stage 1 articulatory bLSTM, except with an MFCC output. Originally we trained the direct acoustic decoder as a six-layer bLSTM that mimics the architecture of the stage 2 decoder with MFCCs as the intermediate layer and as the output. However, we found performance was better with a four-layer bLSTM (no intermediate layer) with 100 hidden units for each layer, 50% dropout and 0.005 learning rate using ADAM optimizer for minimizing mean-squared error. Models were coded using Tensorflow version 1.9 in Python.

Speech synthesis from acoustic features. We used an implementation of the Mel-log spectral approximation algorithm with mixed excitation[47] within Festvox to generate the speech waveforms from estimates of the acoustic features from the neural decoder.

Mel-cepstral distortion. To examine the quality of synthesized speech, we calculated the MCD of the synthesized speech when compared the original ground-truth audio. MCD is an objective measure of error determined from MFCCs and is correlated to subjective perceptual judgments of acoustic quality[21]. For each dimension d ($0 < d < 25$) of reference acoustic features mc^{y} of speaker y, and decoded features $\mathrm{mc}^{\hat{y}}$ we calculate MCD as follows:

$$\mathrm{MCD} = \frac{10}{\ln(10)}\sqrt{\sum_{0<d<25}(\mathrm{mc}_d^{y} - \mathrm{mc}_d^{\hat{y}})^2}$$

Intelligibility assessment. Listening tests using crowdsourcing are a standard way of evaluating the perceptual quality of synthetic speech[48]. To comprehensively assess the intelligibility of the neurally synthesized speech, we conducted a series of identification and transcription tasks on Amazon Mechanical Turk. The unseen test set from participant 1 (101 trials of 101 unique sentences, see Supplementary Information) was used as stimuli for listener judgments. For the word-level identification tasks, we created several cohorts of words grouped by the number of syllables within. Using the time boundaries from the ground-truth phonetic labelling, we extracted audio from the neurally synthesized speech into four classes of one-syllable, two-syllable, three-syllable and four-syllable words. We conducted tests on each of these groups of words that involved identification of the synthesized audio from a group of 10 choices, 25 choices or 50 choices of what they think the word is. The presented options included the true word and the remaining choices were randomly drawn from the other words within the class (see Supplementary Information for class sizes across these conditions). All words within the word groups were judged for intelligibility without any further subselection.

Because the content words in the MOCHA-TIMIT data are largely low-frequency words to assess sentence-level intelligibility, along with the neurally synthesized audio file, we presented the listeners to a pool of words that may be in the sentence. This task is a limited-vocabulary free-response transcription. We conducted two experiments in which the transcriber is presented with pool of 25 word choices or 50 word choices that may be used in the sentence (a sample interface is shown in the Supplementary Information). The true words that make up the sentence are included along with randomly drawn words from the entire test set and displayed in alphabetical order. Given that the median sentence is only seven words long (s.d. = 2.1, minimum = 4, maximum = 13), this task design allows for reliable assessment of intelligibility. Each trial was judged by 10–20 different listeners. Each intelligibility task was performed by 47–187 unique listeners (a total of 1,755 listeners across 16 intelligibility tasks, see Supplementary Information for breakdown per task) making all reported analyses statistically reliable. All sentences from the test set were sent for intelligibility assessment without any further selection. The listeners were required to be English speakers located in the United States, with good ratings (>98% rating from prior tasks on the platform). For the sentence transcription tasks, an automatic spell checker was used to correct misspellings. No further spam detection, or response rejection was done in all analyses reported. The WER metric computed on listener transcriptions is used to judge the intelligibility of the neurally synthesized speech. I is the number of word insertions, D is the number of word deletions and S is the number of word substitutions for a reference sentence with N words, WER is computed as

$$\mathrm{WER} = \frac{I + D + S}{N}$$

Data limitation analysis. To assess the amount of training data that affects decoder performance, we partitioned the data by recording blocks and trained a separate model for an allotted number of blocks. In total, eight models were trained, each with one of the following block allotments: 1, 2, 5, 10, 15, 20, 25 or 28. Each block comprised an average of 50 sentences recorded in one continuous session.

Quantification of silent speech synthesis. By definition, there was no acoustic signal to compare the decoded silent speech. In order to assess decoding performance, we evaluated decoded silent speech with regards to the audible speech of the same sentence uttered immediately before the silent trial. We did so by dynamically time-warping[49] the decoded silent speech MFCCs to the MFCCs of the audible condition and computing Pearson's correlation coefficient and MCD.

