Supplementary materials

Appendix 1:

All results reported in this longitudinal study were based on data collected from the first participant enrolled in the CortiCom Clinical Trial. This participant is a right-handed male, 61 years old at the time of implantation, who was diagnosed with amyotrophic lateral sclerosis (ALS) 8 years earlier. He presented with progressive bulbar dysfunction characterized by mixed spasmodic-flaccid dysarthria, severe dysphagia, and respiratory decline. Before device removal in January 2025, his residual phonatory capacity supported slow but intelligible volitional speech, as indicated by an ALSFRS-R speech domain score of 1/4. Additionally, the participant exhibited progressive weakness in the upper extremity muscles and limited fine motor function of the hands, significantly affecting his ability to perform daily tasks independently. Lower extremity muscle strength remained relatively preserved, though occasional gait instability due to impaired arm swing was noted.

Before enrollment, the participant completed a standard cognitive function assessment that showed no signs of dementia or major cognitive impairment. Monthly cognitive safety screenings during the study consistently found no evidence of significant cognitive decline. Regular oral motor function assessments indicated stable jaw strength and range of motion throughout the study period. The participant underwent cortical implantation surgery in July 2022.

Methods 1:

Signal Acquisition and Audio Synchronization

Raw ECoG signals (sampled at 30kHz) acquired in this study were initially filtered (0.3 Hz-7500 Hz), amplified, and digitized via a NeuroPlex-E headstage (Blackrock Neurotech, Salt Lake City, UT), connected to the implanted subdural electrodes through a transcutaneous 128-channel NeuroPort pedestal (Blackrock Neurotech, Salt Lake City, UT). The same subdural reference wire was used for all recordings, and referential derivations were used for all signal analyses. ECoG signals were transmitted from the Neuroplex-E headstage via an HDMI cable and a digital hub to the NeuroPort Biopotential Signal Processing (NSP) System (Blackrock Neurotech, Salt Lake City, UT), where the digital ECoG signals were downsampled to 1kHz.

Simultaneously, we recorded the participant's speech during the syllable-repetition task using a microphone (BETA® 58A, SHURE, Niles, IL) placed in front of him, with electronic noise isolated. These audio signals were amplified and digitized via an external audio interface (H6 Audio Recorder, Zoom Corporation, Tokyo, Japan) before being routed concurrently to two destinations: (1) one analog input channel of the NSP system, enabling synchronized recording with the neural data at 1kHz, and (2) a Zoom recorder for capturing high-fidelity audio at 48kHz. We then used cross-correlation to align the high-fidelity audio recordings with the synchronized audio signals recorded with the NSP system.

Methods 2:

Tuning fork experiments

To alleviate concerns about potential acoustic artifacts during overt speech attempts, we refer to the methodology of the tuning fork control experiments previously reported by Wilson et al.[1,2] Our results showed that placing a tuning fork at 128 Hz (close to the fundamental frequency of the participant's speech ($F_0 \approx 130$ Hz)) near the participant or in contact with the skull did not result in a measurable energy increase in ECoG signals near that frequency. This result was consistent with previous tests.[2,3]

Methods 3:

tVSA formula derivation

To assess the significance of changes in tVSA, it was necessary to model it statistically. This was done by deriving a model under the same assumptions and methodology as a previously studied model of quadrilateral vowel space area (qVSA)[4]. The determinant form was used to compute the area of the triangle formed by the $(F_1^{\mathrm{a}}, F_2^{\mathrm{a}})$, $(F_1^{\mathrm{u}}, F_2^{\mathrm{u}})$, and $(F_1^{\mathrm{i}}, F_2^{\mathrm{i}})$ points or tVSA (equation 1).

$$A = \frac{1}{2}\left|\underbrace{[F_1^{\mathrm{a}} - F_1^{\mathrm{u}}]}_{T_1} \cdot \underbrace{[F_2^{\mathrm{i}} - F_2^{\mathrm{a}}]}_{T_2} - \underbrace{[F_1^{\mathrm{a}} - F_1^{\mathrm{i}}]}_{T_3} \cdot \underbrace{[F_2^{\mathrm{u}} - F_2^{\mathrm{a}}]}_{T_4}\right| \quad (1)$$

Assuming normality for the formant random variables $F_{1,2}^{\mathrm{x}}$, it becomes possible to define the also normally distributed random variables T_1, T_2, T_3, T_4, defined in equation 1. Assuming independence relationships $T_1 \perp T_2$, $T_3 \perp T_4$, $T_1T_2 \perp T_3T_4$ enables the modelling of tVSA. Denoting $T_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $T_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, as well as $Z = T_1T_2$, the moment generating function $M_Z(s)$ of Z is given by[4],

$$M_Z(s) = \frac{1}{\sqrt{1 - \sigma_1^2\sigma_2^2 s^2}} \exp\left\{\frac{\mu_1\mu_2 s + \mu_1^2\sigma_2^2 s^2/2 + \mu_2^2\sigma_1^2 s^2/2}{1 - \sigma_1^2\sigma_2^2 s^2}\right\} \quad (2)$$

Using the result from equation 2, the moment generating function for tVSA can be determined to be:

$$M_{\mathrm{tVSA}} = M_{T_1T_2}\left(\frac{s}{2}\right) M_{T_3T_4}\left(\frac{-s}{2}\right). \tag{3}$$

The moment generating function described in equation 3 was subsequently derived to obtain the parameters necessary for evaluating significance, *i.e.,* the mean and variance of tVSA:

$$\mu_{\mathrm{tVSA}} = \frac{\mathrm{d}}{\mathrm{d}s}\bigg|_{s=0} M_{\mathrm{tVSA}}(s), \qquad \sigma_{\mathrm{tVSA}^2} = \frac{\mathrm{d}^2}{\mathrm{d}s^2}\bigg|_{s=0} M_{\mathrm{tVSA}}(s) - \mu_{\mathrm{tVSA}^2}. \tag{4}$$

Methods 4:

Modified EEGNet Architecture and Training Protocol

The architecture of this model processes data through a series of distinct modules. First, the input ECoG data is reshaped into the (Batch, 1, Channels, Time) format and fed into the temporal convolution module. This module includes a two-dimensional convolutional layer with a kernel size of (1, 65) (acting as a temporal filter) and a batch normalization layer to stabilize the learning process.

