



# Distributed feedforward and feedback cortical processing supports human speech production

Ran Wang<sup>a</sup>, Xupeng Chen<sup>a</sup>, Amirhossein Khalilian-Gourtani<sup>a</sup>, Leyao Yu<sup>b,c</sup> , Patricia Dugan<sup>b</sup>, Daniel Friedman<sup>b</sup>, Werner Doyle<sup>d</sup>, Orrin Devinsky<sup>b</sup>, Yao Wang<sup>a,c</sup> , and Adeen Flinker<sup>b,c,1</sup>

Edited by Terrence Sejnowski, Salk Institute for Biological Studies, San Diego, CA; received January 6, 2023; accepted July 22, 2023

Speech production is a complex human function requiring continuous feedforward commands together with reafferent feedback processing. These processes are carried out by distinct frontal and temporal cortical networks, but the degree and timing of their recruitment and dynamics remain poorly understood. We present a deep learning architecture that translates neural signals recorded directly from the cortex to an interpretable representational space that can reconstruct speech. We leverage learned decoding networks to disentangle feedforward vs. feedback processing. Unlike prevailing models, we find a mixed cortical architecture in which frontal and temporal networks each process both feedforward and feedback information in tandem. We elucidate the timing of feedforward and feedback-related processing by quantifying the derived receptive fields. Our approach provides evidence for a surprisingly mixed cortical architecture of speech circuitry together with decoding advances that have important implications for neural prosthetics.

speech production | auditory feedback | brain computer interface | electrocorticography | speech motor control

The central sulcus divides the human frontal from the posterior temporal, parietal, and occipital neocortices (1). Traditionally, this divide separates high-order planning and motor execution from sensation. Feedforward execution lies in the frontal cortices in contrast to feedback sensory processing across association cortices for the various modalities (e.g., auditory, visual, somatosensory, etc.) (2). Higher-order capacities such as working memory, cognitive control, and decision-making are often viewed as initiated by frontal cortices with direct influence on sensory cortices (3–5).

Human higher-order cognitive functions include planning and executing complex speech sequences that carry semantic and linguistic meaning (6, 7). Speech production is a complex human motor behavior requiring precise coordination of multiple oral, laryngeal, and respiratory muscles (8). These finely tuned motor actions then produce reafferent feedback in the auditory, tactile, and proprioceptive domains as we process our own speech.

Prevailing models in human speech motor control propose a feedforward system that predicts and generates actions and a feedback system responding to the vocal auditory and somatosensory effects (9–14). Across these models, there is a consensus that the two systems are anatomically separated, with the feedforward system mainly supported by ventral frontal cortices, while feedback processing is carried out across various sensory cortices (e.g., Heschl's and superior temporal gyri for auditory feedback). During speech production, both these systems are engaged in producing articulatory motor movements as well as perceiving the auditory (and somatosensory) feedback produced by our actions. A major challenge in human neuroimaging of speech circuitry is dissociating neural signatures that are due to the feedforward motor plan in contrast to feedback from auditory processing as well as elucidating the exact dynamics of feedforward and feedback engagement across peri-Sylvian cortex.

A growing literature has leveraged unique human electrocorticographic (ECoG) recordings from patients undergoing neurosurgical procedures to obtain a combined spatial and temporal resolution critical for investigating speech production. Studies have detailed the signatures of feedforward speech planning (15) and organization of execution (16–18) in frontal cortices as well as the subsequent auditory feedback architecture in temporal cortices (18–21). Recently, deep neural network approaches have been developed to decode speech represented in auditory (22–25) and sensorimotor (26, 27) cortices from ECoG recordings. Nevertheless, these approaches have not been able to disentangle feedforward and feedback contributions during speech production as the motor and sensory responses co-occur.

## Significance

Human speech production is a complex behavior that involves feedforward control of motor commands as well as feedback processing of self-produced speech. These processes require the engagement of multiple brain networks in tandem. However, it has been hard to dissociate the degree and timing of cortical recruitment for motor control versus sensory processing generated by speech production. We introduce a neural network framework that reconstructs produced speech from neural activity and leverages the learned decoding networks to disentangle these processes. Unlike prevailing models, we find a surprisingly mixed feedforward and feedback cortical recruitment during speech production. These results, together with our speech decoding advances, have important implications for speech motor control and neural prosthetics.

Author contributions: R.W., Y.W., and A.F. designed research; R.W., X.C., L.Y., P.D., D.F., W.D., O.D., and A.F. performed research; R.W., X.C., Y.W., and A.F. contributed new reagents/analytic tools; R.W., X.C., A.K.-G., L.Y., and A.F. analyzed data; P.D. and D.F. provided clinical care; W.D. provided neurosurgical clinical care; O.D. assisted in patient care and consent; and R.W., Y.W., and A.F. wrote the paper.

The authors declare no competing interest.

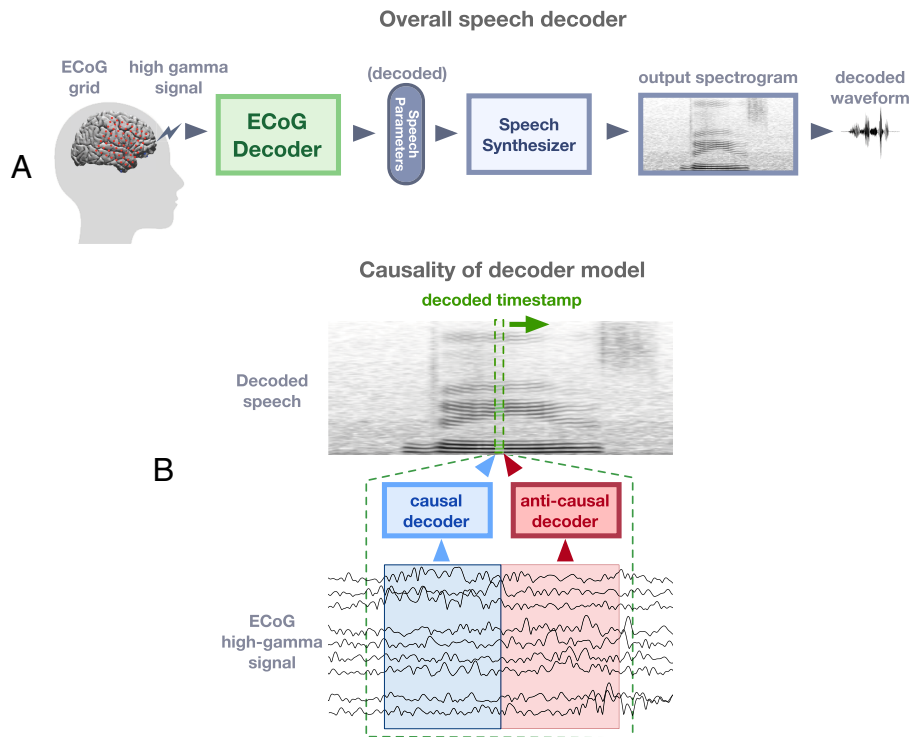
This article is a PNAS Direct Submission.

Copyright © 2023 the Author(s). Published by PNAS. This open access article is distributed under [Creative Commons Attribution License 4.0 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/).

<sup>1</sup>To whom correspondence may be addressed. Email: [adeen.flinker@nyulangone.org](mailto:adeen.flinker@nyulangone.org).

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2300255120/-/DCSupplemental>.

Published October 11, 2023.



**Fig. 1.** Speech decoding framework. (A) The overall structure of the decoding pipeline. ECoG amplitude signals are extracted in the high gamma range (i.e., 70–150 Hz). The ECoG Decoder translates neural signals from the electrode array to a set of speech parameters. This parameter representation is used to drive a speech synthesizer which creates a spectrogram (and associated waveform). (B) Illustration of causal versus anticausal neural decoding, which is leveraged to disentangle feedforward and feedback contributions. In order to reconstruct a speech timestamp, the causal model only uses feedforward neural signals in the past (blue), whereas the anticausal model only uses the neural signals that occur after the timestamp (red). The noncausal model is allowed to use both past and future neural signals to decode the current timestamp.

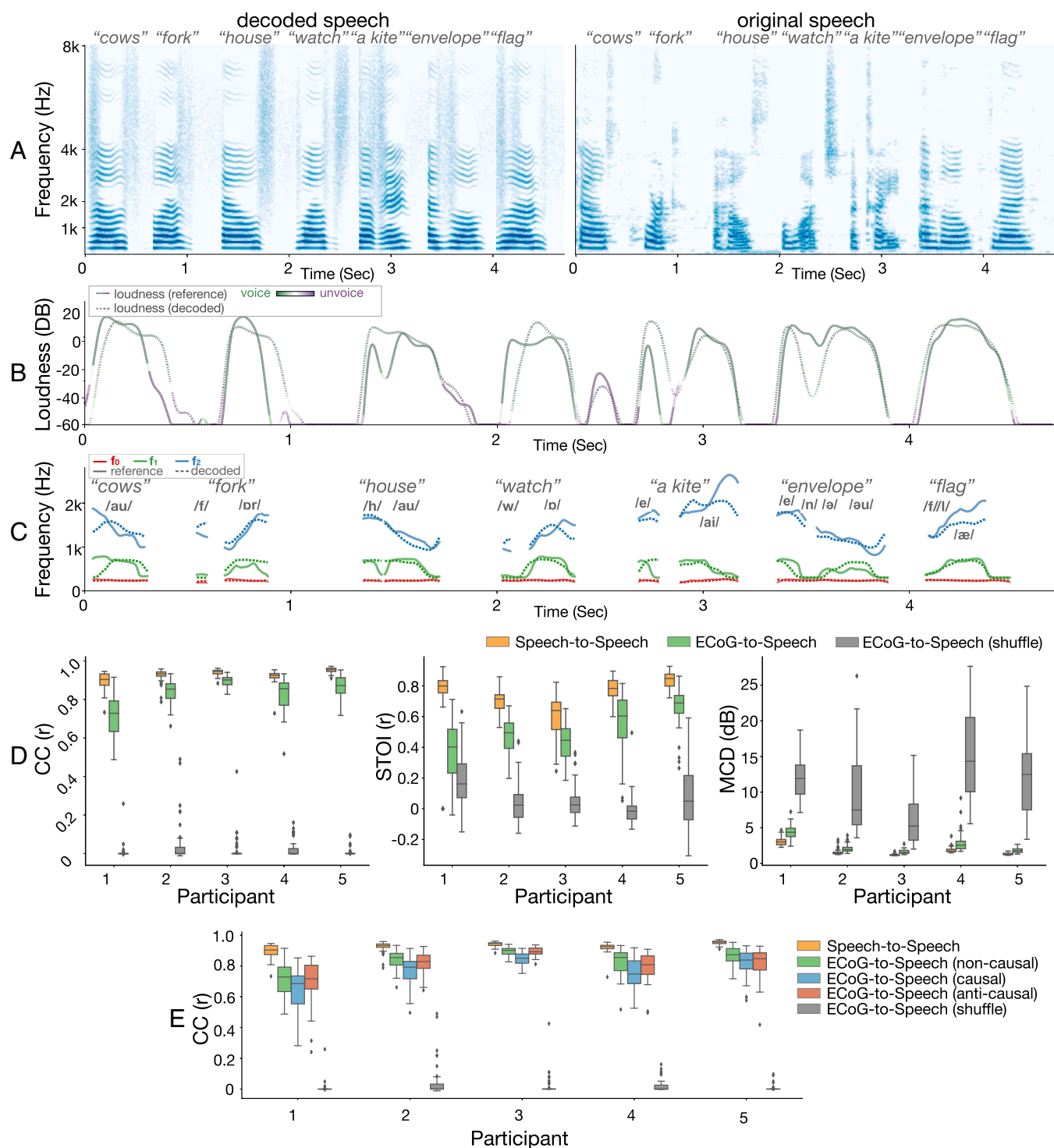
We directly disentangle feedback and feedforward processing during speech production by applying a deep learning architecture on human neurosurgical recordings to decode speech (Fig. 1A; *Materials and Methods, Visualizing Spatial Contribution Map*). Our approach decodes interpretable speech parameters from cortical signals, which drives a rule-based differentiable speech synthesizer. By learning neural network architectures which apply causal (predicting using only the past), anticausal (predicting using the future feedback), or both (noncausal), temporal convolutions (Fig. 1B), we are able to analyze the overall feedforward and feedback contributions, respectively, as well as to elucidate the temporal receptive fields of recruited cortical regions. This approach allows us to operationalize feedforward contributions related to the motor plan with our causal models. The feedback contributions related to hearing auditory feedback are operationalized with our anticausal models. In contrast to current views that separate feedback and feedforward cortical networks, our analyses reveal a surprisingly mixed architecture of feedback and feedforward processing both in frontal and temporal cortices while achieving superb speech decoding performance.

