High-resolution neural recordings improve the accuracy of speech decoding

Suseendrakumar Duraivel[1], Shervin Rahimpour[2,3], Chia-Han Chiang[1], Michael Trumpis[1], Charles Wang[1], Katrina Barth[1], Shivanand P. Lad[2], Allan H. Friedman[2], Derek G. Southwell[2,4,6], Saurabh R. Sinha[7], Jonathan Viventi[1,2,4,6, #], Gregory Cogan[2,5,6,8,9, #]

[#]Corresponding authors

Author affiliations:

1 Department of Biomedical Engineering, Duke University, Durham, NC, USA

2 Department of Neurosurgery, Duke School of Medicine, Durham, NC, USA

3 Department of Neurosurgery, Clinical Neuroscience Center, University of Utah, Salt Lake City, UT, USA

4 Department of Neurobiology, Duke School of Medicine, Durham, NC, USA

5 Department of Neurology, Duke School of Medicine, Durham, NC, USA

6 Duke Comprehensive Epilepsy Center, Duke School of Medicine, Durham, NC, USA

7 Penn Epilepsy Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

8 Department of Psychology and Neuroscience, Duke University, Durham, NC, USA

9 Center for Cognitive Neuroscience, Duke University, Durham, NC, USA

Correspondence to:

Gregory B. Cogan

Department of Neurology

Duke South, Blue Zone 4519, Durham, NC 27710

gregory.cogan@duke.edu

Jonathan Viventi

Biomedical Engineering

1149 CIEMAS, 101 Science Dr, Durham, NC 27705

j.viventi@duke.edu

Running title: Micro-ECoG Speech Decoding

Keywords: Speech, micro-ECoG, decoding, neural prostheses, BCI

Abstract

Patients suffering from debilitating neurodegenerative diseases often lose the ability to communicate, detrimentally affecting their quality of life. One promising solution to restore communication is to decode signals directly from the brain to enable neural speech prostheses. However, decoding has been limited by coarse neural recordings which inadequately capture the rich spatio-temporal structure of human brain signals. To resolve this limitation, we performed novel, high-resolution, micro-electrocorticographic (μECoG) neural recordings during intra-operative speech production. We obtained neural signals with 57× higher spatial resolution and 48% higher signal-to-noise ratio compared to standard invasive recordings. This increased signal quality improved phoneme decoding by 35% compared to standard intracranial signals. Accurate decoding was dependent on the high-spatial resolution of the neural interface. Non-linear decoding models designed to utilize enhanced spatio-temporal neural information produced better results than linear techniques. We show for the first time that μECoG can enable high-quality speech decoding, demonstrating its ability to improve neural interfaces for neural speech prostheses.

Abstract Word Count: 159/175

1 Main

Motor disorders such as amyotrophic lateral sclerosis (ALS) and clinical locked-in syndromes greatly reduce or eliminate patients' ability to verbally communicate and dramatically affect their quality of life. There are ~6000 new ALS diagnoses in the United States each year, resulting in near-complete loss of motor function but preserving cognitive functions[1,2]. Computer aided technologies have been reported to improve quality of life but are often limited by slow processing and inefficacies[3–5]. Neural speech prostheses offer the potential for a faster and more reliable means to decode conversational speech directly from the brain. The development of high-resolution neural recordings would enable accurate decoding of spoken features which is paramount to a successful neural speech prosthesis.

Previous attempts to accurately decode speech have typically utilized invasive macro electrocorticography (macro ECoG, 10 mm inter-electrode spacing and 2.3 mm exposed diameter) and high-density ECoG (4 mm inter-electrode spacing and 1 mm exposed diameter) that target speech motor cortex (SMC) during speech production[6–9]. These studies demonstrated that SMC encodes articulatory properties of speech motor sequences which can form the building blocks of successful speech decoding. These articulatory features are subsequently transformed into acoustic speech[10–12] or can be combined to form fundamental linguistic units such as phonemes[13–17], which can eventually

be aggregated into words[18,19], and sentences[20,21]. Accurate resolution of these features is therefore a crucial component for speech decoding.

The rich spatio-temporal structure of human brain signals occurs at small spatial scales. Recordings of brain signals have previously been limited by ECoG recordings, which are typically measured from 64 – 128 contacts spaced 4 – 10 mm apart. This limitation is particularly relevant for signals that are both spatially specific and highly informative. One such signal is the high gamma band (HG: 70 – 150 Hz), which has been shown to index local neural activity from the surface of the brain[22]. HG has a high correlation with multi-unit firing and also shows high spatial specificity[23,24]. Further, HG has been previously shown to more accurately estimate neural-firing patterns that are stable over longer periods as compared to single units[25]. Micro-scale neural recording of HG could therefore enable accurate resolution of speech-articulatory features.

Information present over small distances in the human brain over is distinct during speech production. Previous methods that quantified HG signal-sharing in SMC during speech articulation have shown low inter-electrode correlation (r = 0.1-0.3 at 4 mm spacing), indicating that articulatory neural properties are distinct at millimeter scale resolutions[26–28]. Consequently, speech decoding studies that utilized HG activity have identified a boost in performance with higher-density cortical sampling[19,20]. Neural speech decoding using 4-mm-spaced arrays showed up to a 5× increase in phoneme prediction compared to 10-mm-spaced arrays[14,17]. These results show that decoding performance improved with increased electrode density, further motivating higher resolution neural recordings to accurately resolve HG for reliable speech decoding.

Previous work using high-density micro-electrocorticographic (μECoG) arrays in other domains have shown the ability to resolve micro-scale neural features. In rodents and non-human primates, sub-millimeter spacing revealed fine-scale sensory topologies consistent with intracortical electrodes.[29–32]. In humans, this high-resolution electrode technology has also enabled the identification of micro-scale epileptic signatures of seizure-onset-zones in epileptic patients[33–37] and have shown improvements in motor neural prostheses by resolving the motor cortex at millimeter-level resolution[38–41]. These results show that high-resolution neural recordings could resolve micro-scale articulatory features from SMC.