Next, the data flows into the deep spatial convolution module. This module is designed to efficiently learn spatial filters across electrode channels on each temporal feature map. It uses deep 2D convolutions with a kernel size of (n_channels, 1), where n_channels is the number of input electrode channels. The output then undergoes batch normalization, an ELU activation function, an average pooling layer for downsampling, and a dropout layer for regularization.

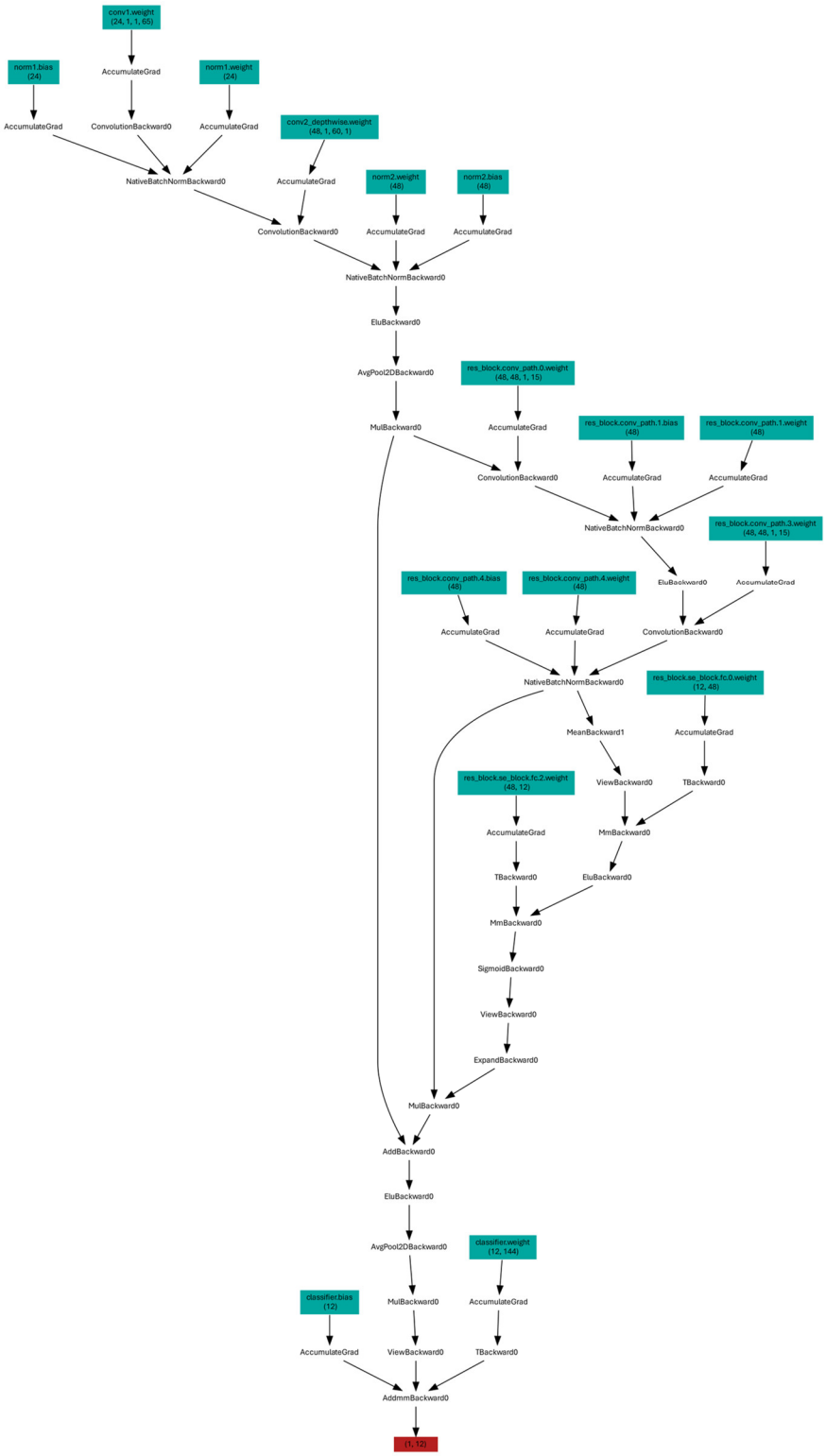
The core of the network is the SE-Residual Block. Feature maps are processed

through a residual path containing two consecutive 2D convolutional layers, both using (1, 15) convolutional kernels followed by batch normalization and ELU activation. Along this path, the SE-Block is applied to the feature map. This block performs a “squeeze” operation via global average pooling to create a channel descriptor, and an “excitation” operation that computes channel weights using two fully connected layers. These weights are multiplied by the feature map to selectively amplify salient information. Finally, a shortcut connection adds the block's original input to the attention-weighted output, and the result is passed to the final ELU activation.

After the residual block, the model proceeds to the final classification stage. Feature maps are downsampled by a final average pooling layer and regularized with dropout. They are then flattened into a vector and passed into a fully connected (linear) layer, which produces classifications for twelve syllable categories.