## Results

A major challenge in speech production research is dissociating neural signatures that are feedforward, or motor in nature, rather than auditory and somatosensory feedback produced by speech articulation. Given that these neural signals co-occur, we chose to disentangle their contributions by leveraging deep neural network architectures (i.e., Speech Decoding, Fig. 1A) that can decode

speech acoustics from the neural responses in different temporal directions (i.e., feedforward and feedback, Fig. 1B; *Materials and Methods, Speech Decoding Framework*). We leverage neurosurgical ECoG data obtained from five participants who took part in a battery of tasks: Auditory Repetition (AR), Auditory Naming (AN), Sentence Completion (SC), Word Reading (WR), and Picture Naming (PN). These were designed to elicit the same set of spoken words across tasks while varying the stimulus modality (28) and provided 50 repeated unique words (400–800 total trials per participant), all of which were analyzed locked to the onset of speech production.

**Robust Speech Decoding from Neural Signals.** We first demonstrate that our neural network approach produces accurate speech decoding with detailed acoustic features. The model's decoded spectrogram preserves the spectro-temporal structure of the original speech. It reconstructs both vowels and consonants (Fig. 2A) as well as the overall spectral energy distribution (*SI Appendix, Fig. S1*). These acoustic details result in a reconstruction that preserves the speakers' timbre (*Movies S1 and S2*) and leads to naturalistic voice decoding. Our model's speech parameters which include loudness, formant frequency, and the mixing parameter (i.e., the relative weighting between voiced and unvoiced components) are decoded accurately with the correct temporal alignment of each word onset and offset (Fig. 2B and C). The overall accuracy of the fundamental frequency (i.e., pitch), the first two modeled formants (i.e., F1 and F2), and the transition between voiced and unvoiced sounds are a major driving force for accurate speech decoding as well as naturalistic reconstruction that mimics the patient's voice.



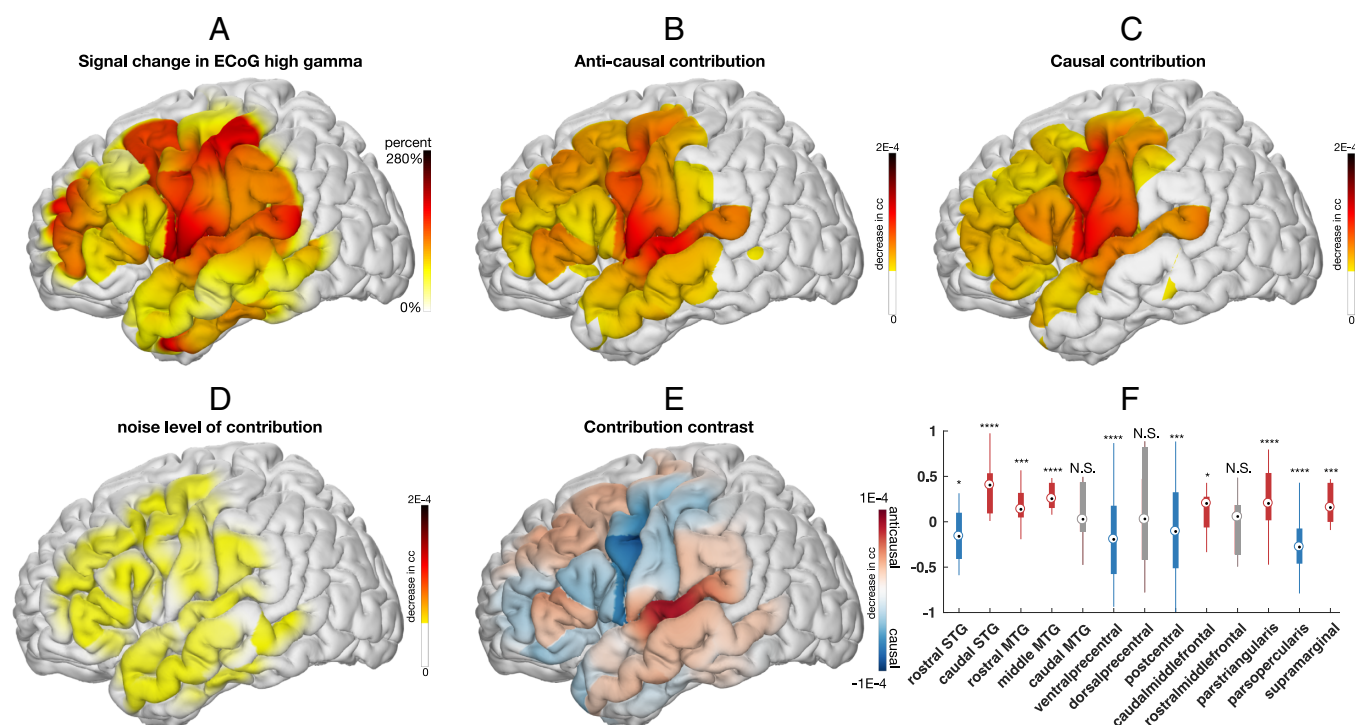
**Fig. 2.** Comparison of original and decoded speech produced by the model. (A) Spectrograms of decoded (Left) and original (Right) speech exemplar words. (B) Decoded loudness parameter with the voiced (mostly vowel) or unvoiced (mostly consonant) mixing parameter color-coded over the loudness curves. The same color spread and amplitude trend between decoded (dashed) and reference (solid) curves reflect accurate decoding of voiced and unvoiced phonemes with correct energy and temporal alignment. (C) Frequencies of the first two formants (F1, F2) and the pitch (F0). The matching between decoded (dashed) curves and reference (solid) curves in both frequencies during each phoneme and the overall temporal dynamic leads to intelligible and naturalistic decoding of voiced sounds. (D) Evaluation of the decoded speech quality in objective metrics. The correlation coefficient of spectrograms (CC, Left), short-time objective intelligibility (STOI, Middle), and Mel cepstral distortion (MCD, Right) is used for the evaluation. Note that lower MCD values represent better performance. Both the reconstructed speech from the speech autoencoder (yellow) and the speech decoded by the ECoG decoder (green) are reported. Additionally, the performance of a model trained on shuffled data (trained by matching the decoded spectrogram from the neural signal in a given duration to a randomly selected segment of spectrograms during the entire recording session) is also reported as a control. (E) Comparison of the CC metric among noncausal (green), causal (blue), and anticausal (red) models. Compared to the shuffled model (the same shuffled model as in D), the comparable performance across noncausal, causal, and anticausal models demonstrates sufficient information for decoding speech from both feedforward and feedback signals during speech production.

In order to evaluate the performance and quality of speech, we used several objective metrics, including the correlation coefficient (CC) between the decoded spectrogram and actually produced speech (23, 29, 30), an objective measure for speech intelligibility known as the Short-Time Objective Intelligibility (STOI) (23, 31), and a measure of spectral distortion, Mel-cepstral distortion (MCD) (26, 32). Across all participants and metrics, our neural decoding results performed well above chance (Fig. 2D in gray; estimated using shuffled data; *Materials and Methods, Visualizing Spatial Contribution Map*) and approached an upper bound of performance based on the unsupervised autoencoder (i.e., speech-to-speech) which did not use neural data. In order to verify that our model can generalize well to unseen words, we also performed stricter cross-validation providing a similar performance (*SI Appendix, Fig. S7*). The performance range across metrics and our participants were equal to and often better than the current literature (23, 26, 29, 30). Critically, all these models represent the noncausal case (Fig. 2D) that uses data both from the past (feedforward) and the future (feedback), as is currently a common practice (22, 23, 26, 29, 33) except a few nominal models (30).

In order to directly assess the performance of the causal (predicting using only the past) and anticausal (predicting using the future feedback) models and compare them with the noncausal (using past and future) model, which is standard in the field, we trained three separate models varying the temporal convolution direction. Our results (Fig. 2E) show a slight decrease in performance with the causal model. However, it performs close to the other models while providing a causal interpretation, which only uses past signals to predict future

speech. This is encouraging, as it suggests that, with additional improvement in the decoder design and training, it is possible to design practically applicable neuroprosthetic speech synthesizers. Also, comparable performance between causal, anticausal, and noncausal approaches indicates a similar amount of information contained by feedforward and feedback signals.