In the present work, we demonstrate the first use of μECoG arrays for speech decoding in humans. We recorded intra-operatively from speech-abled patients using liquid crystal polymer thin-film (LCP-TF) μECoG arrays (1.33 – 1.72 mm inter-electrode distance, 200 μm exposed diameter electrodes) placed over SMC during a speech production task. These recording devices enable the high-resolution spatio-temporal sampling of local neuronal activity which produced superior speech decoding. We decoded speech by predicting the actual spoken phonemes from HG neural activations. We compared our

results from μECoG arrays to neural decoding from macro intracranial electroencephalographic (IEEG) recordings, to empirically validate our improved decoding results. We also show that μECoG decoding relies on the ability to resolve micro-scale spatial and temporal features of the neural signal. Lastly, we leverage this access to micro-scale neural signals by using a nonlinear decoding model to decode entire speech sequences. We show the first use of μECoG arrays for neural decoding of speech. This novel technology could lead to improved neural speech prostheses.

2 Results

2.1 High resolution neural activity of speech

We examined micro-scale speech neural activity from four subjects recorded using LCP-TF μECoG electrodes in the intraoperative setting. We used two versions of the LCP-TF μECoG electrode arrays to record speech neural activations from SMC: a 128-channel subdural array (**Fig. 1A & 1B (top)** 8 x 16 array; inter-electrode distance: 1.33 mm) and a 256-channel subdural array (**Fig. 1A & 1B (bottom)** 12 x 24 array; inter-electrode distance: 1.72 mm). The μECoG electrode arrays had up to 57× electrode density with respect to macro-ECoG arrays and up to 9× higher density compared to high-density ECoG arrays (**Fig. 1C**). Subjects S1 (M: 61 years old), S2 (M: 63 years old), and S3 (M: 62 years old) underwent surgery for the treatment of movement disorders and were implanted with a 128-channel array that was tunneled through the burr-hole during deep brain stimulator (DBS) implantation (see Methods). Each of the subjects completed three task blocks (52 trials per block) of a speech repetition task, during which the subjects were asked to listen and repeat back auditorily presented nonwords. Subject S4 (F: 26 years old) underwent surgery for a tumor resection and was implanted with a 256-channel array. This subject completed one block of the same speech repetition task as other subjects (**Fig. 1D**, Supplementary Table 1 & 2). Subjects took on average 1.1 seconds (range = 0.7 to 1.5 s) to repeat auditorily presented nonwords and had an average utterance duration of 450 ms (range = 300 to 700 ms) (**Fig. 1E,** Supplementary Table 3). Subjects correctly repeated the nonwords on more than 95% of the trials (**Fig. 1F,** S1: 96%, S2: 98%, S3: 98%, S4: 100%), indicating that subjects understood and could complete the task in the intraoperative setting. The overall experiment time lasted up to 15 minutes and total utterance duration lasted 0.47 to 1.5 minutes (Supplementary Table 3).

We observed uniform *in vivo* impedance across the array (S1: 81.3 ± 3.8 kOhm, S2: 12.9 ± 1.3 kOhm, S3: 27.5 ± 4.8 kOhm, S4: 19.7 ± 4.2 kOhm, mean ± standard deviation; Supplementary Fig. 1), and discarded electrodes with higher impedance (> 1 MOhm) from neural analysis. We observed significant modulation of spectro-temporal neural activations during speech articulation, including prominent HG band power increases. These distinct spatial patterns were seen in each patient and are shown in example arrays for 128 channel (S1) and 256 channel (S4) arrays in **Fig. 2A**. HG power increases were

aligned to the speech utterance in individual electrodes (-250 ms to 250 ms, **Fig. 2B,** Supplementary Fig. 2) and were identified as statistically significant as compared to a pre-stimulus baseline using a non-parametric permutation test with an FDR-corrected alpha threshold of $p < 0.05$ (see Methods). Significant electrodes are highlighted in Figure 2 with black borders: S1 111/128 significant channels, S2 111/128, S3 60/128, and S4 149/256.

Next, we sought to determine the benefit of μECoG for recording neural signals at higher fidelity as compared to standard methods. We examined the evoked-signal-to-noise ratio (ESNR) of HG power from μECoG neural recordings (-250 ms to 250 ms with respect to speech utterance onset) and compared it to standard IEEG (see Methods). We pooled within-subject IEEG electrodes (electrodes implanted during clinical pre-operative epileptic monitoring on separate patients: ECoG and stereo-electroencephalography – SEEG) that were anatomically localized to SMC and which exhibited significant HG power during speech articulation. Neural signals from μECoG recordings demonstrated a 48% (1.7 dB) increase in measured signal-to-noise ratio (median HG ESNR = 2.9 dB, S1: 4.2 dB, S2: 3.5 dB, S3: 1.9 dB, and S4: 2.1 dB, $p< 0.05$, Mann-Whitney U test) as compared to recordings from standard IEEG (median ESNR = 1.2 dB, **Fig. 2C**).

To quantify the spatial resolution of μECoG neural signals, we computed the Pearson correlation coefficient of the HG envelope[26] between each micro-electrode pair. The correlation increased with proximity from 10 mm to 1.33 mm and remained largely uncorrelated at the equivalent spatial resolutions of macro ECoG for all 4 subjects (**Fig. 2D**: $r < 0.2$ at 10 mm spacing). HG neural signals at less than 2 mm were more correlated ($r = 0.6$), but not fully correlated, suggesting that speech information contained in HG neural activations are spatially discriminative at scales below 2 mm.