Implementation and training were conducted with rigorous data selection and hyperparameter optimization. The model input consisted of the HG band-power extracted from 0 to 3 seconds following the speech cue. We applied baseline normalization to each trial by subtracting the mean and dividing by the standard deviation of the pre-cue baseline period (-0.5 to 0 seconds). The final dataset used 60 cortical channels for model training and testing, excluding those on the upper grid and several channels identified as exhibiting anomalous signal quality. We identified the optimal model configuration through a systematic Bayesian hyperparameter optimization of 50 iterations, searching for the ideal learning rate, weight decay, batch size, number of filters, and dropout rate.[5] We trained the final model with the optimal hyperparameter combination identified through the search. During training, we employed the AdamW optimizer[6] and the Cosine annealing learning

rate scheduling.[7] To improve learning on complex samples, we used Focal Loss as the loss function (α=0.25, γ=2.0).[8]



Methods 5:

Electrode Saliency Analysis

To assess the spatial distribution of cortical contributions to decoding performance, we estimated electrode-level saliency based on perturbation sensitivity.[9,10] Saliency is defined as the sensitivity of model predictions to input perturbations.[11] Specifically, for each electrode channel, when there is a small neural signal perturbation in our input, the method quantifies the relative change in the model output probability distribution caused by the perturbation, thereby assessing the extent to which the electrode contributes to the current classification result.[11] We calculated the L1 norm of the predicted category probability gradient relative to the input activity for each electrode. This gradient was temporally aggregated throughout the trial to generate a saliency score for each electrode per trial. The final electrode contribution score was obtained by averaging all trials within each of the 20 consecutive sessions. Then, for the early-trained model, we had six groups of sessions, and for the late-trained model, we had four groups in total. For visualization, preoperative volumetric MRI and postoperative CT scans were co-aligned using FreeSurfer to locate electrode grids on the cortical surface.[12]

Methods 6:

Neural Signal Analysis

To further understand the ECoG signal variations for each syllable, we divided the experimental data into subsets, each consisting of 10 consecutive sessions corresponding to 50 trials per syllable. We computed the average HG response Z-score for the trials within each subset and assessed the degree of stabilization by measuring the variability of the

active-period HG response waveforms between each subset and its subsequent subset using the root-mean-square error (RMSE),

$$RMSE = \sqrt{\frac{\sum(x_i - y_i)^2}{n}}. \quad (5)$$

x_i and y_i denote the corresponding sampling points of the active HG Z-score waveforms between successive subsets, respectively, and n is the number of sampling points.

Supplementary Figures

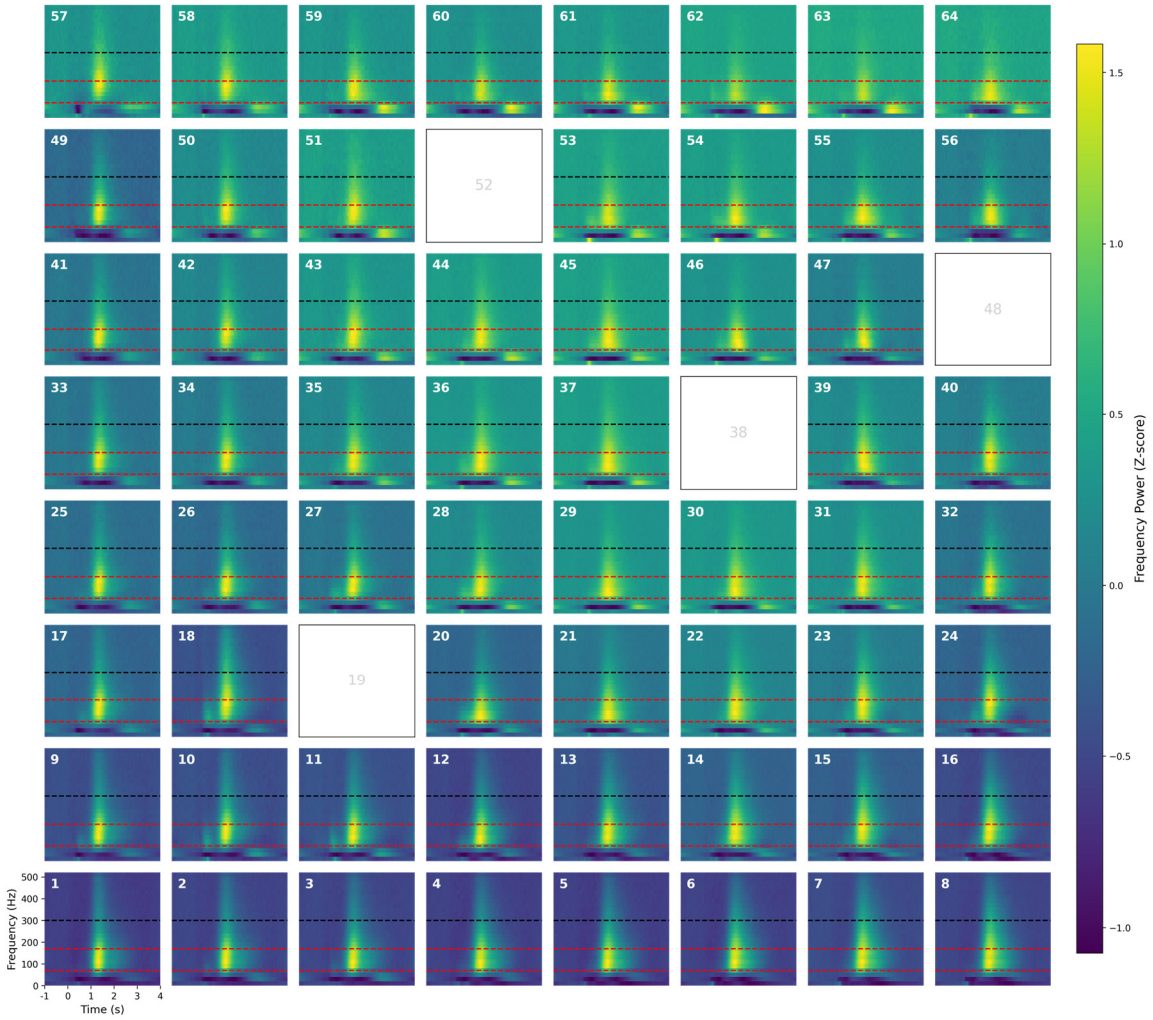


Figure 1. Trial-averaged spectrograms across a 64-channel chronic ECoG array. Each panel shows the trial-averaged spectrogram for a single electrode over two years. Time zero marks the onset of the auditory cue. Color intensity indicates the Z-scored power change relative to a -0.5 to 0s baseline period, with yellow signifying a power increase. Horizontal dashed lines mark the frequency bands used for subsequent analysis: the high-gamma band (70–170 Hz, red) for neural decoding and a high-frequency noise band (300–499 Hz, black) for white noise. White panels indicate channels excluded from analysis due to high impedance or poor signal quality.