**Feedforward and Feedback Cortical Contributions to Speech Production Are Mixed.** To elucidate the feedforward and feedback contribution of different cortical regions to speech production, we examined the relative contribution of each electrode to decoding speech in our models. We derived the relative contribution by quantifying how the input signal at a particular electrode affects the overall accuracy (measured by the CC) of the reconstructed speech in the causal and anticausal models, respectively (*Materials and Methods, Visualizing Spatial–Temporal Contribution Receptive Fields*). This analysis isolates feedforward (i.e., causal) and feedback (i.e., anticausal) neural activity contributing to the decoding from the entangled ECoG neural signal, shown in Fig. 3A. In both the causal and anticausal models, peri-Sylvian electrodes were important for speech decoding; however, there was surprising recruitment of frontal regions when decoding speech based on the feedback (anticausal model, Fig. 3B) as well as recruitment of temporal sites when decoding speech based on the feedforward signals (causal model, Fig. 3C). We only show significant contributions that are above a threshold derived from the shuffled model (depicted in Fig. 3D). In order to quantify the prevalence of feedforward or feedback processing, we directly contrasted the two and projected the results onto the cortex (Fig. 3E and *SI Appendix, Fig. S5*). To



**Fig. 3.** (A) Averaged signal of input ECoG projected on the standardized MNI anatomical map. The colors reflect the percentage change of high gamma (–250 to 750 ms relative to speech onset) during production compared to the baseline level during the prestimulus baseline period. Note that (B) shows the anticausal contribution of different cortical locations (red indicates higher contribution), while (C) illustrates the causal contribution. (D) The noise level of the contribution analysis was evaluated by the contributions from the shuffled model. Contributions below noise level are not shown in (B) and (C). (E) The contrast is obtained by taking the difference between the anticausal and causal contribution maps (red means higher anticausal contribution, while blue means higher causal contribution). The boxplots (F) show the average difference in each cortical region (\*P-value < 0.05, \*\*P-value < 0.01, \*\*\*P-value < 0.001, \*\*\*\*P-value < 0.0001). For purposes of visualization, we normalize each electrode by the local grid density, since our ECoG grid has hybrid density. This removes the effect of nonuniform density on the projected results (A–E).

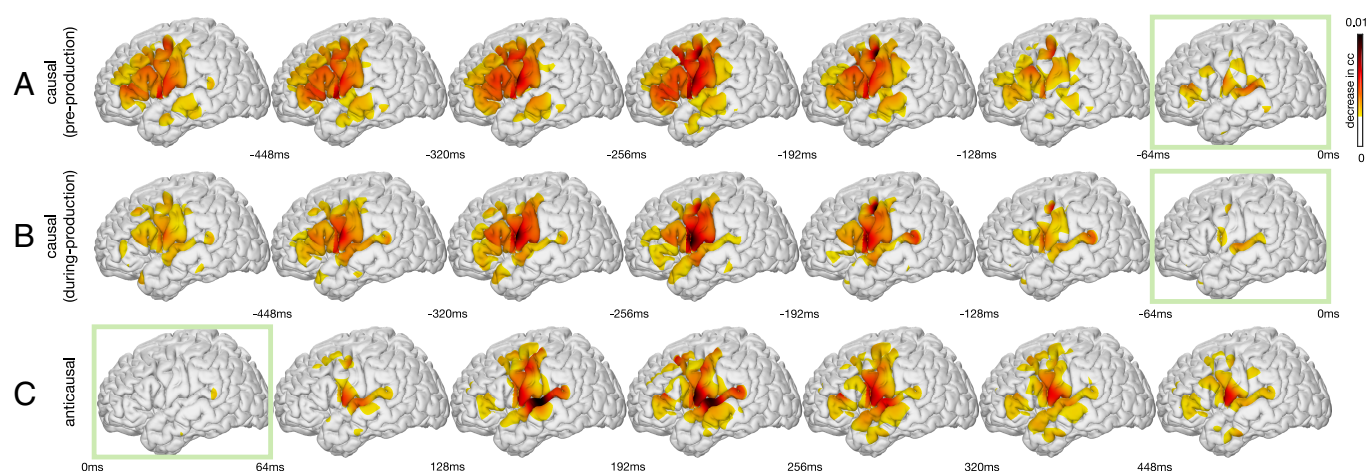
ascertain regions that contribute significantly more to feedback or feedforward processing, we conducted a region of interest analysis based on within-subject anatomical labels of each electrode (*Materials and Methods, Electrode Localization*), testing for an increase in causal or anticausal contributions across trials (nonparametric paired Wilcoxon test; Fig. 3, as well as ANOVA controlling for the subject as a random effect; *SI Appendix, Fig. S9*). We found a surprisingly mixed distribution of causal and anticausal contributions within both temporal and frontal cortices. A majority of the temporal cortex was predominantly anticausal, including the caudal superior temporal gyrus (STG; Wilcoxon sign rank,  $P = 1.607E - 15$ ,  $Z = 9.6234$ ) and portions of the middle temporal gyrus (MTG; rostral MTG: Wilcoxon sign rank test  $P = 2.5108E - 04$ ,  $Z = 4.9359$ , and middle MTG: Wilcoxon sign rank test  $P = 1.5257E - 13$ ,  $Z = 9.0185$ ) as well as supramarginal cortex (Wilcoxon sign rank test  $P = 1.1144E - 04$ ,  $Z = 5.3919$ ), implicating it in processing the auditory feedback signals for speech production. However, there was also a significant causal contribution in rostral STG (Wilcoxon sign rank test  $P = 0.0332$ ,  $Z = -2.9628$ ). Similarly, the majority of the sensorimotor cortex was predominantly causal, implicating it in processing the motor speech commands, including ventral precentral (Wilcoxon sign rank,  $P = 4.9511E - 08$ ,  $Z = -7.1409$ ) and postcentral gyri (Wilcoxon sign rank,  $P = 6.419E - 04$ ,  $Z = -4.9612$ ). However, the dorsal division of the precentral gyrus was equally causal and anticausal (Wilcoxon sign rank,  $P = 0.4349$ ,  $Z = 0.6525$ ), implicating it in processing both feedforward and feedback information equally. Within the inferior frontal cortex, we found a striking division of function wherein the pars opercularis was significantly causal (Wilcoxon sign rank test,  $P = 8.0693E - 15$ ,  $Z = -9.6185$ ) while the pars triangularis was significantly anticausal (Wilcoxon sign rank test,  $P = 2.6715E - 06$ ,  $Z = 6.3518$ ). Our results showed similar recruitment of the cortex when we considered the normalized causal to anticausal contribution of each individual electrode (*SI Appendix, Fig. S2*) as well as when the contribution analysis was weighted by the neural high gamma signal (*SI Appendix, Fig. S6*). Overall, these findings provide evidence for a mixed feedforward and feedback

processing of speech commands and their refference across temporal and frontal cortices, in contrast to a dichotomous view.

### Temporal Dynamics and Receptive Fields of Speech Production.

Speech production includes articulatory planning and executing the motor commands, processes that recruit distinct regions of the frontal cortex (15). However, their exact temporal receptive fields remain poorly understood. Earlier, we examined the causal and anticausal cortical contributions during speech articulation. Next, we examine articulatory planning and articulation of speech production stages and derive the related temporal receptive fields across the cortex. We leverage the receptive fields to test how cortical regions contribute differently to speech decoding with time and how frontal cortex dynamics change when feedback is introduced (after articulation starts). Both feedforward and feedback information are processed in tandem.

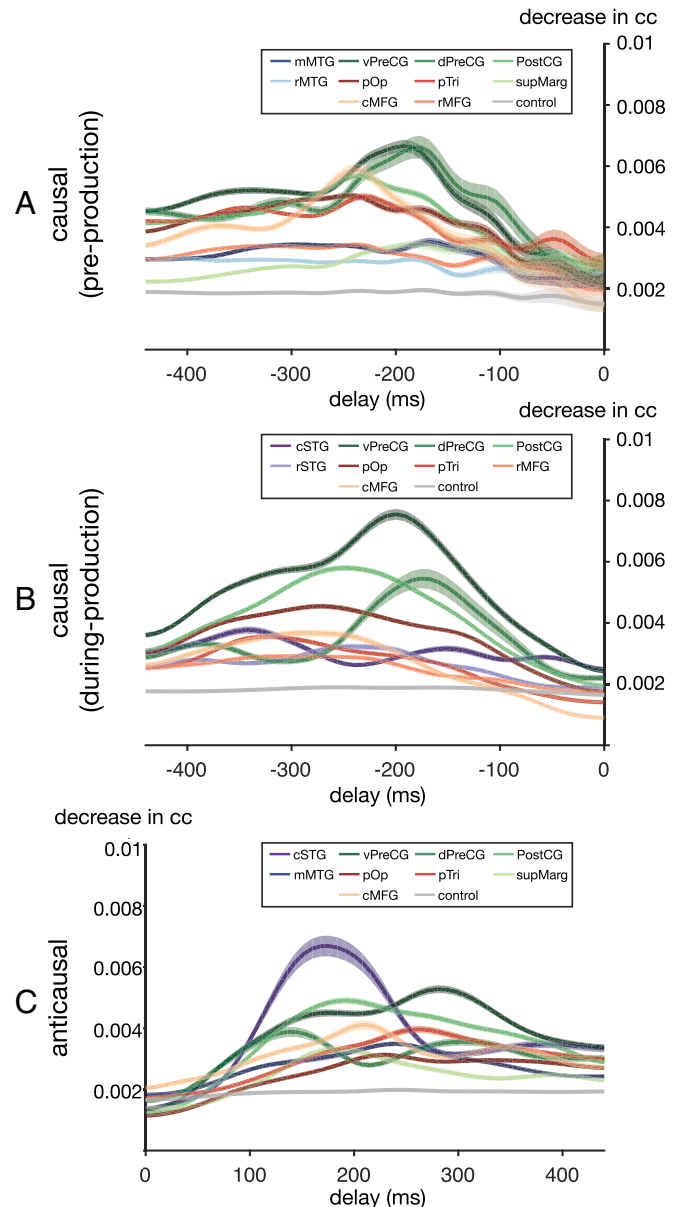
We employed a similar occlusion approach to derive the temporal receptive fields as in the previous section. However, we quantified how the input signal at a particular electrode affects the accuracy of the reconstructed speech across varying delays (*Materials and Methods, Visualizing Spatial–Temporal Contribution Receptive Fields*). In brief, we employ the same trained models (Figs. 2 and 3) but occlude neural signals at different time points and quantify the change in speech decoding. We always quantify speech-decoding changes during speech; however, the neural signal occluded could be preproduction (e.g., premotor) or during-production (e.g., motor). Importantly, we quantify these contributions across all relative delays between the neural signal and speech decoding (i.e., negative in the causal direction and positive in the anticausal direction). This approach allowed us to quantify the contribution of a specific electrode in the model as a function of delay relative to speech decoding, similar to classical temporal receptive fields (i.e., TRF). We conducted this analysis for both the trained causal and anticausal models applied to data during two epochs—one prior to production ( $-512 \text{ ms} \sim 0 \text{ ms}$ ; Fig. 4*A*) and the other during production, which included both causal and anticausal components ( $0 \text{ ms} \sim 512 \text{ ms}$ ; Fig. 4*B* and *C*). The projection of all the temporal receptive fields onto the cortex, which were significantly above a threshold derived from