To investigate micro-scale neural information, we examined spatio-temporal activations specific to key units of speech production: articulatory features of individual phonemes. We first examined spatio-temporal activity patterns by averaging normalized HG activity across trials specific to first-position phonemes of the nonword utterances. All four subjects exhibited distinct spatio-temporal patterns for four different articulators (/a/ - low, /i/ - high-vowel, /b/ - labial, /g/ - dorsal tongue) with respect to the speech utterance onset (**Fig. 3** & Supplementary Fig. 3), as highlighted in **Fig. 3A** for example subject S1.

We next sought to investigate how motor articulatory features were organized in neural space. Previous studies examining speech production have identified functional organization patterns of phonemes based on their composite articulators in SMC[42,43]. We sought to identify similar population-driven articulator patterns for μECoG HG (500 ms window; 100 time points) in a cortical state-space and determine if this organization is dependent on spatial resolution. To model cortical state-space, we

used singular value decomposition (SVD) to transform spatio-temporally covarying HG activity into low-dimensional features (80% of variance explained), and used t-distributed Stochastic Neighbor Embedding (tSNE)[44] to visualize the transformed principal-component scores in a two-dimensional vector space. **Fig. 3B** shows clear separation of speech trials in tSNE state-space for both articulators as well as individual phonemes. We examined the contribution of the high spatial-resolution of μECoG by repeating the state-space analysis of spatially subsampled HG (Poisson disc sampling spatially constrained across the array, Supplementary Fig. 4A). We observed decreased distinction for both articulators and phonemes with reduced spatial sampling (**Fig. 3C**). This effect is further quantified by a significant decrease in silhouette scores for sub-sampled versions in both articulator and phoneme space (**Fig. 3D,** Supplementary Fig. 4B). To statistically assess this difference, we performed a one-way ANOVA to examine the effects of spatial subsampling on state-space clustering which revealed a significant main effect of subsampling in both articulator ($F_{3, 196} = 114$, $p = 9.6e\text{-}43$) and phoneme space ($F_{3, 196} = 47.8$, $p = 3e\text{-}23$). Post hoc t-tests with Bonferroni correction showed that mean values were significantly different between all groups in the articulatory space ($p < 0.01$). Differences were also significant in the phoneme space, except for between 100% and 50% subsampling ($p = 0.17$). These results clearly demonstrate that high-resolution neural recordings enable more accurate representation of motor features in SMC.

2.2 Decoding phonemes using μECoG spatio-temporal features

Successful neural speech prostheses will require the ability to decode the full range of human speech and language. One solution to enable this ability is to focus on compositional units that enable generative language. We therefore sought to evaluate the ability of μECoG to decode the smallest compositional unit of speech: the phoneme. We performed a 20-fold, nested, cross-validated decoding model of manually aligned phonemes based on a μECoG, HG, low-dimensional subspace (SVD decomposed) using a supervised linear discriminant model (LDA; see Methods). We selected the eigenvalues based on nested cross-validation that explained 80% of neural variance for phoneme prediction. Benefiting from high spatial sampling, we observed strong decoding performance in all subjects for predicting phonemes in all positions (**Fig. 4A & 4 B**, $p < 0.01$, Binomial test against chance model[45]). All subjects had higher accuracies for decoding phonemes in the first position (average P1 across subjects: 49.9%, chance = 11.11%) when compared to phonemes in the second (P2: 45%, $p = 0.0097$) and third positions (P3: 38.8%, $p = 0.0250$; one-sided paired t-test). S1 and S2 exhibited higher decoding performances across all positions which we attribute to higher HG-SNR and increased utterance duration for these subjects (**Fig. 4B**). S1-P1 (Subject 1, phoneme position 1) had a maximum prediction score of 57% with the best consonant (/g/) obtaining a decoding score of 84.2% and the best vowel (/i/) obtaining a decoding score of 75.3% (**Fig. 4A**). Similarly, S2-P1 had

similar maximum prediction scores with the best consonant (/g/) with a decoding score of 72.7% and the best vowel (/u/) with a decoding score of 83.3%. Decoding accuracies were reduced when manually aligned phonemes were pooled across all positions (S1: 40.72%, S2: 42.96%, S3: 20.92%, and S4: 41.67%). This reduction in performance suggests that there is unique position-based information present in the signal, indicating a need for position-dependent decoding. The lower prediction quality of S3 was attributed to poor contact between the electrodes and SMC (as indicated by low HG-SNR). S4 had comparable SNR and spatial coverage as S1 and S2, however this subject completed only one out of three blocks during the intraoperative experiment resulting in reduced data duration for the decoding model and highly varying performance across phoneme positions. To show that the reduced performance obtained from S4 was likely due to the this decreased recording duration, we subsampled trials across all subjects and phoneme positions and found comparable decoding accuracy when the recording duration for other subjects was also reduced (Supplementary Fig. 5, median accuracy at 50 subsampled trials – S1: 36%, S2: 27%, S3: 15% and, S4: 44%). Our decoding results represent the best phoneme decoding scores for surface neural recordings and validate the use of μECoG for the development of neural speech prostheses.

To directly assess the benefit of recording neural signals for speech at the micro-scale, we compared μECoG decoding to standard IEEG recordings (SEEG/ECoG). We performed the same phoneme decoding analysis using neural signals recorded from eleven epilepsy patients (7 females, mean age = 28.5, 4 ECoG, 7 SEEG) who were implanted with standard IEEG electrodes during pre-operative epilepsy monitoring. For comparison between electrodes, we performed a 4-way decoder (using SVD-LDA; see Methods) to decode between vowels and quantified the decoding performance. For each patient, we anatomically restricted electrodes to SMC and selected the electrodes that exhibited significant HG ESNR during the speech utterance (see Methods). Our results showed that on average, μECoG outperformed standard clinical recordings by 36% (**Fig. 4C**: ECoG: 36.7%, SEEG: 36.5%, μECoG: 71%, chance 25%, $p < 0.05$, one-sided permutation tests with 10,000 iterations). This boost in performance results show the ability of μECoG to enable more accurate speech decoding as compared to standard clinical recording techniques.