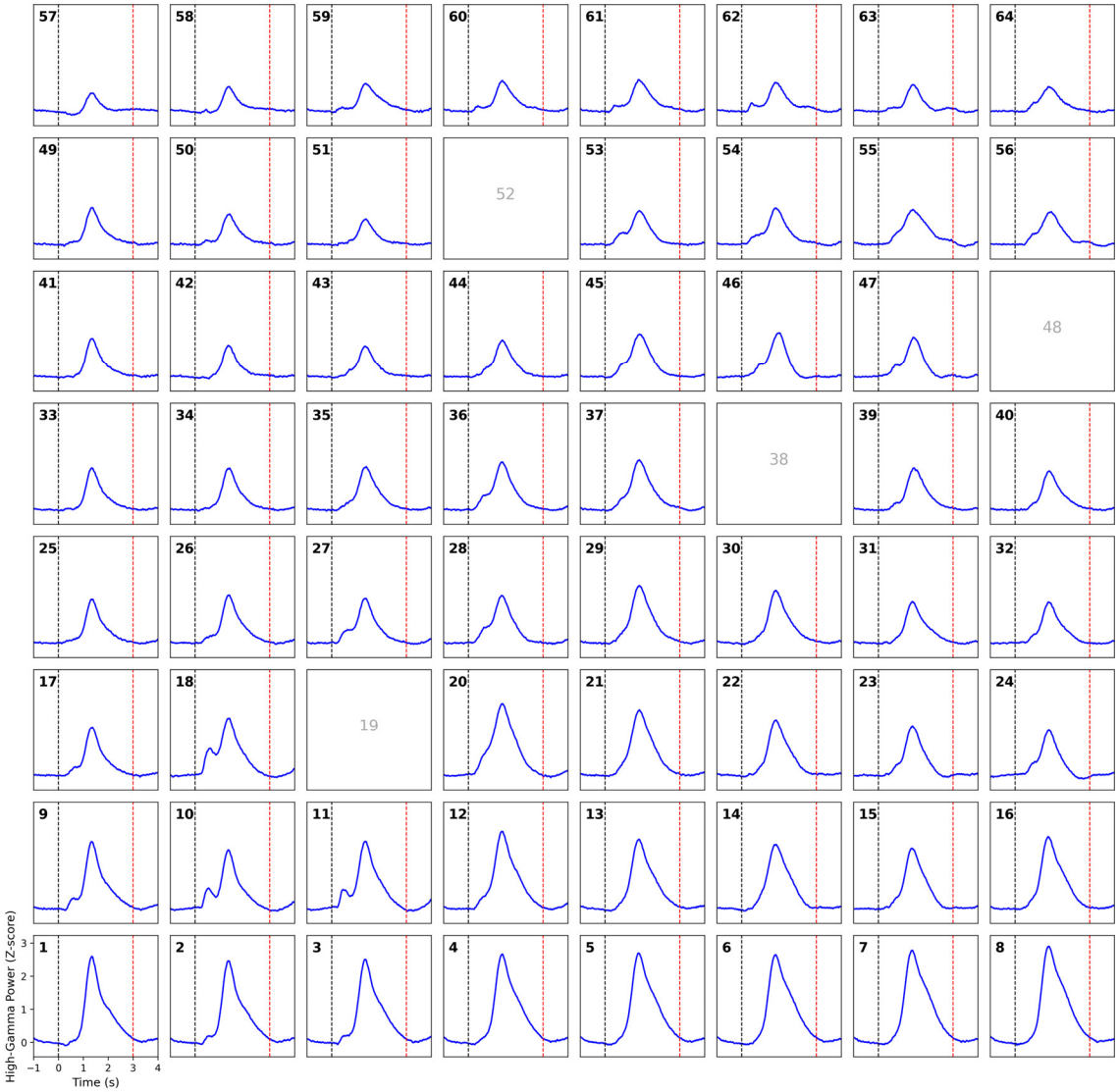


Figure 2. Trial-averaged high-gamma responses across all channels. Each panel displays the HG (70 -170Hz) time course. The blue traces show the normalized amplitude (Z-score) over time. The vertical black dashed line in each panel indicates stimulus onset (0s), and the red dashed line marks the end of the active speech period (3s). White panels denote channels excluded from the analysis.