**Fig. 4.** Spatial-temporal receptive fields based on decoding contribution. The contribution to speech decoding from cortical neural responses as a function of temporal delays. (A) The causal model is applied to preproduction data ( $-512 \text{ ms} \sim 0$ ) while occluding neural data and quantifying future speech decoding across multiple delays (between neural occlusion and speech decoding). These feedforward spatial-temporal receptive fields quantify the contribution of past neural signals, prior to production, to speech decoding (negative delays). (B) Similarly, the causal model is applied to during-production data ( $0 \sim 512 \text{ ms}$ ) representing feedforward spatial-temporal receptive fields that quantify the contribution of past neural signals (during speech production) to speech decoding (negative delays). (C) Representation of the feedback spatial-temporal receptive fields derived from the anticausal model that quantifies the contribution of future (positive delays) neural signals during feedback in the during-production data ( $0 \sim 512 \text{ ms}$ ). These feedback receptive fields quantify the contribution of future neural signals to (past) speech decoding (positive delays). Green boxes denote the temporal window closest to zero-lag. Contributions below significance ( $P < 0.05$ ) representing the noise level are clipped and not shown in the plots.

the shuffled model, is plotted in Fig. 4 as a function of delay. We found an increased frontal and MTG contribution prior to production (Fig. 4A) compared with during production (Fig. 4B). These processes are likely related to articulatory planning and lexical retrieval prior to speech production. During production, there was a prominent sharpening of ventral precentral gyrus receptive fields marked by a significant increase in contribution compared with preproduction (Wilcoxon sign rank test,  $P = 8.3979E - 05$ ,  $Z = 5.4203$ ). While a majority of prefrontal regions engaged prior to production, there was a significant decrease in contribution across pars triangularis (Wilcoxon sign rank test,  $P = 1.8493E - 32$ ,  $Z = -13.6074$ ) and middle frontal gyri (MFG; Wilcoxon sign rank test,  $P = 3.9177E - 09$ ,  $Z = -7.6103$  for caudal and  $P = 4.1581E - 04$ ,  $Z = -4.8311$  for rostral) except for pars opercularis (Wilcoxon sign rank test,  $P = 0.4819$ ,  $Z = 0.2066$ ). Similarly, to our previous results (Fig. 3 E and F), during production, we found a significant increase in anticausal contribution for caudal STG (Wilcoxon sign rank test,  $P = 2.6789E - 17$ ,  $Z = 9.6711$ ), pars triangularis (Wilcoxon sign rank test,  $P = 0.0162$ ,  $Z = 3.9003$ ), and caudal MFG (Wilcoxon sign rank test,  $P = 0.0045$ ,  $Z = 3.9862$ ) compared with causal contributions. This confirms the anatomical–functional division of the inferior and middle frontal gyri as well as caudal (Wilcoxon sign rank test,  $P = 2.6789E - 17$ ,  $Z = 9.6711$ ) and rostral separation of STG (Wilcoxon sign rank test,  $P = 0.0343$ ,  $Z = -2.9457$ ).

Next, we conducted a region of interest analysis, based on within-subject anatomical labels of each electrode, in order to derive the temporal receptive curves per region (Fig. 5). This approach provides critical insight as to the temporal tuning and peak recruitment of various regions to feedforward processing prior to (Fig. 5A) and during production (Fig. 5B) as well as feedback processing (Fig. 5C). We found a shift in receptive field tuning for the two subdivisions of the precentral gyrus. Prior to production, dorsal and ventral precentral gyri were not significantly different from each other (Wilcoxon sign rank test,  $P = 0.454$ ,  $Z = -0.36103$ ) and had close peak times ( $-196$  ms,  $-192$  ms prior to speech for ventral and dorsal precentral gyri, respectively). However, during production, these dynamics shifted and we found a significant decrease in dorsal precentral causal contribution (Wilcoxon sign rank test,  $P = 4.7575E - 05$ ,  $Z = -5.6272$ ) accompanied by temporal separation of peaks ( $-208$  ms,  $-184$  ms for ventral and dorsal precentral gyri, respectively; Fig. 5 A and B). Within the inferior frontal gyrus, we found that the pars opercularis was recruited similarly both prior to production and during production for feedforward processing (Wilcoxon sign rank test,  $P = 0.5922$ ,  $Z = 1.7462$ ) at a peak delay of  $-248$  ms and  $-280$  ms, respectively. During production, the pars triangularis had a selective increase in recruitment for anticausal compared with causal contributions (Wilcoxon sign rank test,  $P = 0.0162$ ,  $Z = 3.9003$ ), implicating it in increased feedback processing (Fig. 4C and SI Appendix, Tables S2 and S3). The anticausal receptive fields during production provide evidence for feedback processing most strongly contributed by caudal STG, with the earliest peak in contributions seen in the dorsal precentral gyrus (144 ms) and caudal STG (168 ms) followed by parietal (supramarginal 184 ms, postcentral 192 ms) and ventral precentral (280 ms) gyri (SI Appendix, Table S3). In order to ensure that these significant peaks were not within a temporal window possibly influenced by the autocorrelation structure of the speech or neural signal, we conducted a correlation analysis (SI Appendix, Fig. S11) and found that our results (Figs. 4 and 5) were outside the upper



**Fig. 5.** The temporal receptive field across anatomical regions. The contribution to decoding the current speech from cortical neural responses with certain temporal delays. (A) and (B) are the feedforward temporal receptive fields derived from the causal model by evaluating the contribution of past (negative delays) neural signals during a period before production onset (A) and after onset (B). (C) Representation of the feedback temporal receptive fields derived from the anticausal models that evaluate the contribution of future (positive delays) neural signals during feedback after articulation. The temporal propagation of the shuffled model estimates the noise level dynamics (gray curves in plots). Only regions significantly above noise level (Wilcoxon sign rank test on across-time averaged data,  $P < 0.05$ ) are reported.

bound of signal autocorrelations (i.e., all results were greater than the near-zero correlation at 136 ms for speech spectra and 48 ms for the neural signal). Taken together, our contribution analyses suggest preferential recruitment of prefrontal cortices in feedforward processing prior to production followed by a shift in dynamics during production when feedforward and feedback signals are jointly processed with anatomical divisions of labor.

## Discussion

Our study leverages a deep learning approach together with neurosurgical recordings and dissociates direct feedforward and feedback cortical contributions during speech production. Our neural network architecture achieves exceptional decoding of speech production by tapping an interpretable compact speech representation and can be altered to focus on causal, anticausal, and noncausal decoding. Our analyses of the cortical contributions driving the performance of these models reveal a mixed distribution of feedforward and feedback processing during speech production. This was prominent in inferior, middle frontal, and superior temporal gyri which exhibited an anatomical division between feedforward and feedback processing. Last, we show a change in the temporal dynamics of prefrontal recruitment during speech planning through production, characterized by a shift of inferior frontal and precentral gyri recruitment, processing both feedforward and feedback information at different time points and spatial locations.

A growing number of studies have leveraged deep neural networks for cortical speech decoding. Convolutional neural networks (CNN) (22–25) and recurrent neural networks (RNN) (26) have mapped ECoG signals into speech and text (33, 34). However, our approach diverges from these studies. First, we develop a differentiable speech synthesizer that can generate natural speech from a compact set of interpretable speech parameters based on several signal-processing equations. This rule-based synthesizer allows for unsupervised pretraining of meaningful encoded representations (reference speech parameters) as well as reduces the capacity of the entire model and increases training data efficiency. Our speech synthesizer provides a direct mapping from the speech parameters to a participant's voice, by using a set of participant-specific hyperparameters for the speech synthesizer that is obtained using unsupervised learning from the participant's speech. These hyperparameters include the parameters for the prototypes of the voiced and unvoiced filters. This eliminates the need for labeled articulatory data that maps speech to articulatory dynamics as proposed by Anumanchipalli et al. (26). Second, our compact speech representation leverages an interpretable decomposition of speech into voiced and unvoiced components. This decomposition is biologically necessary and has been reported in neural representations across frontal and temporal cortices (17, 35) and stands in contrast to other traditional speech synthesizing approaches (36, 37). Last, the speech neural decoding models to date mostly employ noncausal operations. Since such decoders require both past and future information for decoding, they are not applicable to real-time speech prosthetic applications. However, a select few studies have employed causal decoding (27). Importantly, mixed operations hinder disentangling feedforward and feedback cortical contributions. In addition to providing a causal model which directly translates to practical speech prosthetics, our approach can dissociate feedforward and feedback cortical contributions during speech production.

During speech production, we process feedforward and feedback signals in tandem. Previously, the two have not been disentangled. Attempts have focused on experimental manipulations which change the feedback by shifting frequency (38) or time (39). However, these manipulations change the cortical dynamics and introduce other cognitive processes due to hearing one's own voice altered as well as induced motor compensation. We applied convolution filters with different causality to directly train models to disentangle feedforward (i.e., causal models) and feedback (i.e., anticausal models) contributions of cortical regions. Feedforward

and feedback processes are critical for driving articulatory vocal tract movement. The feedforward pathway generates an initial articulatory command and predicts sensory (auditory and somatosensory) targets; the feedback pathway compares the targets with the perceived sensory feedback and updates subsequent feedforward commands. The exact mapping between anatomical regions and their contribution to specific functional roles differ across speech motor control models (10, 11). Further, these findings have been developed based mostly on noninvasive studies which have low temporal (e.g., fMRI) or spatial resolution (e.g., M/EEG). Our high spatiotemporal resolution ECoG data together with advanced deep neural networks provide a fine-grained mapping of the cortical feedforward and feedback speech networks.

Consistent with the predominant speech motor control models, our results showed a dominant feedforward process in the ventral motor and pars opercularis of the inferior frontal gyrus, while posterior superior temporal and supramarginal gyri in the parietal lobe showed feedback (6, 11, 40). However, in contrast to these models, we found that cortices in the frontal lobe, including the pars triangularis and caudal middle frontal, are predominantly feedback in nature, while rostral STG appears feedforward. This feedback processing across frontal cortices became even stronger when we limited our analyses to the speech production epoch (Fig. 4C and *SI Appendix, Table S3*). Additionally, most gyri (inferior frontal, caudal middle frontal, superior temporal, precentral, and postcentral cortices; *SI Appendix, Table S2*) had both feedforward and feedback contributions above the noise level derived from the shuffled model, suggesting that feedforward and feedback processing can mix in these regions.

Our findings are likely driven by sensory-motor signaling shared across the animal kingdom, referred to as a corollary discharge (41, 42). Such a discharge from motor cortices acts to inform sensory cortices of future self-generated reafferent stimulation, often in the form of cortical suppression (19). Our peak feedforward results in the ventral precentral gyrus are consistent with such a framework (11) as well as a recent report providing direct evidence for the source of a corollary discharge in human speech (43). Interestingly, our data suggest that feedforward processing might be more distributed than previously assumed given our significant contributions across the cortex (Fig. 3C) as well as large causal contributions in the anterior temporal cortex (Fig. 3E). Further, the stark asymmetry of feedforward and feedback contributions within adjacent regions of frontal cortex (e.g., pars opercularis and pars triangularis) suggests the possibility that cortical regions are processing feedforward corollary discharge signals while also updating representations based on feedback processing.