Since high-density μECoG demonstrated both a greater ability to resolve higher ESNR and stronger decoding performance as compared to standard clinical recordings, we next sought to directly determine the relationship between decoding performance and ESNR. To quantify this relationship, we calculated the univariate phoneme decoding performance for every individual μECoG electrode along with their corresponding HG-ESNR (dB) value (**Fig. 4D**). Linear model fits indicate moderate but significant correlations for S1 (Pearson r = 0.23, $p < 0.01$), S2 (r = 0.23, $p < 0.01$), and S3 (r = 0.19, $p < 0.05$), and a strong correlation for S4 (r = 0.44, $p < 0.001$). Further, we observed increases in accuracy

as a function of HG-ESNR at a mean rate of 0.4% per dB for S1, S2, and S3, and 1.1 % per dB for S4. These results show a moderate but significant relationship between the ability to capture higher SNR micro-scale activations using high-density μECoG and the corresponding ability to decode speech.

Finally, we sought to characterize the types of decoding errors that were present in high-density μECoG recordings. We reasoned that if our ability to decode phonemes was largely based on our ability to resolve their constituent articulatory features in the motor cortex, decoding errors should not be uniform across phonemes, but rather reflect similar articulatory properties from other phonemes. For example, in **Fig. 4A**, we see increased confusion between phonemes with similar phonological features (bilabial plosives /p/ vs. /b/)[6,14,16]. To systemically characterize these types of errors, we calculated each phonemes' phonological feature set[46] by assigning a 17-bit categorical vector to each phoneme (see Methods). We used Hamming distance to calculate phonological distance for each phoneme pair (Supplementary Fig. 6) so that phonemes with similar phonological features had lower distances (e.g., /p/ & /b/ differed by 1 bit, whereas, /p/ & /g/ differed by 5 bits). We found that our decoding models' confusion/error rate decreased with phonological distance across all phoneme positions (**Fig. 4E**, P1: $F_{1,79} = 90.4$, $p = 9.9e\text{-}15$, P2: $F_{1,79} = 19.7$, $p = 2.8e\text{-}5$, P3: $F_{1,79} = 28.6$, $p = 8.4e\text{-}7$; F-statistic test against constant model). We observed similar inverse relationships for all subjects and the errors were lower than chance, (**Fig. 4F** chance error: 2.42 bits; uniform confusion matrix). S1 and S2 had the lowest phonological distance error across all phoneme positions (less than 1 bit; except S1-P3), indicating that phoneme misclassifications were predominantly biased by similar phonological feature articulation. This analysis showed that decoding using high-density μECoG exhibited errors that were systemic, largely due to speech articulatory structure and not random misclassifications.

2.3 Improved decoding with increased resolution, coverage, and contact size

We next sought to directly study the effect of spatial resolution and spatial coverage of neural signals and their effect on decoding. For our subsequent analyses, we focused on S1 and S2 as they both had the requisite SNR and amount of data for our decoding analyses (see Supplementary Figure 5 for a direct comparison). First, we examined the effect of spatial resolution on decoding by sampling electrodes (n ≥ 30 samples) across the array, using Poisson disc sampling (Supplementary Fig. 4A). For each sampling, the resultant spacing in millimeters was calculated using the formula:

$$pitch = \sqrt{\frac{X * Y}{ne}} \tag{1}$$

where, X and Y are length and width of the array in mm, and ne is the total number of electrodes in the sample. Accuracy values from the resultant spacing (rounded to the nearest integer) were then

compared to the 95% threshold of maximum decoding accuracy values obtained from the full array. We found a sharp increase in decoding performance with an increase in spatial resolution. The subset of electrodes with spacing less than 1.5 mm reached the 95% threshold of the full array (**Fig. 5A** left, & Supplementary Fig. 7A). The subsampling distributions from the 4 to 10 mm range (electrode distance range for standard IEEG) had median accuracy values of only 20-35% as compared to 55 - 60% for the full micro-array, an improvement from 17% to 29%. Next, to demonstrate the importance of spatial coverage, we subsampled electrodes from rectangular subgrids at various fixed spatial resolutions. Each of these subgrids were two dimensional windows with dimensions from 2 x 4 up to 8 x 16 electrodes. The decoding performance improved with increased coverage (**Fig. 5A** middle & Supplementary Fig. 7A). Finally, to assess the impact of micro-scale contact-size on speech decoding, we estimated various sized recording contact areas of recording by using rectangular subgrids with dimensions from 1 x 1 (effective 200 μm contact size) to 8 x 8 (effective 10 mm contact size) of spatially averaged raw neural recordings. HG signals extracted from these spatially averaged neural recordings were then used to decode phonemes. Decoding accuracy decreased with increased estimated contact size. The highest performance was at the smallest contact size (**Fig. 5A** – right & Supplementary Fig. 6A). We validated the robustness of this spatial analysis by showcasing similar decoding trends across all phoneme positions for both S1 and S2 (Supplementary Fig. 7B).

Since previous decoding work has characterized a widely different amount of phonemes, we next sought to characterize this spatial degradation of different N-way decoding models. For each electrode subsampling, we averaged the decoding accuracies of all possible N-way classifiers and obtained the simulated spacing required to achieve 90% and 95% threshold of full array decoding performance. We found that the spacing required to obtain the full array decoding performance was inversely proportional to the number of unique phonemes, indicating that the benefit of using high resolution neural recordings increases with the number of uniquely decoded phonemes (Supplementary Fig. 7C), further suggesting that as models scale up in size, the importance of high spatial sampling will be even larger. Together, these results directly validate the benefits of μECoG electrode arrays for highly accurate neural decoding through their high-spatial resolution, large spatial coverage, and micro-scale electrodes.