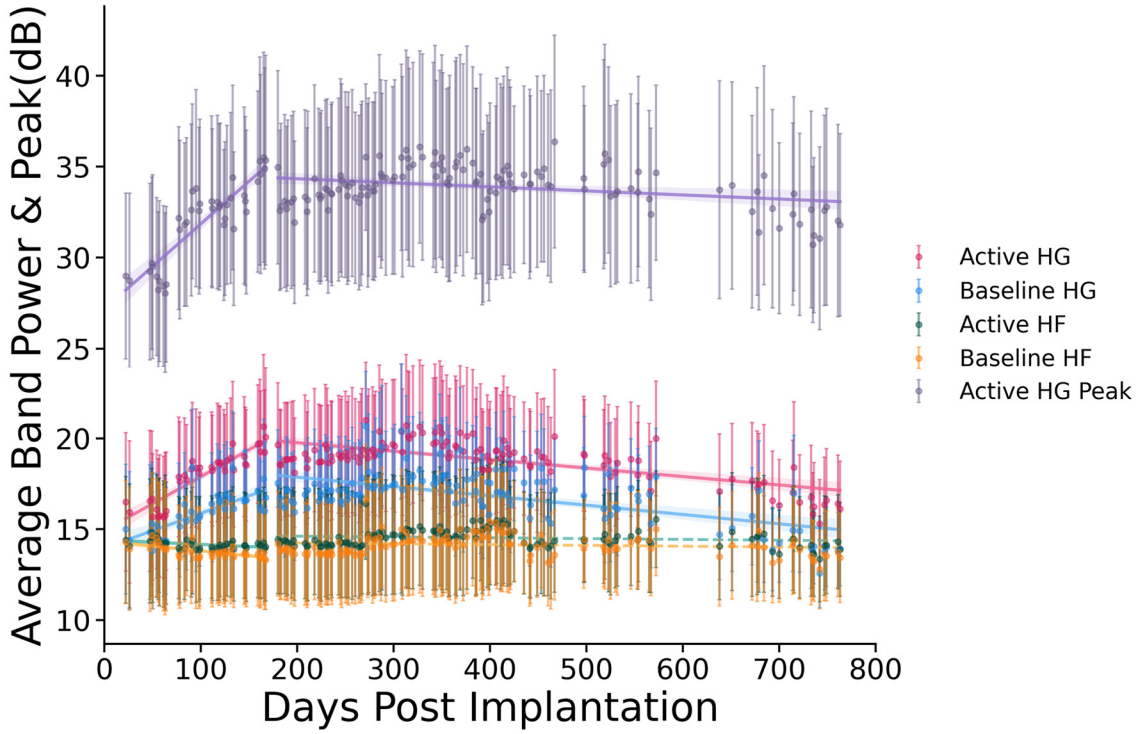


Figure 3. Longitudinal analysis of ECoG signal power across early and late post-implantation phases. Mean HG and HF band power are presented during the active (speech) and baseline (rest) periods, along with HG peak responses (peak amplitude) during speech trials. Each data point represents the average of 60 electrodes, and vertical bars indicate standard deviations. Solid lines represent statistically significant time trends ($\alpha < 0.005$, two-tailed t-test, with Bonferroni correction, k = 10), while dashed lines indicate non-significance. The shaded band surrounding the regression line denotes the 95% confidence interval.

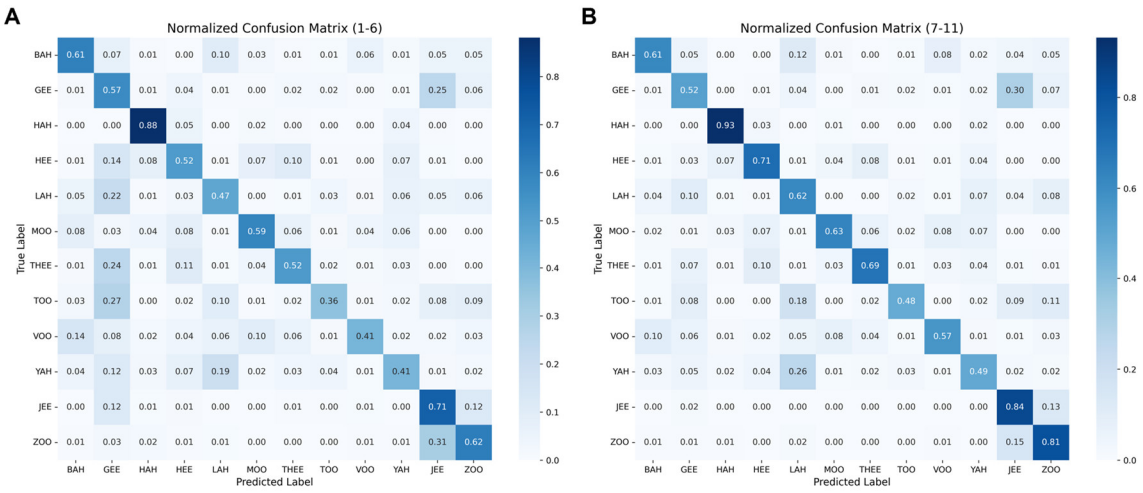


Figure 4. Confusion matrices show classification performance across all 12 syllables for models trained on data from the early (1-6) and late (7-11) periods. Results from training with late-stage data (B) generally outperformed those from early-stage data (A) on nearly all syllables, although the misclassification patterns remained similar across both periods.