Our results highlight the anticausal feedback signature exhibited by sensorimotor and frontal cortices. While this goes against the canonical model of the frontal cortex in an action–perception loop (44), our findings complement a growing body of evidence showing specific responses in the frontal cortex to auditory stimuli during perception. Cheung et al. (45) found distinct auditory receptive fields as well as robust passive listening responses in the ventral precentral gyrus. Similarly, the dorsal division of the precentral gyrus has recently been implicated in processing auditory feedback of altered speech as well as responding robustly during passive listening (39). However, this begs the question as to why the speech motor cortex is processing auditory information. Our feedback contribution analysis suggests that auditory processing is specifically leveraged for anticausal processing of the reafferent signals during production. Indeed, our results show that the

dorsal precentral gyrus decreases feedforward processing while engaged in actual speech production (Fig. 5B) and is recruited for feedback at an early time point together with temporal cortices (Fig. 5C). Under this view, the auditory frontal responses seen during passive listening may constitute a representation dedicated to feedback processing when speech is produced.

To summarize, we provided an approach to decode speech production and interrogate the recalcitrant problem of mixed feedforward and feedback processing during speech production. We were able to leverage feedforward processing only in causal models to drive neural speech prosthetics [as opposed to the literature using noncausal processing (22, 23, 26, 29, 33)] as well as provide insights into the underpinning cortical drivers. Our results suggest a mixed cortical architecture in frontal and temporal cortices that dynamically shifts and processes both feedforward and feedback signals across the cortex. This is in contrast to previous views associating feedforward processing with primarily prefrontal and motor cortices while feedback processing is associated with the superior temporal cortex.

## Materials and Methods

### Experimental Model and Subject Details.

**Participants and experiments.** The neural data were obtained from five patients (three male) who were native English speakers, undergoing treatment for refractory Epilepsy at NYU Langone Hospital, with implanted electrocorticographic (ECoG) subdural electrode grids. All experimental procedures were approved by the NYU Grossman School of Medicine Institutional Review Board. Patients provided written and oral consent at least one week prior to surgery by a research team member after separate consultation with the clinical care provider. The subjects were instructed to complete five tasks to pronounce the target words in response to certain auditory or visual stimuli. The subjects were asked to freely respond after the stimuli were presented without any cue or artificial delay introduced. The five tasks were

- Auditory Repetition (AR, i.e., to repeat the auditory words),
- Auditory Naming (AN, i.e., name a word based on an auditory presented definition sentence),
- Sentence Completion (SC, i.e., complete the last word of an auditorily presented sentence),
- Visual Reading (VR, i.e., read aloud visually presented word in written form), and
- Picture Naming (PN, i.e., naming a word based on a visually presented color line drawing).

Each task contained the same 50 unique target words while varying stimulus modalities (auditory, visual, etc.). Each word appeared once in the AN and SC tasks and twice in the other tasks. For participants 1–3, the five tasks included 400 trials of the produced words and the corresponding ECoG recordings. The produced speech in each trial has an average duration of 500 ms. We repeated the same five tasks twice for participant 4 and collected data from 800 trials. For participant 5, we collected 800 trials by repeating the tasks twice, and we also ran an additional AR task (200 trials) which provided 1,000 trials in total.

**Data collection and preprocessing.** A microphone recorded the subject's speech during the tasks and was synchronized to the clinical Neuroworks Quantum Amplifier (Natus Biomedical, Appleton, WI), which records ECoG. The recordings sampled the peri-Sylvian cortex, including STG, IFG, precentral, and postcentral gyri. The implanted ECoG array contained total of 128 electrode channels, including standard 64 clinical  $8 \times 8$  macrocontacts (2 mm exposed, 10 mm spacing) as well as 64 additional interspersed smaller electrodes (1 mm exposed) between the macrocontacts (providing 10-mm center-to-center spacing between macrocontacts and 5-mm center-to-center spacing between micro/macro contacts, PMT corporation, Chanassen, MN). This FDA-approved array was manufactured for the study, and a member of the research team explained to patients that the additional contacts were for research purposes during consent. The ECoG arrays were implanted on the left hemisphere in all

participants' brains, and placement location was solely dictated by clinical care. More detailed illustration of the electrodes' coverage is shown in [SI Appendix, Fig. S3](#). We trained separate sets of decoding models for each participant. We randomly selected 50 out of all trials from the five tasks for testing and used the remaining data for training. The results reported are for testing data.

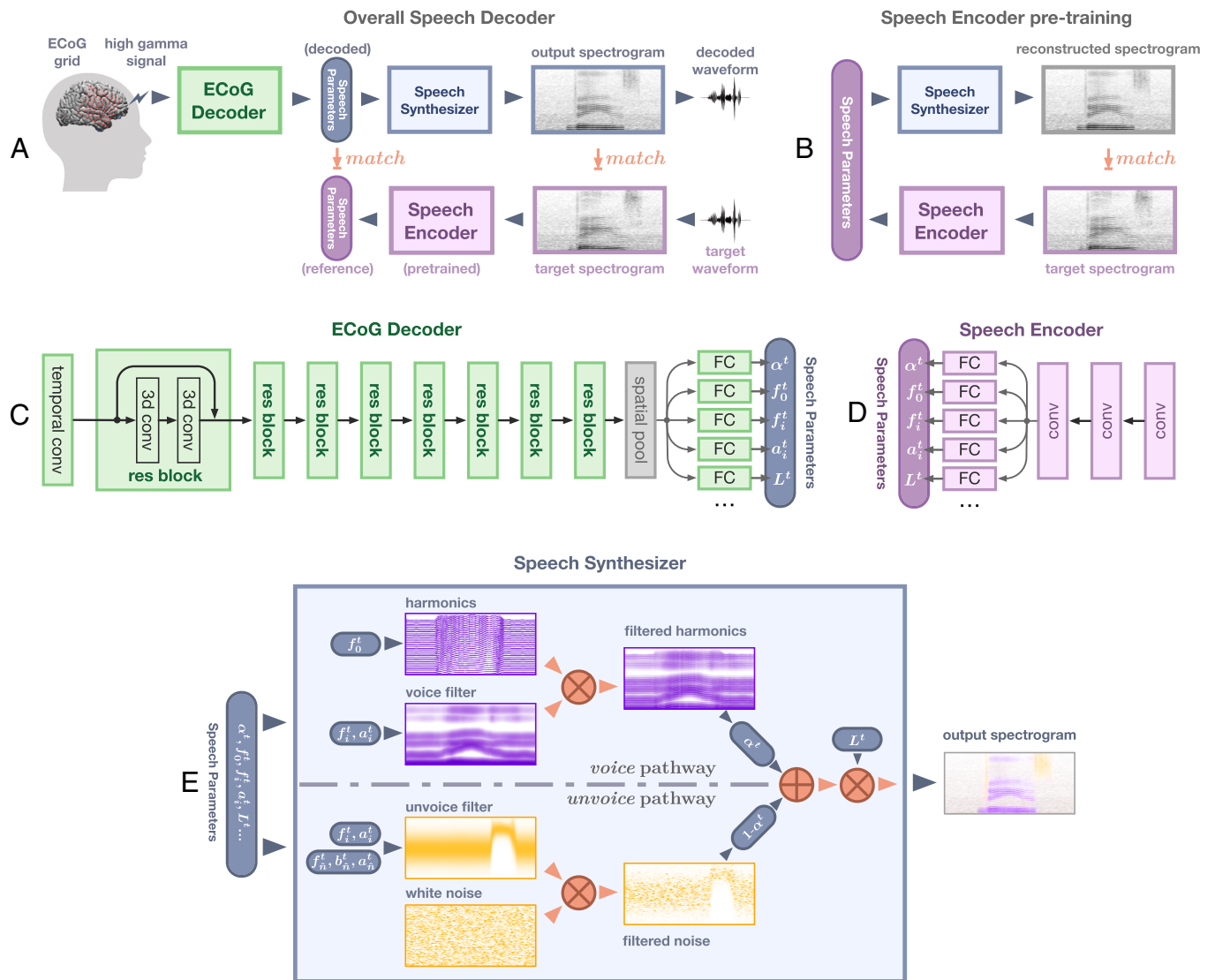
Each electrode sampled ECoG signals at 2,048 Hz, which was decimated to 512 Hz prior to processing. After rejecting electrodes with artifacts (i.e., line noise, poor contact with the cortex, and high amplitude shifts), we subtracted a common average reference (across all valid electrodes and time) from each individual electrode. Electrodes with interictal and epileptiform activity were removed from the analysis (note that a large number of temporal electrodes were removed from patients 4 and 5 for this reason). We then extracted the envelope of the high gamma (70–150 Hz) component from the raw signal with the Hilbert transform and further downsampled it to 125 Hz. The signal of each electrode over the silent baseline of 250 ms before the stimulus was used as the reference signal, and each electrode's signal was normalized to the reference mean and variance (i.e., z-score).

**Electrode localization.** Electrode localization in subject space, as well as MNI space, was based on coregistering a preoperative (no electrodes) and postoperative (with electrodes) structural MRI (in some cases, a postoperative CT was employed depending on clinical requirements) using a rigid-body transformation. Electrodes were then projected to the surface of the cortex (preoperative segmented surface) to correct for edema-induced shifts following previous procedures (46) (registration to MNI space was based on a nonlinear DARTEL algorithm). Based on the subject's preoperative MRI, the automated FreeSurfer segmentation (Destrieux) was used for labeling within-subject anatomical locations of electrodes. The anatomical region segments can be found in [SI Appendix, Fig. S8](#).

### Speech Decoder Details.

**Speech decoding framework.** The backbone of our decoding framework is constructed by an ECoG decoder and a speech synthesizer (Fig. 1A or 6A). During testing, the decoder generates a set of speech parameters from the high gamma components of the neural signal, that drive a differentiable speech synthesizer to generate speech spectrograms (and corresponding waveforms by the Griffin-Lim algorithm). All models are trained on data –250 to 750 ms relative to speech onset.