We next used our recordings of micro-scale signals to identify articulatory motor maps in SMC. These maps show the specific neural organization of effectors for motor speech output. Previous studies have identified average maps across multiple subjects in SMC for articulator encoding[43], however, subject-specific maps have not been well characterized due to the relatively sparse coverage using standard electrode arrays. Individual subject maps would better enable the development of subject-specific neural prostheses. To quantify these micro-scale neural articulatory maps, we trained an SVD-LDA model for each electrode to decode four articulatory features (low vowels, high vowels, labial

consonants, and dorsal-tongue consonants). The receiver-operator-characteristic area-under-the-curve (ROC-AUC) metric was used to evaluate each electrode's encoding preference for each articulator, and the values from all the electrodes were spatially arranged to derive micro-scale neural articulatory maps. We observed overlap of informative electrodes across each of the articulators. Despite this overlap, high-density μECoG arrays exhibited distinct clustering of articulatory features in all subjects, indicated by red contours (90 percentile of AUC) and corresponding centroids (**Fig. 5B &** Supplementary Fig. 8). For S1 with an 8 x 16 grid, AUC values were dispersed throughout the array for vowels, whereas there was distinct clustering for consonants (**Fig. 5B**, labial-consonants with 72.5 mm^2 surface area and dorsal-consonants with 17.7 mm^2 coverage, supplementary table 4). For subject S4 which had a 12 x 24 grid, we observed strong spatial clustering including both vowels and consonants (labials: 56.2 mm^2, dorsals: 79.9 mm^2, Supplementary Fig. 8 and Supplementary Table 4). These subject-specific, high-resolution articulator maps show the ability of high-density μECoG to resolve fine-scale motor representations for speech production in individual subjects, which could enable better individualized neural speech prostheses.

2.4 Sequential neural modeling reveals positional encoding of phonemes

Thus far our analyses have highlighted the ability of micro-scale recordings to resolve spatially specific neural signals. However, these signals reflect complex neural dynamics which also evolve in time. We sought to leverage our ability to resolve detailed spatial recordings of neural signals by examining how these signals evolve and change in time during speech utterances. We assessed these temporal dynamics through our ability to resolve phoneme sequences in time. We constructed an SVD-LDA model to train and test on HG neural time-segments with respect to the utterance. Sequences of the specific ordered phonemes within the utterance should only be resolvable if the sequence information is present in spatio-temporal HG neural activation patterns.

We obtained stable temporal decoding for phonemes across all positions. Further, the peak decoding values for each phoneme occurred sequentially times reflecting their order in the utterance (**Fig. 6A**: S1: 0.015, 0.125, and 0.425 s from the utterance onset for Position 1, 2, and 3 respectively, and for S2: 0.04, 0.28, and 0.39 s). This analysis demonstrates the ability to decode sequential phonemes in an utterance from the rich spatio-temporal neural features from micro-scale recordings. This ability to resolve these phoneme sequences also demonstrates our ability to resolve entire speech utterances from the decoding of phoneme units, a requirement for successful neural speech prostheses.

2.5 Phoneme sequence generation using seq2seq recurrent neural network (RNN)

Having demonstrated the presence of sequential phoneme information in the micro-scale neural signal, we sought to decode the entire utterance via the individual constituent phonemes without explicit

phoneme onset information. We reasoned that this scenario would more accurately reflect the use-case of a neural speech prosthetic for entire utterances, where manual phoneme boundaries are not available. We developed a phoneme generation technique to sequentially decode positional phonemes within the utterance. We used a seq2seq RNN model to translate HG neural activity to the entire phoneme sequence directly[20]. For each spoken utterance, HG activations (1 second window; 200 time points) from all significant µECoG electrodes were passed as input into an 'encoder-decoder' recurrent neural-network (**Fig. 6B**). The model was executed in three stages: 1) The temporal convolution layers (green arrows) to extract low dimensional neural features, 2) The bidirectional encoder RNN (blue boxes) to encode neural features along the temporal-domain. 3) The decoder RNN (orange boxes) to transform encoded neural features to phoneme sequences.

We trained the entire network using Adam[47] gradient descent optimization (learning rate = 1e-3) for 800 epochs (number of visits of training data), with $\ell 2$ regularization applied to all layers. The network hyperparameters were estimated using nested cross-validation (see Supplementary table 5).

We quantified the decoding performance by calculating the percentage of correctly predicted phonemes for each position. The resultant decoding accuracies and errors were considerably better for seq2seq models for both S1 (mean accuracy: 46.1% vs. 38.9%, mean phoneme error: 1.2 bits vs. 1.6 bits) and S2 (accuracy: 51.6% vs. 43.1%, phoneme error: 0.89 bits vs. 1.4 bits) as compared to equivalent linear decoding models (**Fig. 6C** and supplementary table 6). This result demonstrates that non-linear sequential decoding can more accurately reconstruct spoken sequences than linear methods that require explicit phoneme onset times. To further validate the importance of micro-spatial sampling for these results, we performed an analysis that spatially subsampled our electrodes which resulted in a reduction of the sequential decoding performance (**Fig. 6D** S1: $F_{3, 99} = 40.47$, $p = 5.4e\text{-}17$, S2: $F_{3, 99} = 99.97$, $p = 2e\text{-}79$). Post hoc t-tests with Bonferroni correction showed significant differences in mean across all subsampling groups for both subjects ($p < 0.01$). This sequential decoding method provides the basis for an algorithmic framework for the decoding of entire speech utterances using high-resolution neural recordings.