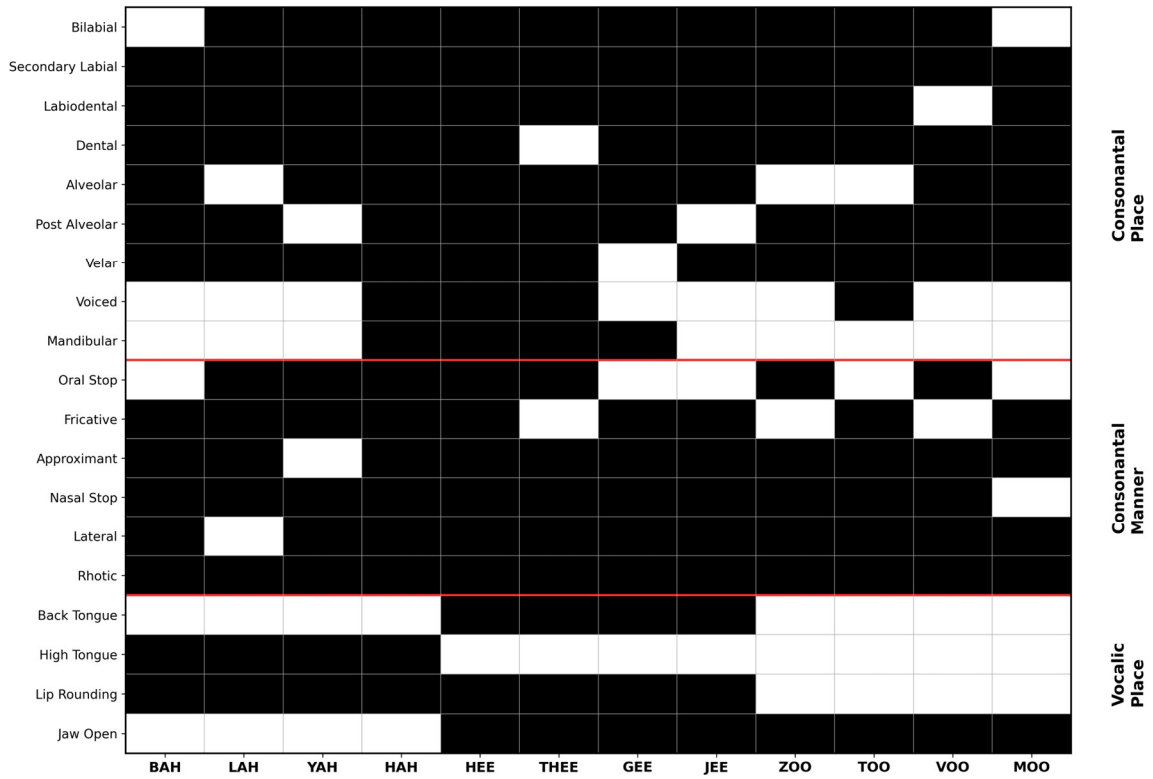


Figure 5. Binary International Phonetic Alphabet (IPA) feature matrix for selected syllables. Each column corresponds to one of the 12 syllables, and each row represents a distinct articulatory feature derived from the IPA. The rows are grouped into three functional classes indicated on the right-hand axis: consonantal place, consonantal manner, and vocalic place. Red horizontal rules mark the boundaries between the groups. A white cell denotes that the feature is present in the corresponding syllable, whereas a black cell indicates its absence.

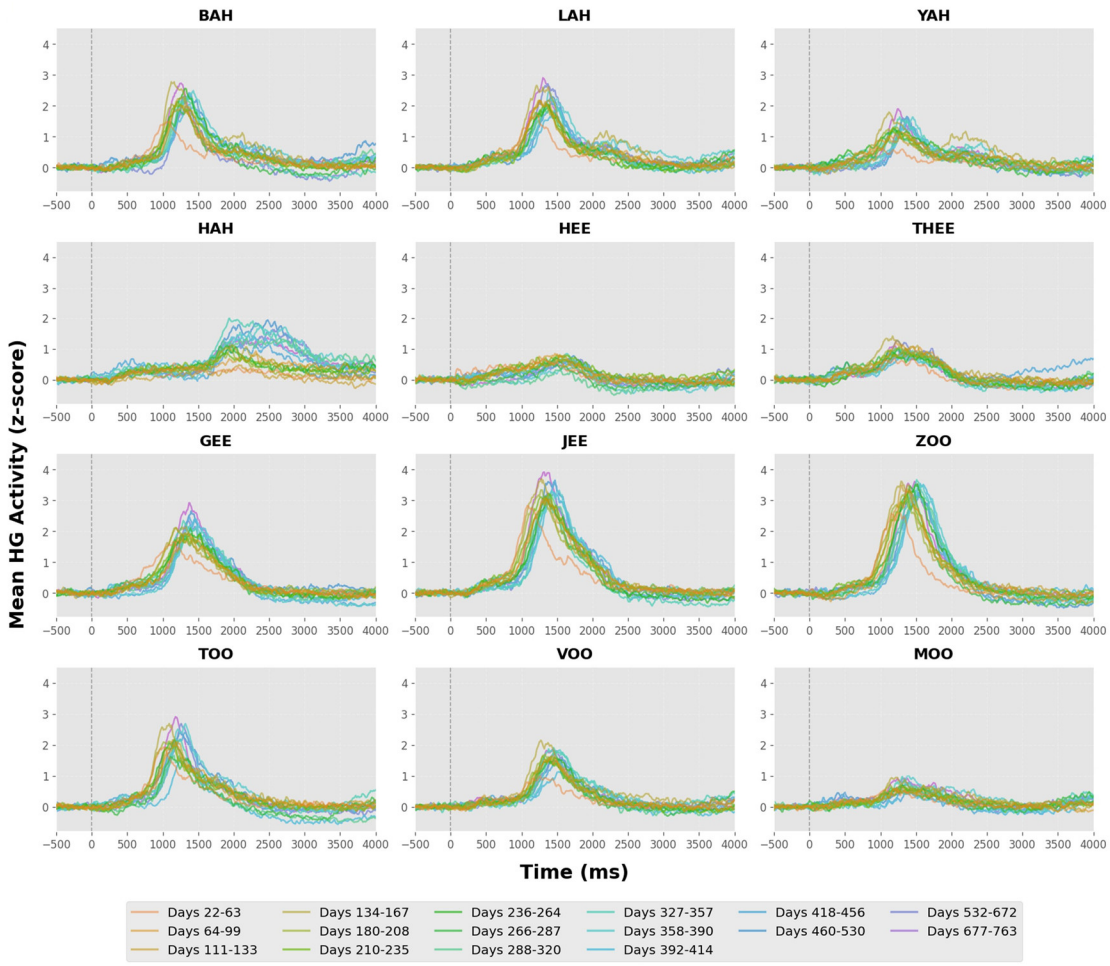


Figure 6. Temporal dynamics of HG activity across syllables. Each subplot represents the average HG response of all electrodes for each syllable, with the dashed line indicating cue onset. Each colored line corresponds to the trial-averaged activity within a sliding window of 10 consecutive recording sessions. Although there were differences in peak latency and amplitude across syllables, most syllables triggered a significant HG peak approximately 1000- 2000ms after cue onset.