- ECoG decoder. The decoder maps the ECoG signals to a set of speech parameters (describing both the voiced and unvoiced components) which are then synthesized into speech spectrograms. The ECoG decoder architecture (Fig. 6C) is based on recent advances in convolutional neural networks leveraging the ResNet approach (47). We construct a modified ResNet model with nine layers that treat the cortical input as a spatiotemporal three-dimensional tensor (two dimensions for the electrode array and one for time). The decoder is trained such that its output parameters match the reference parameters derived from a speech encoder (which is learned separately in an unsupervised manner). Furthermore, our approach ensures that the speech spectrogram derived from these parameters and constructed by the speech synthesizer matches the actual speech spectrogram. We use this approach to be more data-efficient and allow us to train on a small set of samples for each patient.
- Speech parameters. Our speech representation is motivated by the vocoders used for low-bit-rate speech compression dating back to the 1980s. We model speech signals as a mixture of voiced and unvoiced components, with the voiced component described by a source-filter model (dynamically filtered harmonic signals) (37) and the unvoiced component generated by white noise broadband filtering. In addition to the mixing parameter, our representation includes speech formant information (frequency, bandwidth, etc.) and loudness (i.e., the energy of speech). Overall, speech parameters are sampled at 125 Hz, and each time step involves 18 parameters to model speech. Details are found in [SI Appendix, Additional Decoding Framework Details](#).
- Synthesizer. We use a set of signal processing equations (such as harmonic oscillation, noise generation, filtering, etc.) to synthesize the spectrogram



**Fig. 6.** Structure of the decoding framework. (A) The overall network architecture. The ECoG decoder is trained to match reference speech parameters and reconstructed by a pretrained speech encoder shown in panel (B). (B) The autoencoder is used to pretrain the speech encoder. The speech encoder is trained to generate proper speech parameters that can reconstruct input spectrograms through the speech synthesizer. (C) The ECoG decoder is a modified spatiotemporal residual network. After an initial temporal convolutional layer and eight residual blocks (constructed by three-dimensional convolution layers), multiple subnetworks (using one or two fully connected layers) generate speech parameters separately. (D) The speech encoder in (B) has three convolutional layers followed by the same multihead output structure as in (C). (E) Illustration of the processes within the speech synthesizer. The harmonics (in voice pathway) and white noise (unvoice pathway) are generated and filtered (multiplication in spectrogram domain) by voice and unvoiced filters, respectively. The filtered results are then weighted and averaged according to the mixing parameter and then amplified by the loudness parameter. Abbreviations, FC: fully connected layer, conv: convolutional layer, res block: residual block.

from our proposed speech parameters (Fig. 6E). We can train the ECoG decoder with a limited amount of training data by limiting the number of speech parameters and using differentiable signal processing equations. It is noteworthy that the equations we use are differentiable (*Differentiable Speech Synthesizer* in *SI Appendix, Additional Decoding Framework Details*), which allows for backpropagation from the spectrogram to the actual learning of the decoder.

- **Speech encoder.** The speech encoder (Fig. 6B) is pretrained using an independent unsupervised approach before the ECoG decoder training. The encoder is trained to generate a set of speech parameters from a given spectrogram, from which the aforementioned speech synthesizer can reproduce the spectrogram. This pretrained encoder generates reference speech parameters from actual speech signals used for the training of the ECoG decoder. The unsupervised process can be easily used to train the speech encoder from any set of speech signals, including patient-specific speech. Importantly, this process constrains the speech parameter space to

optimize the training of our ECoG decoder, and the parameters can directly drive a speech synthesizer based on differential equations.

We ruled out the possibility that our neural data suffer from a recently reported acoustic contamination (*SI Appendix, Fig. S4*) by following published approaches (48).

**Revealing delay-dependent decoding contributions on cortex.** To quantify electrode contributions, we use an occlusion-based approach where the entire electrode is occluded and we calculate the relative change of decoding correlation coefficient (i.e., Fig 3), or a small temporal segment of the electrode is occluded relative to speech production and we calculate the relative change of decoding correlation coefficient (i.e., Figs. 4 and 5). The temporal segment zeroed out relative to speech is varied and represents the delay in Figs. 4 and 5. Importantly, each delay represents all decoding changes during speech (at any point) and the neural signal relative to that delay (i.e., the neural signal could be prior to or during speech). This approach allows us to construct a neural receptive

field of contribution as a function of relative delay compared to speech (i.e., similar to spectro-temporal receptive fields in the auditory domain and motor receptive fields in the motor domain). Before formally defining the various contribution scores, we introduce the following notations:  $A_{\text{ref}}[s]$ : the ground truth spectrogram over a time duration  $S$  (i.e., window size) centered at time  $s$ , i.e., from  $s - S/2$  to  $s + S/2$ , derived by the speech-to-speech autoencoder.  $A_{\text{intact}}[s]$ : the model output with “intact” input (i.e., all ECoG signals are used).  $A_{\text{occlude}}^i[s; t]$ : the model output at time duration centered at  $s$  when the  $i$ th ECoG electrode signal in the time duration centered at  $t$  from  $t - \frac{T}{2}$  to  $t + \frac{T}{2}$  is occluded (i.e., set to zeros).  $S$  and  $T$  define the temporal refinement of the analysis and are independent of the receptive field of the ECoG decoder (i.e., delay from temporal convolutions).  $r(\cdot, \cdot)$ : correlation coefficient between two signals. We define the contribution of  $i$ th electrode in time duration centered at  $t$  to the output over duration centered at  $s$  by the reduction in the correlation coefficient between the output signal with the reference signal over the duration  $s$  when the  $i$ th electrode signal in duration  $t$  is occluded. Specifically,

$$C^i[s, t] = \text{Mean}\{r(A_{\text{ref}}[s], A_{\text{intact}}[s]) - r(A_{\text{ref}}[s], A_{\text{occlude}}^i[s; t])\},$$

where  $\text{Mean}\{\cdot\}$  denotes averaging across all testing samples.

To generate the contribution map, we first determine the contribution of each electrode (with a corresponding location in the MNI coordinate), which is then diffused into the surrounding area in the same anatomical region using a Gaussian kernel. Since our ECoG grid has hybrid density, to remove the effect of nonuniform density on the diffused result, we normalize the result of each region by the local grid density. The results shown in Figs. 3–5 are obtained by averaging the contribution maps obtained for all test samples for all participants (Figs. 4 and 5 do not include one patient due to coverage).

**Visualizing spatial contribution map.** The contribution of the entire signal at the  $i$ -th electrode to the entire output signal,  $C^i$ , is obtained by using the method in the section above (i.e., revealing delay-dependent decoding contributions on cortex), with  $S$  and  $T$  covering the entire input and output of the signal duration. The causal and anticausal contribution plots in Fig. 3 are generated by applying this analysis to the learned anticausal model (Fig. 3B) and causal model (Fig. 3C), respectively. The contrast of the anticausal and causal contribution (Fig. 3E) is the difference between the causal and anticausal contribution maps. The noise level for the contribution analysis (Fig. 3D) is generated from the shuffled model using noncausal processing (the shuffled model is trained on an artificial dataset with temporally misaligned input-output, and thus models of different causality are equivalent). To generate the box plots per cortical region (Fig. 3F), we calculate the contrast contribution averaged over electrodes of the same within-subject anatomical region.

The contrast of the anticausal and causal contributions (as shown in Fig. 3E) for each electrode  $i$  is defined as

$$C_{\text{contrast}}^i = C_{\text{anticausal}}^i - C_{\text{causal}}^i.$$

In order to examine the polarized electrode contributions (anticausal or causal), we calculate the normalized version of anticausal and causal contribution contrast:

$$C_{\text{polar}}^i = \frac{C_{\text{anticausal}}^i - C_{\text{causal}}^i}{C_{\text{anticausal}}^i + C_{\text{causal}}^i}.$$

By normalizing the contrast of anticausal and causal contribution,  $C_{\text{polar}}^i$  emphasizes the directionality of anticausal or causal, rather than their absolute contrast. This is visualized in *SI Appendix, Fig. S2 A and B* [only electrodes with either anticausal contribution attribute ( $C_{\text{anticausal}}^i$ ) or causal contribution ( $C_{\text{causal}}^i$ ) above the noise level determined by the shuffled model].

**Visualizing spatial-temporal contribution receptive fields.** When evaluating the contribution over a finite duration (temporal occlusion), we use small temporal scope  $S = T = 64$  ms (i.e., this represents the size of the neural signal occluded,  $T$ , as well as the size of speech decoding assessed,  $S$ ). To evaluate the contribution of an electrode signal to the output with various delays, denoted

by  $\tau$ , we average  $C^i[s, s + \tau]$  for all  $s$  in a certain duration (from time  $s_0$  to  $s_1$ ) leading to

$$\tilde{C}^i(\tau) = \frac{1}{s_1 - s_0} \sum_{s=s_0}^{s_1} C^i[s, s + \tau].$$

Here, we assume that the effect of the delay is independent of the actual time of speech  $s$ . Negative  $\tau$  values denote neural signals prior to a specific speech segment ( $s$ ) and thus represent the causal direction. In other words, when  $\tau \leq 0$ ,  $\tilde{C}_{\text{causal}}^i(\tau)$  reveals the causal contribution of electrode  $i$  to the speech output at the different delays (Fig. 4 A and B). Conversely, positive  $\tau$  values denote neural signals after a specific speech segment ( $s$ ) and thus represent the anticausal direction (i.e.,  $\tau \geq 0$ ,  $\tilde{C}_{\text{anticausal}}^i(\tau)$  reveals the anticausal contribution shown in Fig. 4C). Causal neural contributions to speech can originate prior to speech onset (prearticulation) or during speech production (during-articulation). To investigate the prearticulation contribution, we restrict  $s + \tau$  to the prearticulation period so that neural signals are only occluded before speech onset. Similarly, to investigate during-articulation contribution, we restrict  $s + \tau$  to postspeech onset. Importantly, speech decoding ( $s$ ) is always assessed during speech, while the neural signals are occluded either prearticulation or during-articulation time periods. This occlusion analysis can be performed prearticulation for the causal model only (past neural signals prior to speech onset contributing to future speech decoding) and during-articulation for both the causal model (past neural signals during speech production contributing to future speech decoding) and anticausal model (future neural signals during speech production contributing to past speech decoding) shown in Fig. 4.

**Visualizing per region temporal contribution receptive field.** Similarly to the per-region plot in Fig. 3F, we average the spatial-temporal receptive field data (Fig. 4) over within-subject anatomical region labels to generate a temporal contribution curve for each region (Fig. 5). The control curve is generated by applying the same method for the shuffled model (gray curves in Fig. 4). We omit curves that are not significantly above noise level by Wilcoxon sign rank testing between averaged (over time) region contribution curves and the averaged (over time) noise level curve (*SI Appendix, Table S2*). In order to calculate peak contributions, we smooth the curves using an 88-ms Hann window and take the maximal value. The motivation of smoothing the temporal curves was to provide unique estimates for peak contribution (i.e., maximum). The smoothed curves are plotted in Fig. 5, and the original unsmoothed, raw, data show the same dynamics and are plotted in *SI Appendix, Fig. S10*.