3 Discussion

We demonstrate the benefits of using micro-scale neural recordings for human speech decoding and as an important component of neural speech prostheses. We successfully performed intra-operative recordings from four speech-abled patients (3 movement disorder patients and 1 tumor patient) who completed a nonword speech repetition task. Even with the time-constrained experimental procedure (up to 15 minutes of recording duration/ 1.04 minutes of total spoken duration; **Fig. 1E,** Supplementary Table 3), our µECoG electrodes recorded neural signals in SMC that produced significantly higher HG-

ESNR as compared to standard intracranial methods (ECoG/SEEG). The measured HG activations from SMC were distinctive at the millimeter scale during the speech utterance, as validated by inter-electrode correlation of r ~= 0.6 (**Fig. 2D**) at <2mm spacing, which both extends previous work with coarser spatial sampling and demonstrates the informative nature of signals at the millimeter scale[26,27]. This result is also in keeping with previous work that sampled microscale epileptic activity using high-density μECoG arrays which revealed a striking heterogeneity in the spatio-temporal dynamics of interictal events[48,49]. Importantly, in that work, microseizures were most often observed occurring on just one to two electrodes spaced 750 μm apart and were not visible on neighboring electrodes. The higher SNR combined with evidence for distinct neural patterns of activations at the micro-scale motivates the benefits of using micro-recording technologies for human speech decoding.

On examining the spatio-temporal activation of μECoG HG, we performed a cortical state-space analysis that resulted in clear clustering of articulatory features and phonemes, which is consistent with previous work[42]. We built on this finding by showing that this clustering was dependent on micro-scale spatial resolution as a decrease in sampling produced significantly worse clustering results. This inverse relationship shows the benefit of high-resolution neural recordings to accurately model speech cortical dynamics.

Previous speech decoding studies with phoneme classification results employed electrode technologies ranging from standard ECoG[13–15,17–20] to penetrating intracortical arrays[16]. These studies involved subjects performing speech production tasks in an epilepsy monitoring unit, with total utterance durations (actual speaking time) of >~10 minutes. The corresponding neural recordings enabled speech decoding by classifying up to 39 English phonemes. With our initial experiments constrained to 15 minutes of intra-operative time, we developed a 9-way phoneme decoder and reported stronger accuracy metrics from intra-operative data with total spoken durations of ≤ 2.5 minutes (supplementary table 3). We observed robust decoding for classifying spoken phonemes across all subjects, with maximum accuracy reaching up to 57% (chance: 11.11%) in predicting both consonants and vowels, and with S1 and S2 achieving the highest decoding performances (**Fig. 4**). Subject S3 obtained lower decoding values likely due to poor contact with SMC (indicated by lower SNR) that could be due to issues related with tunneling of the electrode array through the burr hole. Subject S4 obtained decoding performance that was on par with S1 and S2. This subject however, performed only one block of our speech repetition task. On examining the impact of recording duration, we observed increased decoding performance with more trials/time for training, as expected (Supplementary Figure 5). Further, S4's decoding performance with a 256-channel array at 50 subsampled trials was higher than the other subjects with 128 channel arrays, indicating that improvements in decoding accuracy may be possible with increased channel counts. We also found

that spacing was inversely proportional to the number of unique phonemes, suggesting that with the benefits of μECoG recordings would be even greater for larger datasets with a greater number of phonemes (Supplementary Fig. 7C). We directly compared the decoding ability of μECoG electrodes to standard IEEG (ECoG/SEEG) electrodes located and showed that even when using short time periods, micro-scale neural recordings using high-density μECoG outperformed current standard technology. Future work will focus on using longer recordings, more naturalistic speech, and direct reconstruction of utterances from micro-scale neural recordings.

When we analyzed decoder performance, we found that the classification errors varied systematically as a function of the utterance. Phonemes with similar phonological features (e.g. /p/ and /b/) were more often confused compared to dissimilar phonemes that shared fewer features (/p/ and /g/)[14,16]. We quantified this phonological error using Hamming distance (bits) and showed that classifier errors were inversely proportional to phonological distance. Evaluating speech decoding models with phonological distance could lead to error correction that comparatively weighs similar phonemes with a lower cost versus phonemes with dissimilar features[50]. This phonological error metric could be used in addition to classification accuracy to optimize cost-functions for speech decoding algorithms.

We characterized the importance of spatial characteristics of micro-scale neural dynamics for speech decoding. We observed improved phoneme decoding performance that increased with spatial resolution, spatial coverage, and decreased contact size. There was up to 29% improvement in decoding performance as compared to simulated spatial resolutions of standard IEEG systems (4 – 10 mm; supplementary fig. 7). Higher spatial resolution was required to achieve better decoding as more unique phonemes were included in the analysis. On a subject-specific scale, we observed high-resolution speech articulator maps, analogous to previously established group-level organization[43]. We observed that the surface area for each articulator ranged from 11 mm^2 to 80 mm^2 depending on the location and coverage of each μECoG grid (Supplementary table 4). These subject-specific maps could be a strong feature for future neural prostheses for individual patients.

Speech production involves the temporal sequencing of articulatory movements[42,43], and consequently the neural properties of articulators should be sequentially encoded. With our micro-scale neural recordings, we demonstrated sequential dynamics of articulation by successfully decoding individual phonemes in time. This temporal decoding reflected the correct order of serial phoneme execution during the speech utterance. These decoding activation patterns overlapped in time for each successive phoneme which could be evidence for co-articulatory dynamics during the speech utterance[42,51]. We then leveraged this ability to decode the temporal dynamics of the speech utterance to decode the entire utterance by using a deep-neural-net based sequential model that was able to

decode phoneme sequences in the nonwords with only 2 minutes of training data (Supplementary table 3). We found a significant improvement when compared to our linear model that required explicit phoneme onset information. The sequential decoding performance was dependent on high resolution sampling, further highlighting the benefits of μECoG recordings. Decoding using deep-neural network requires a high number of parameters[16,17,20,52], and future work will study the relationship between high-fidelity cortical recordings and deep network complexity for successful speech decoding. These results are dependent on micro-neural activations without the addition of an explicit language model[18,21] to boost performance, thereby highlighting the role of the higher resolution of neural recordings from μECoG for arrays for entire speech utterance decoding without strong priors. Future work however, will explore this use of state-of-the-art, deep-learning-based, speech-reconstruction models with large vocabulary language models.