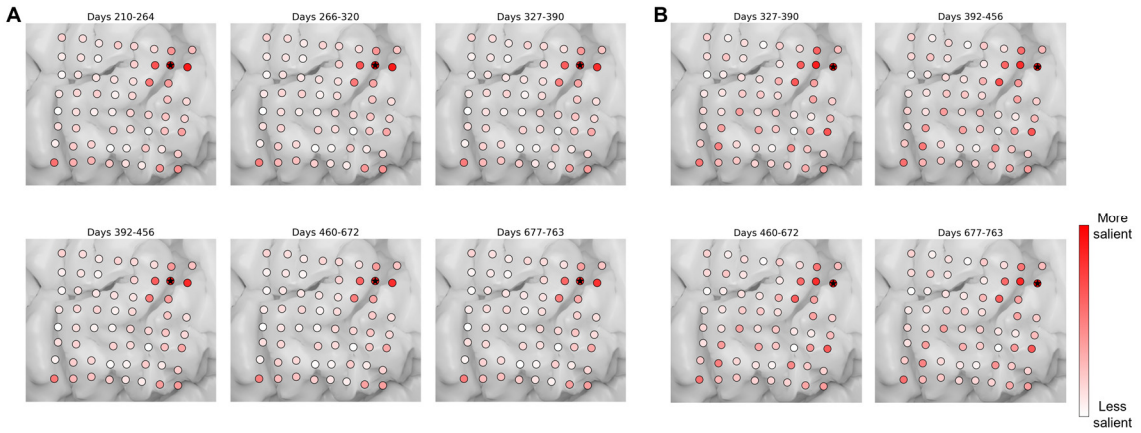


Figure 7. Electrode saliency maps across 20-session windows. (A) Electrode-wise saliency maps were computed from the CNN model trained on months 1-6. (B) Corresponding maps based on the model trained on months 7–11. Each dot represents an electrode, with color intensity indicating its contribution to decoding accuracy (red: more salient). Electrodes marked with a black star denote the highest saliency in each window. The most salient electrodes were consistently located in the dorsal lateral portion of the ECoG grid, which is associated with lip movements.

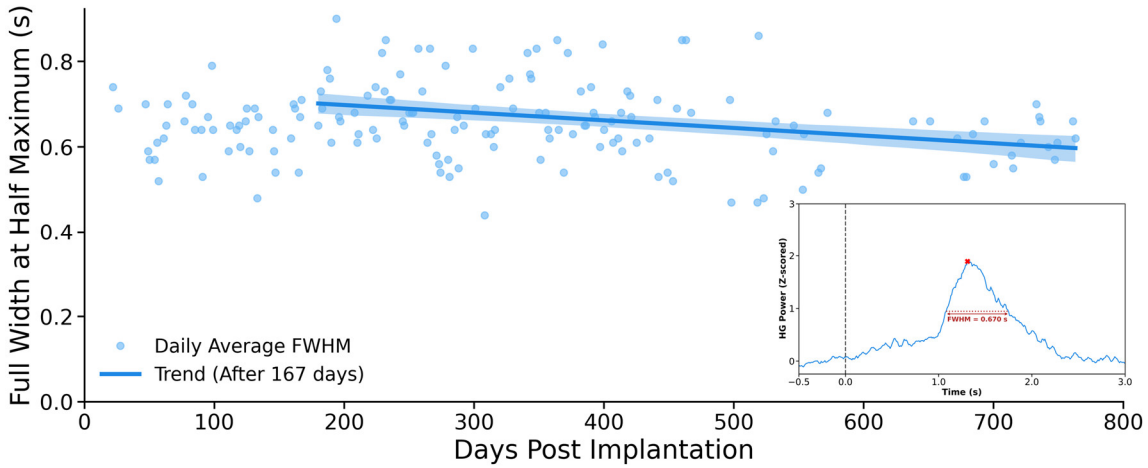


Figure 8. Progressive shortening of the z-scored HG waveform duration over 2 years. The scatter plot shows the daily average full width at half maximum (FWHM) of the z-scored HG power envelope for each recording session. The solid line and shaded area represent the linear regression fit and 95% CI (slope = -0.00018s/day, 95% CI: [-0.00027, -0.00008], P = 0.0003), respectively, for the period after the first six months. (Inset) Example of FWHM calculation on a representative daily-averaged HG waveform. The FWHM (red dotted line) is measured at the 50% value of the peak amplitude (red cross). The x-axis represents time relative to the cue onset.