#### Additional Decoding Framework Details.

**Differentiable speech synthesizer.** In a traditional vocoder, speech is generated by switching between voiced and unvoiced content. Each content comes from an autoregressive system driven by a certain excitation signal that is either a harmonic signal or a white noise signal (49). Inspired by such a process, we construct our speech synthesizer shown in Fig. 6. It consists of two pathways. The voice pathway generates a voiced component by driving a harmonic excitation with time-varying fundamental frequency (i.e., pitch)  $f_0^t$  through a voiced filter consisting of  $N$  formant filters, each described by a center frequency  $f_i^t$  and an amplitude  $a_i^t$ ,  $i = 1, 2, \dots, N$ . Each formant filter can be viewed as an IIR filter defined explicitly in the frequency domain and is applied to the excitation signal (spectrogram) by a frequency-wise multiplication. Note that we parameterize the bandwidth  $b_i^t$  as a function of the center frequency  $f_i^t$ . The unvoice pathway generates an unvoiced component by driving a white noise through an unvoiced filter described as a center frequency  $f_n^t$ , bandwidth  $b_n^t$  and amplitude  $a_n^t$  (in addition to the  $N$  formant filters for the voice pathway). These two components are adaptively combined with a time-varying mixing factor  $\alpha^t$ , controlling the relative contribution between voiced sounds (for sonorant phonemes including vowels and nasals) and unvoiced sounds (for voiceless plosives and fricatives, such as /p/ and /s/, respectively). The voiced plosives and fricatives (such as /b/ and /z/, respectively) can be generated as a combination of voiced and unvoiced components. Finally, the combined signal is amplified by a loudness parameter  $L_t$ . In our study, we used  $N = 6$  formants. The synthesizer is driven by a total of 18 time-varying speech parameters, including the fundamental

(or pitch) frequency  $f_0^t$ , the mixing factor between the two pathways  $\alpha^t$ , the 12 parameters for the voiced filter ( $f_i^t, a_i^t$ ) and the three parameters for the unvoiced filter  $f_h^t, b_h^t, a_h^t$ , and the loudness  $L^t$ . Given the parameter values at each time sample, the synthesizer can generate a spectrogram sample. The spectrogram is a differentiable function of the speech parameters so that we can backpropagate the gradient of the training loss in terms of the predicted spectrogram to the speech parameters, which can then be backpropagated to either the speech encoder or the ECoG decoder parameters. Specifically, let the  $V^t(f)$  represent the spectrogram of the voicing component,  $U^t(f)$  that of the unvoicing component, and  $\alpha^t \in [0, 1]$  the mixing factor. The combined spectrogram can be written as  $S^t(f) = \alpha^t V^t(f) + (1 - \alpha^t) U^t(f)$ . Finally, the synthesized speech spectrogram is  $\tilde{S}^t(f) = L^t S^t(f)$ , where  $L^t$  is the loudness that modulates the signal cross-time.

**Formant filters in the voice pathway.** The filter in the voice pathway consists of multiple formant filters, corresponding to the multiple formants associated with vowels. The formant filter shape over frequency, which is related to the resonance property of the vocal tract, is closely related to the timbre of speakers' voice (50). We have found that a predefined analytic form such as generalized Gaussian cannot cover all feasible filter shapes. Instead, we learn a speaker-dependent prototype filter for each formant based on the speaker's natural speech. We represent the prototype filter  $G_i(f)$  for the  $i$ -th formant as a piecewise linear function, linearly interpolated from  $g^i[m]$ ,  $m = 1 \dots M$ , the amplitudes of the filter at  $M$  uniformly sampled frequencies up to  $f_{max}$ . We restrict the resulting filter  $G_i(f)$  to be unimodal (with a single peak of value 1) by properly constraining  $g[m]$ . Given  $g[m]$ ,  $m = 1 \dots M$ , the peak frequency  $f_i^{proto}$  and the half-power bandwidth  $b_i^{proto}$  can be determined. The actual formant filter at any time can be written as a shifted and scaled version of  $G_i(f)$ . Specifically, at time  $t$ , given an amplitude ( $a_i^t$ ), a center frequency ( $f_i^t$ ), and a bandwidth ( $b_i^t$ ), the  $i$ -th formant filter is given by

$$F_i^t(f) = a_i^t \cdot G_i \left( \frac{b_i^{proto}}{b_i^t} \cdot (f - f_i^t) + f_i^{proto} \right). \quad [1]$$

Then, the filter for the voice pathway with  $N$  formant filters can be written as  $F_h^t(f) = \sum_{i=1}^N F_i^t(f)$ . We learn the parameters  $g[m]$ ,  $m = 1 \dots M$  for  $G_i(f)$  during the unsupervised pretraining of the speech encoder, which does not require neural data. Fitting such a prototype filter is not data-hungry even with a relatively large  $M$ . We used  $M = 20$  in our experiment. Although two formants ( $N = 2$ ) have been shown to suffice for intelligible reconstruction (7), we use  $N = 6$  in our experiments for more accurate synthesis. We denote the parameter set for the voiced filter at time  $t$  by  $S^t = \{(f_i^t, a_i^t, b_i^t) | i \in \{1, \dots, N\}\}$ . The bandwidth  $b_i^t$  parameters are not independent speech parameters, rather functions of the center frequencies  $f_i^t$ .

**Unvoiced filter.** For the unvoice pathway, we add a broadband filter described by  $\{(f_h^t, a_h^t, b_h^t)\}$ . The shape of this filter  $F_h^t(f)$  follows Eq. 1 but with the filter coefficients  $(\alpha_i^t, f_i^t, b_i^t)$  replaced by  $(\alpha_h^t, f_h^t, b_h^t)$ . The bandwidth is constrained to satisfy  $b_h^t > 2,000$  Hz, following the broadband nature of obstruent phonemes. We also keep the multiple formant filters in the voiced filter described by  $S^t$ . This is motivated by the fact that human beings differentiate consonants with similar sounds such as /p/ and /d/, not only by the immediate burst of these sounds but also the development of the following formant frequency until the next vowel (51). To encode such formant transitions, we use the same formant filter parameters for modeling the narrow bandpass in both the voiced component and the unvoiced component. The parameter set for the unvoiced component is thus  $T^t = S^t \cup \{(f_h^t, a_h^t, b_h^t)\}$ . The overall filter for the unvoice pathway is  $F_h^t(f) = F_h^t(f) + \sum_{i=1}^N F_i^t(f)$ .

To further reduce the parameter space dimension, we model the bandwidth  $b_i^t$  of a formant filter as a piecewise linear function of the center frequency  $f_i^t$ . We assume

$$b_i^t = \begin{cases} a(f_i^t - f_\theta) + b_0, & \text{if } f_i^t > f_\theta \\ b_0, & \text{otherwise,} \end{cases}$$

where threshold frequency  $f_\theta$ , slope  $a$ , and baseline bandwidth  $b_0$  are three parameters that can be learned during unsupervised pretraining, shared among all formant filters.

**Harmonic excitation.** In the voice pathway, the voiced filter is applied on the harmonic excitation. This pathway models the human production of vowels and nasals, which results from the voice excited by the vocal cord shaped by the vocal tract. The excitation is constructed by sinusoidal harmonic oscillations with a time-varying fundamental frequency  $f_0^t$ . Inspired by the formulation in ref. 52, we define the harmonic excitation  $h^t$  as  $h^t = \sum_{k=1}^K h_k^t$ , where  $K$  is the total number of harmonics ( $K = 80$  in our experiment). Assuming the initial phase is 0, each harmonic resonance  $h_k^t$  at time step  $t$  has an instant phase that is the accumulation of resonance frequency in the past. Specifically, the  $k$ -th resonance at time step  $t$  is  $h_k^t = \sin(2\pi \sum_{\tau=0}^t f_k^\tau)$ , where  $f_k^{(t)} = kf_0^{(t)}$ . Denoting the spectrogram of  $h^t$  as  $H^t(f)$ , the spectrogram of the voiced component is the multiplication of  $H^t(f)$  and the voiced filter, i.e.,  $V^t(f) = H^t(f) F_h^t(f)$ .

**Noise excitation.** The unvoiced pathway models consonants like plosives and fricatives, where the vocal tract and human mouth filter the airflow through the mouth. It follows a similar process as in the harmonic counterpart. The major difference is that the excitation being filtered becomes stationary white Gaussian distributed noise  $\hat{n}(t) \sim \mathcal{N}(0, 1)$ , with a corresponding spectrogram  $N^t(f)$ . The filtered noise spectrogram (i.e., the unvoiced component) is  $U^t(f) = N^t(f) F_h^t(f)$ .

**ECoG decoder and speech encoder.** The ECoG decoder is constructed by a three-dimensional ResNet that treats time-varying signals on an ECoG grid array as spatiotemporal three-dimensional tensors (width  $\times$  height  $\times$  time duration). As is depicted in Fig. 6C, after an initial temporal convolutional layer [with 128 output features, each corresponding to a convolution kernel of size  $1 \times 1 \times 9$  (72 ms)], the signal passes through eight residual blocks. Each block contains two three-dimensional convolutional layers [with 128 output features, each corresponding to a convolution kernel of size  $3 \times 3 \times 5$  (40 ms)]. The output of the residual blocks creates a shared latent representation consisting of 128 features (each is a one-dimensional temporal signal by average pooling the two spatial dimensions), which is then fed into different output heads (each applies each consists of one or two fully connected layers acting on the 128 features at the same time point) to generate speech parameters. The overall temporal receptive field for generating one speech parameter sample is 73 temporal samples of 584 ms.

The speech encoder network architecture we choose is as simple as possible to demonstrate the effectiveness of the speech synthesizer design. We use three layers of temporal convolution (we treat the frequency axis of the spectrogram as the feature dimension) to generate a latent representation (Fig. 6D). Each convolutional layer has 128 output features and a temporal kernel size of three frames (24 ms). To output the speech parameter, we apply the same multihead structure to the latent representation as in the last layer of the ECoG decoder.

In order to implement models with different temporal causality, each convolutional layer of the speech encoder and ECoG decoder models were implemented as either causal, anticausal, or noncausal corresponding to the model causality.

**Loss and training hyperparameters.** The speech encoder is trained with a weighted average of the mixed spectral and parameter loss. The mixed spectral loss (52) is defined as

$$L_{MSS}(\tilde{S}^t(f), S^t(f)) = L_{lin}(\tilde{S}^t(f), S^t(f)) + L_{mel}(\tilde{S}^t(f), S^t(f)),$$

in which,

$$L_{lin}(x, y) = \|x - y\|_1 + \|\log x - \log y\|_1$$

$$L_{mel}(x, y) = \|x_{mel} - y_{mel}\|_1 + \|\log x_{mel} - \log y_{mel}\|_1,$$

where  $S^t(f)$  and  $\tilde{S}^t(f)$  denote the ground truth and reconstructed spectrograms, respectively, subscript *lin* means that the frequency is in the linear scale, while the subscript *mel* means the frequency is in the mel scale. In our experiments, we use 256 frequency samples (ranging from 0 to 8,000 Hz) for both linear and mel scale speech spectrograms.