Our results highlighted the benefits of high-density, high-channel-count LCP-TF μECoG arrays for human speech decoding. Micro-scale recordings demonstrated more accurate speech decoding than standard IEEG electrodes and better resolved the rich spatio-temporal dynamics of the neural underpinnings of speech production. These findings motivate the use of high-density μECoG for brain-computer-interfaces to restore speech for patients suffering from debilitating motor disorders who have lost verbal communication abilities[53].

4 Methods

4.1 Participants

We recorded neural activity from speech motor cortex (SMC) of four awake patients with no speech-impairments using LCP-TF μECoG arrays, during intraoperative procedures at the Duke University Medical Center (DUMC - Supplementary table 1). Three of the four subjects (S1, S2, and S3) were being treated for movement disorders and the intraoperative recordings were performed during awake deep brain stimulator (DBS) implantation surgery for movement disorder treatment. The fourth subject (S4) underwent a brain tumor resection and neural data were recorded during the awake craniotomy prior to resection. As a control, we recorded from eleven patients with epilepsy (no speech-impairment) implanted with conventional electrodes (ECoG or SEEG) and recorded with a clinical IEEG system (Natus Quantum), during preoperative intracranial monitoring at DUMC. All subjects were fluent in English, and we obtained written informed consent in accordance with the Institutional Review Board at the DUMC.

4.2 Intra-operative μECoG array implantation

We performed intraoperative recordings using non-commercial, custom designed LCP-TF μECoG arrays that were fabricated by Dyconex Micro Systems Technology (Bassersdorf, Switzerland) with two

different designs. For the first design, we used an array with 128 microelectrodes (200 μm electrode diameter) in an 8 x 16 design with 1.33 mm center-to-center spacing (**Fig. 1A & 1B,** top).

The 128-channel array design is narrow and flexible enough to be implanted through the burr-hole opened during DBS implantation. Each subject's cortical anatomy was assessed using preoperative MR imaging with contrast, to ensure the absence of large cortical bridging veins. Following the burr hole craniotomy, the surgeon tunneled the electrode array in the subdural space using neuro-navigation (Stealth Station Surgical Navigation, Medtronic, Minnesota) and preoperatively determined the cortical target (SMC). The tail of the array was then reflected such that the burr hole remained open for the placement of the DBS electrodes. A clinically-indicated intraoperative CT was obtained to ensure accurate placement of DBS electrodes which was also used to localize the placement of the array. The DBS insertion canula was used as the ground/reference.

The second LCP-TF μECoG design had 256 microelectrodes (200 μm diameter) in a 12 x 24 array (1.72 mm spacing) and was used to target anatomy directly over eloquent cortex exposed during an awake craniotomy (**Fig. 1A & 1B,** bottom). Following clinically indicated cortical mapping, the array was placed over the functionally relevant cortex (SMC) and affixed to a modified surgical arm. A grounding/reference wire was connected to a surgical retractor attached to the subject's scalp. Finally, the four corners of the array were registered using Brainlab (Munich, Germany) neuronavigation, and were colocalized to the subject specific anatomical T1 structural image.

The arrays were connected to custom headstages with Intan Technologies RHD chips for amplification and digitization. These headstages were connected to an Intan Technologies RHD recording controller through micro- high-definition multimedia interface (μHDMI) cables passed outside of the sterile zone. The entire μECoG set up was sterilized using ethylene oxide gas at DUMC prior to surgical implantation.

In order to co-register the arrays between subjects, the array locations were first assessed using Brainlab neuronavigation software. The array templates (for both 128 & 256 channels) were then manually mapped to a Montreal Neurological Institute (MNI) average brain using key anatomical landmarks such as distance to the central sulcus and to the Sylvian fissure. **Fig. 1D** shows three templates of 128-channel grids (violet, green, and blue for S1, S2, and S3 respectively), and a 256-channel template (S4: purple) implanted over the SMC of an average MNI brain.

4.3 Neural Recording

Neural data from the μECoG arrays were recorded with an Intan RHD recording system (Intan Technologies, Inc.). The recordings were analog filtered between 0.1 Hz and 7500 Hz and digitized at 20,000 samples per second. The control subjects implanted with the standard IEEG system (ECoG or SEEG) during pre-operative epileptic monitoring. For ECoG monitoring, the subjects had Ad-Tech

electrode grid (48 or 64 channels; Ad-Tech Medical Instrument Corporation) with 10 mm inter-electrode spacing and 2.3 mm exposed diameter. The subjects with SEEG monitoring were implanted with either PMT (PMT Corporation) or Ad-Tech depth electrodes with 3.5 mm and 5 mm inter-electrode spacing, respectively (Supplementary Table 1). The intracranial IEEG data were recorded using a Natus Quantum LTM amplifier (Natus Medical, Inc.), with an analog filter between 0.01 Hz and 1000 Hz and digitized at 2,048 samples per second.