Phoneme acoustic similarity analysis. We compared the acoustic properties of decoded phonemes to the ground-truth to better understand the performance of our decoder. To do this, we sliced all time points for which a given phoneme was being uttered and used the corresponding time slices to estimate its distribution of spectral properties. Using principal components analyses, the 32 spectral features were projected onto the first four principal components before fitting the Gaussian kernel density estimate (KDE) model. This process was repeated so that each phoneme had two KDEs representing either its decoded and or ground-truth spectral properties. Using Kullback–Leibler divergence, we then compared each decoded phoneme KDE to every ground-truth phoneme KDE, creating an analogue to a confusion matrix used in discrete classification decoders. Kullback–Leibler divergence provides a metric of how similar two distributions are to one another by calculating how much information is lost when we approximate one distribution with another. Lastly, we used Ward's method for agglomerative hierarchical clustering to organize the phoneme similarity matrix.

To understand whether the clustering of the decoded phonemes was similar to the clustering of ground-truth phoneme pairs (Extended Data Fig. 7), we used the cophenetic correlation to assess how well the hierarchical clustering determined from decoded phonemes preserved the pairwise distance between original phonemes, and vice versa[24]. For the decoded phoneme dendrogram, the cophenetic correlation for preserving original phoneme distances was 0.71 compared to 0.80 for preserving decoded phoneme distances. For the original phoneme dendrogram, the cophenetic correlation for preserving decoded phoneme distances was 0.64 compared to 0.71 for preserving original phoneme distances. $P < 1 \times 10^{-10}$ for all correlations.

State–space kinematic trajectories. For state–space analysis of kinematic trajectories, a principal components analysis was performed on the 33 kinematic features using the training dataset from participant 1. Figure 4a, b shows kinematic trajectories (original, decoded (audible and mimed) projected onto the first two principal components. The example decoded mimed trajectory occurred faster in time by a factor of 1.15 than the audible trajectory, so we uniformly temporally stretched the trajectory for visualization. The peaks and troughs of the decoded mimed trajectories were similar to the audible speech trajectory ($r = 0.65$ and $r = 0.55$, principal components 1 and 2, respectively) although the temporal locations are shifted relative to one another, probably because the temporal evolution of a production, whether audible or mimed, is inconsistent across repeated productions. To quantify the decoding performance of mimed trajectories, we used the dynamic time-warping approach described above, although in this case, temporally warping with respect to the inferred kinematics (not the state–space) (Fig. 4e).

For analysis of state–space trajectories across participants (Fig. 4f), we measured the correlations of productions of the same sentence, but across participants. Since the sentences were produced at different speeds, we dynamically time-warped them to match and compared these against correlations of dynamically time-warped mismatched sentences.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The data that support the findings of this study are available from the corresponding author upon request.

Code availability

All code may be freely obtained for non-commercial use by contacting the corresponding author.

40. Prahallad, K., Black, A. W. & Mosur, R. Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis. *In Proc. 2006 IEEE International Conference on Acoustics Speech and Signal Processing* (ICASSP, 2006).
41. Anumanchipalli, G. K., Prahallad, K. & Black, A. W. *Festvox: tools for creation and analyses of large speech corpora*. http://www.festvox.org (2011).
42. Hamilton, L. S., Chang, D. L., Lee, M. B. & Chang, E. F. Semi-automated anatomical labeling and inter-subject warping of high-density intracranial recording electrodes in electrocorticography. *Front. Neuroinform*. **11**, 62 (2017).
43. Richmond, K., Hoole, P. & King, S. Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. In *Proc. Interspeech* **2011** 1505–1508 (2011).
44. Paul, B. D. & Baker, M. J. The design for the Wall Street Journal-based CSR corpus. In *Proc. Workshop on Speech and Natural Language* (Association for Computational Linguistics, 1992).
45. Abadi, M. et al. TensorFlow: large-scale machine learning on heterogeneous systems. http://www.tensorflow.org (2015).
46. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
47. Maia, R., Toda, T., Zen, H., Nankaku, Y. & Tokuda, K. An excitation model for HMM-based speech synthesis based on residual modeling. In *Proc. 6th ISCA Speech synthesis Workshop (SSW6)* 131–136 (2007).
48. Wolters, M. K., Isaac, K. B. & Renals, S. Evaluating speech synthesis intelligibility using Amazon Mechanical Turk. In *Proc. 7th ISCA Speech Synthesis Workshop (SSW7)* (2010).
49. Berndt, D. J. & Clifford, J. Using dynamic time warping to find patterns in time series. In *Proc. 10th ACM Knowledge Discovery and Data Mining (KDD) Workshop* 359–370 (1994).