Let us denote the  $j$ -th reconstructed speech parameter as  $\tilde{P}_j^t$  and its reference  $P_j^t$ ; the overall training loss for the ECoG decoder becomes

$$L = L_{\text{spectrogram}} + L_{\text{speechparameters}} \\ = \lambda_0 L_{\text{MSS}}(\tilde{S}^t(f), S^t(f)) + \sum_j \lambda_j \left( \left\| \tilde{P}_j^t - P_j^t \right\|_2^2 \right),$$

where  $\lambda_j$  balance the contribution from different loss terms since they have different physical meanings and scales.

Both the speech encoder and ECoG decoder are fitted by the Adam optimizer with hyperparameters:  $lr = 10^{-3}$ ,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ . We train an individual ECoG decoder and speech encoder per patient. The pretraining of the speech encoder and the training of the ECoG decoder share the same training/testing set partition.

**Quantification and Statistical Analysis.** We perform most of our statistical tests with Wilcoxon sign rank. The one-way ANOVA test is used to confirm the causal and anticausal contributions per-region with the subject treated

as a random variable. Prior to the ANOVA test, we first verified that causal and anticausal follow a normal distribution (using a Kolmogorov–Smirnov test). Statistical significance is indicated as \* $P$ -val  $< 0.05$ , \*\* $P$ -val  $< 10^{-2}$ , \*\*\* $P$ -val  $< 10^{-3}$ .

**Data, Materials, and Software Availability.** Some study data available (the code and data reported in this paper will be shared by the lead contact upon request, due to the sensitivity of patient data and ability to anonymize all data. Code to generate the results can be found at <https://github.com/flinkerlab/DistributedFeedforwardFeedbackProcessing>) (53).

**ACKNOWLEDGMENTS.** We would like to thank Robert Knight and Sasha Devore for providing helpful comments. This work was supported by the NSF under Grant No. IIS-1912286 (Y.W. and A.F.) and NIH R01NS109367, R01NS115929, R01DC018805 (A.F.).

Author affiliations: <sup>a</sup>Electrical and Computer Engineering Department, New York University, New York, NY 11201; <sup>b</sup>Neurology Department, New York University, New York, NY 10016; <sup>c</sup>Biomedical Engineering Department, New York University, New York, NY 11201; and <sup>d</sup>Neurosurgery Department, New York University, New York, NY 10016

1. E. R. Kandel *et al.*, *Principles of Neural Science* (McGraw-Hill, New York, NY, 2000), vol. 4.
2. J. M. Fuster, The prefrontal cortex—An update: Time is of the essence. *Neuron* **30**, 319–333 (2001).
3. B. T. Miller, M. D'Esposito, Searching for “the top” in top-down control. *Neuron* **48**, 535–538 (2005).
4. J. M. Fuster, The prefrontal cortex in the neurology clinic. *Handb. Clin. Neurol.* **163**, 3–15 (2019).
5. D. T. Stuss, R. T. Knight, *Principles of Frontal Lobe Function* (Oxford University Press, 2013).
6. G. Hickok, D. Poeppel, The cortical organization of speech processing. *Nat. Rev. Neurosci.* **8**, 393 (2007).
7. E. F. Chang, K. P. Raygor, M. S. Berger, Contemporary model of language organization: An overview for neurosurgeons. *J. Neurosurg.* **122**, 250–261 (2015).
8. K. Simonyan, H. Ackermann, E. F. Chang, J. D. Greenlee, New developments in understanding the complexity of human speech production. *J. Neurosci.* **36**, 11440–11448 (2016).
9. F. H. Guenther, A neural network model of speech acquisition and motor equivalent speech production. *Biol. Cybernet.* **72**, 43–53 (1994).
10. F. H. Guenther, *Neural Control of Speech* (MIT Press, 2016).
11. J. H. Houde, S. S. Nagarajan, Speech production as state feedback control. *Front. Hum. Neurosci.* **5**, 82 (2011).
12. G. Hickok, Computational neuroanatomy of speech production. *Nat. Rev. Neurosci.* **13**, 135–145 (2012).
13. G. Hickok, The cortical organization of speech processing: Feedback control and predictive coding the context of a dual-stream model. *J. Commun. Disorders* **45**, 393–402 (2012).
14. G. Hickok, The architecture of speech production and the role of the phoneme in speech processing. *Lang. Cognit. Neurosci.* **29**, 2–20 (2014).
15. A. Flinker *et al.*, Redefining the role of Broca's area in speech. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 2871–2875 (2015).
16. K. E. Bouchard, N. Mesgarani, K. Johnson, E. F. Chang, Functional organization of human sensorimotor cortex for speech articulation. *Nature* **495**, 327–332 (2013).
17. J. Chartier, G. K. Anumanchipalli, K. Johnson, E. F. Chang, Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex. *Neuron* **98**, 1042–1054 (2018).
18. X. Pei *et al.*, Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition. *Neuroimage* **54**, 2960–2972 (2011).
19. A. Flinker *et al.*, Single-trial speech suppression of auditory cortex activity in humans. *J. Neurosci.* **30**, 16643–16650 (2010).
20. J. D. Greenlee *et al.*, Human auditory cortical activation during self-vocalization. *PLoS ONE* **6**, e174744 (2011).
21. J. D. Greenlee *et al.*, Sensory-motor interactions for vocal pitch monitoring in non-primary human auditory cortex. *PLoS ONE* **8**, e60783 (2013).
22. H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, N. Mesgarani, Towards reconstructing intelligible speech from the human auditory cortex. *Sci. Rep.* **9**, 874 (2019).
23. M. Angrick *et al.*, Speech synthesis from ECoG using densely connected 3D convolutional neural networks. *J. Neural Eng.* **16**, 036019 (2019).
24. R. Wang, Y. Wang, A. Flinker, “Reconstructing speech stimuli from human auditory cortex activity using a WaveNet approach” in *2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)* (IEEE, 2018), pp. 1–6.
25. R. Wang *et al.*, “Stimulus speech decoding from human cortex with generative adversarial network transfer learning in 2020” in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* (2020), pp. 390–394.
26. G. K. Anumanchipalli, J. Chartier, E. F. Chang, Speech synthesis from neural decoding of spoken sentences. *Nature* **568**, 493–498 (2019).
27. M. Angrick *et al.*, Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity. *Commun. Biol.* **4**, 1–10 (2021).
28. J. Shum *et al.*, Neural correlates of sign language production revealed by electrocorticography. *Neurology* **95**, e2880–e2889 (2020).
29. M. Angrick *et al.*, Interpretation of convolutional neural networks for speech spectrogram regression from intracranial recordings. *Neurocomputing* **342**, 145–151 (2019).
30. C. Herff *et al.*, Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices. *Front. Neurosci.* **13**, 1267 (2019).
31. C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech” in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, 2010), pp. 4214–4217.
32. J. Kominek, T. Schultz, A. W. Black, “Synthesizer voice quality of new languages calibrated with mean MEL cepstral distortion” in *Spoken Languages Technologies for Under-Resourced Languages* (2008).
33. J. G. Makin, D. A. Moses, E. F. Chang, Machine translation of cortical activity to text with an encoder-decoder framework. *Nat. Neurosci.* **23**, 575–582 (2020).
34. D. A. Moses *et al.*, Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New Engl. J. Med.* **385**, 217–227 (2021).
35. C. Humphries, M. Sabri, K. Lewis, E. Liebenthal, Hierarchical organization of speech perception in human auditory cortex. *Front. Neurosci.* **8**, 406 (2014).
36. M. Fleischer, S. Pinkert, W. Mattheus, A. Mainka, D. Mürbe, Formant frequencies and bandwidths of the vocal tract transfer function are affected by the mechanical impedance of the vocal tract wall. *Biomech. Model. Mechanobiol.* **14**, 719–733 (2015).
37. J. L. Flanagan, *Speech Analysis Synthesis and Perception* (Springer Science & Business Media, 2013), vol. 3.
38. E. F. Chang, C. A. Niziolek, R. T. Knight, S. S. Nagarajan, J. F. Houde, Human cortical sensorimotor network underlying feedback control of vocal pitch. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 2653–2658 (2013).
39. M. Ozker, W. Doyle, O. Devinsky, A. Flinker, A cortical network processes auditory error signals during human speech production to maintain fluency. *PLoS Biol.* **20**, e3001493 (2022).
40. J. P. Rauschecker, S. K. Scott, Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nat. Neurosci.* **12**, 718 (2009).
41. E. von Holst, H. Mittelstaedt, Das Reafferenzprinzip. *Naturwissenschaften* **37**, 464–476 (1950).
42. T. B. Crapse, M. A. Sommer, Corollary discharge across the animal kingdom. *Nat. Rev. Neurosci.* **9**, 587–600 (2008).
43. A. Khalilian-Gourtani *et al.*, A corollary discharge circuit in human speech. *bioRxiv* [Preprint] (2022). <https://doi.org/10.1101/2022.09.12.507590> (Accessed 28 September 2023).
44. J. M. Fuster, Upper processing stages of the perception-action cycle. *Trends Cognit. Sci.* **8**, 143–145 (2004).
45. C. Cheung, L. S. Hamilton, K. Johnson, E. F. Chang, The auditory representation of speech sounds in human motor cortex. *eLife* **5**, e12577 (2016).
46. A. I. Yang *et al.*, Localization of dense intracranial electrode arrays using magnetic resonance imaging. *Neuroimage* **63**, 157–165 (2012).
47. K. He, X. Zhang, S. Ren, J. Sun, “Deep residual learning for image recognition” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778.
48. P. Roussel *et al.*, Observation and assessment of acoustic contamination of electrophysiological brain signals during speech production and sound perception. *J. Neural Eng.* **17**, 056028 (2020).
49. L. Deng, D. O'Shaughnessy, *Speech Processing: A Dynamic and Optimization-Oriented Approach* (CRC Press, 2018).
50. E. J. Hunter, J. G. Svec, I. R. Titze, Comparison of the produced and perceived voice range profiles in untrained and trained classical singers. *J. Voice* **20**, 513–526 (2006).
51. J. Jiang, M. Chen, A. Alwan, On the perception of voicing in syllable-initial plosives in noise. *J. Acoust. Soc. Am.* **119**, 1092–1105 (2006).
52. J. Engel, L. Hantrakul, C. Gu, A. Roberts, DDSP: Differentiable digital signal processing. *arXiv [Preprint]* (2020). <http://arxiv.org/abs/2001.04643> (Accessed 7 December 2021).
53. R. Wang, A. Flinker, DistributedFeedforwardFeedbackProcessing Repository. GitHub. <https://github.com/flinkerlab/DistributedFeedforwardFeedbackProcessing>. Deposited 28 September 2023.