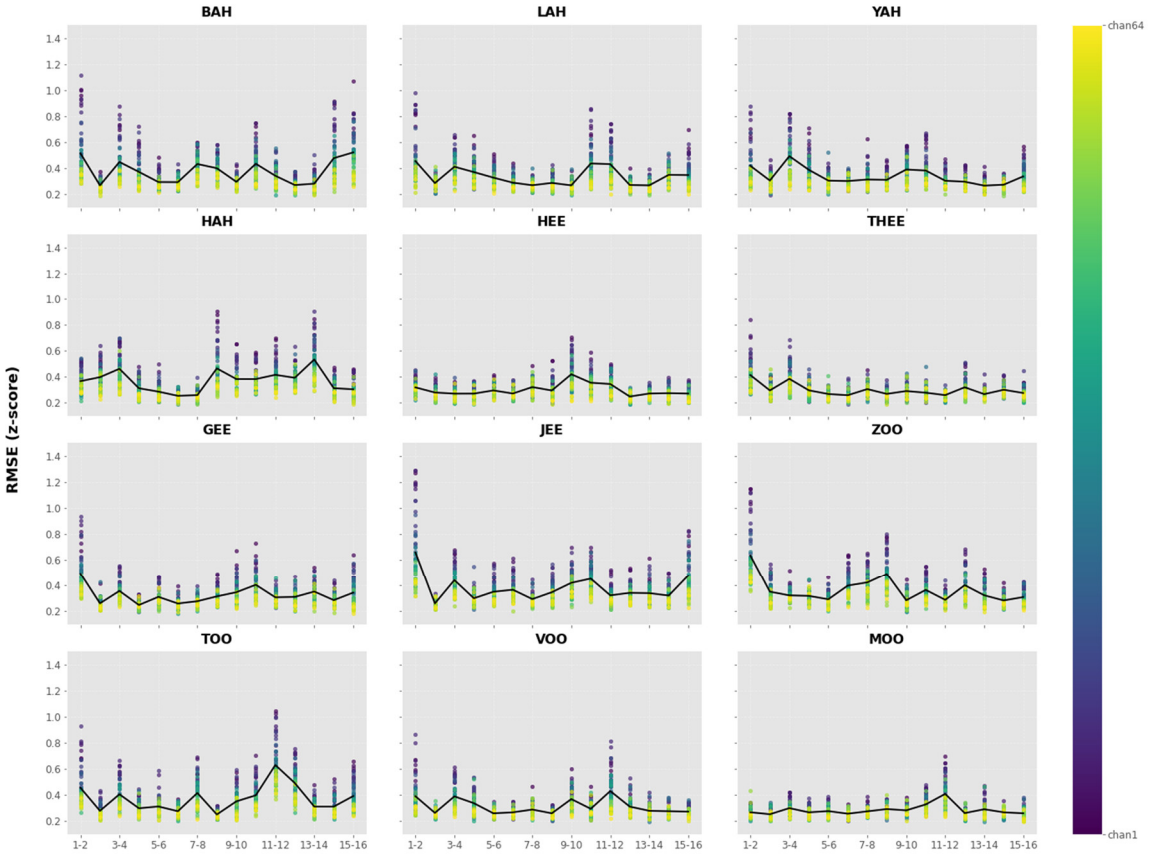


Figure 9. Longitudinal stability of RMSE of HG responses for each syllable. The RMSE of HG responses between adjacent session bins for each syllable is computed independently for each electrode (eMethods 4). Colored points represent per-channel RMSEs, with channel color denoting spatial identity (see the colormap). Black lines show the mean RMSE across all electrodes. Stability remained high across most time points (average RMSE below 0.6 SD), with limited fluctuations and no systematic drift, indicating the preservation of the representational structure of speech-related activity over long-term implantation.

Supplementary Table 1:

Table 1 Slopes of regression lines for all metrics.

Metrics	1 – 6 Months	7 – 25 Months
Active HG	**0.029711**	**-0.004657**
Baseline HG	**0.019162**	**-0.005140**
Active HF	-0.002416	-0.000448
Baseline HF	**-0.005378**	-0.000454
Active HG peak	**0.046740**	**-0.002228**
ActR	**0.008789**	**0.000412**
SNR_{active}	**0.032127**	**-0.004209**
$SNR_{baseline}$	**0.025888**	**-0.004579**
HG Response Peak (Z-score)	**0.004765**	**0.000840**

All metrics have slopes in dB/day, except for HG Response Peak, which has the slope expressed in Z-score/day.
Bolded values highlight statistical significance, $\alpha < 0.0028$, two-tailed t-tests with Holm-Bonferroni correction, k = 18. The precise p-value for all significance data is lower than 0.0001.

Reference

1. Wilson GH, Stavisky SD, Willett FR, Avansino DT, Kelemen JN, Hochberg LR, et al. Decoding spoken English from intracortical electrode arrays in dorsal precentral gyrus. J Neural Eng. IOP Publishing; 2020;17:066007. https://doi.org/10.1088/1741-2552/abbfef

2. Wyse-Sookoo K, Luo S, Candrea D, Schippers A, Tippett DC, Wester B, et al. Stability of ECoG high gamma signals during speech and implications for a speech BCI system in an individual with ALS: a year-long longitudinal study. J Neural Eng. 2024;21. https://doi.org/10.1088/1741-2552/ad5c02

3. Luo S, Angrick M, Coogan C, Candrea DN, Wyse-Sookoo K, Shah S, et al. Stable Decoding from a Speech BCI Enables Control for an Individual with ALS without Recalibration for 3 Months. Adv Sci Weinh Baden-Wurtt Ger. 2023;10:e2304853. https://doi.org/10.1002/advs.202304853

4. Berisha V, Sandoval S, Utianski R, Liss J, Spanias A. Characterizing the distribution of the quadrilateral vowel space area. J Acoust Soc Am. 2014;135:421–7. https://doi.org/10.1121/1.4829528

5. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework. Proc 25th ACM SIGKDD Int Conf Knowl Discov Data Min [Internet]. New York, NY, USA: Association for Computing Machinery; 2019 [cited 2025 Sept 3]. p. 2623–31. https://doi.org/10.1145/3292500.3330701

6. Loshchilov I, Hutter F. SGDR: Stochastic Gradient Descent with Warm Restarts [Internet]. arXiv; 2017 [cited 2025 Sept 3]. https://doi.org/10.48550/arXiv.1608.03983

7. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization [Internet]. arXiv; 2019 [cited 2025 Sept 3]. https://doi.org/10.48550/arXiv.1711.05101

8. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection. 2017 IEEE Int Conf Comput Vis ICCV [Internet]. 2017 [cited 2025 Sept 3]. p. 2999–3007. https://doi.org/10.1109/ICCV.2017.324

9. Anumanchipalli GK, Chartier J, Chang EF. Speech synthesis from neural decoding of spoken sentences. Nature. 2019;568:493–8. https://doi.org/10.1038/s41586-019-1119-1

10. Boersma P, van Heuven V. PRAAT, a system for doing phonetics by computer. Glot Int. 2001;5:341–5.

11. Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. Int Conf Learn Represent ICLR. Banff, Canada: Banff, Canada; 2014.

12. Fischl B. FreeSurfer. NeuroImage. 2012;62:774–81. https://doi.org/10.1016/j.neuroimage.2012.01.021