4.4 Speech task

Each subject performed a speech production task during their intra-operative procedure. In each trial, they listened to and repeated one of 52 nonwords that were designed to sample the phonotactics of American English (Supplementary Table 2). The nonwords contained an equal number of consonant-vowel-consonant (CVC: 26) and vowel-consonant-vowel (VCV: 26) tokens, which included a total of nine different phonemes (four vowels: /a/, /ae/, /i/, /u/ and five consonants: /b/, /p/, /v/, /g/, /k/). Each trial lasted between 3.5 – 4.5 s including the auditory stimulus (mean stimulus duration: 0.5 second) followed by a three second response window with a 250 ms jitter between trials. The trials were presented in three consecutive blocks with 52 unique nonword tokens shuffled for each block. The overall task time took approximately 15 minutes for each subject. The control subjects implanted with ECoG/SEEG during preoperative clinical monitoring in the epilepsy monitoring unit at the DUMC performed a similar speech production task, but some repeated different speech tokens. Equivalent phonemes and number of trials were used for the direct comparison between recording techniques.

4.5 Audio recording

Auditory stimuli were presented using a battery-powered stereo-speaker (Logitech) through a USB DAC and audio amplifier (Fiio). The beginning of each trial was marked by a flashing white circle on the stimulus laptop screen which was detected by a photodiode (Thorlabs, Inc.) attached to an auxiliary channel on the Intan recording system. The audio waveforms were recorded using a clip-on microphone (Movophoto, Inc.) which was connected to a pre-amplifier (Behringer), and digitized at 44,100 Hz by the Psychtoolbox module in MATLAB on the recording laptop. This microphone recording was also digitized on the Intan Technologies RHD recording system at 20,000 samples per second (along with the photodiode). The auditory stimulus, spoken nonword production response, and phoneme onsets were manually identified using the Audacity software package.

4.6 Neural data analysis and decoding models

For preprocessing, the neural recordings from μECoG electrodes were decimated to 2,000 samples per second using an anti-aliasing filter followed by common-average-referencing (CAR). The electrodes

with impedance greater than 1 MOhm or with recordings greater than 3 * log-RMS were excluded from CAR. Further, we removed trial epochs (after CAR) with excessive artifacts (greater than 10 RMS).

4.6.1 Spectrogram visualization

We computed spectrograms for each electrode to determine the presence of speech-neural activity. The spectral information was extracted within a 500 ms analysis window (50 ms step-size) with 10 Hz frequency smoothing using a multi-taper time-frequency analysis method[8]. The resultant spectrograms were normalized with respect to a baseline period (500 ms window pre-stimulus across trials) by dividing the resultant spectrogram by its average baseline power-spectral-density to remove *1/f* broadband activity. The normalized spectrograms were then averaged across trials and the log-power was computed around the speech utterance (-500 ms to 500 ms) to quantify neural activations in decibels.

$$HG - ESNR\ (dB) = 20 * \log_{10}\left(\sum_{t=-0.5}^{0.5}\sum_{f=70}^{150}\frac{Spect(t,f)}{\frac{1}{\tau}\sum_{\tau}Spect_{baseline}(\tau,f)}\right) \quad (2)$$

4.6.2 Neural feature extraction

Neural high gamma signals (HG: 70 – 150 Hz) have been previously shown to characterize localized neural activity from cortical regions and are highly correlated with multi-unit firing[22,23]. They therefore serve as a robust marker for localizable neural activity. The pre-processed neural signals from each electrode were band-pass filtered into 8 center frequencies (logarithmically arranged) between 70 and 150 Hz[54]. The envelope of each frequency band was extracted using the Hilbert transform and was averaged across bands. The resultant envelope was then downsampled to 200 Hz using an anti-aliasing filter and mean-normalized with respect to the baseline period (500 ms window pre-stimulus across trials). For each trial, we used a one second window of normalized HG centered at phonemic onset (-500 ms to 500 ms) to represent speech neural activity.

4.6.3 Cortical state space analysis using SVD-tSNE

We used a singular value decomposition (SVD) to transform normalized HG activity into a low-dimensional sub-space[40,55]. SVD was carried out on the covariance matrix of the HG activity to decompose highly covarying space-time activations across the array.

$$H: U\ X\ ET \quad (3)$$

U – number of utterance trials, E – number of significant HG μECoG electrodes, T – time duration of HG.

$$E[H^T H] = Q\lambda Q^T \quad (4)$$

The eigen-vectors (**Q**) served as basis for spatio-temporal activations with each trial receiving principal component (PC) scores corresponding to the eigen-vectors. To control for HG variability across subjects, we selected the set of ordinal eigen-vectors that cumulatively explained 80% of neural variance. The resultant PC scores were further projected to a two-dimensional subspace using t-distributed Stochastic Neighbor Embedding (t-SNE; Barnes-Hut algorithm, perplexity = 30)[44], a non-linear affine transformation technique for visualizing high-dimensional data. Finally, the trial embeddings were coded to represent either phoneme or articulatory features.

To evaluate the clustering of trials with respect to either articulatory features or phonemes, we performed a silhouette analysis (Euclidean distance)[56] for each state-space trial that measured the closeness of trial to a given cluster with respect to the other clusters. We further determined two metrics to evaluate the clustering strength of the cortical state-space: 1) silhouette ratio: ratio of the number of trials with positive silhouette score to the total number of trials, and 2) silhouette score: the average silhouette values of positive silhouette trials. Finally, to motivate the benefit of high-resolution neural recordings, we repeated the cortical state-space analysis on spatially sub-sampled electrodes extracted using Poisson disc sampling[31], by spatially constraining the minimum distance $\left(r = 2 * \sqrt{\frac{(X*Y)}{n*3.5}}\right)$ between subsampled electrodes (n)

4.6.4 Linear decoding model using SVD-LDA

We constructed a linear-classification model with 20-fold cross validation to decode phonemes from normalized HG neural activations. First, we used SVD to compress the HG matrix (training set) to a low-dimensional subspace as explained in the previous section (Equations 3 & 4). Then, we utilized the resultant principal component scores to classify phonemes using a supervised linear-discriminant (LDA) model. The optimal number of eigenvectors (or principal components) was identified using a nested cross-validation procedure, where a separate 20-fold cross-validation with grid-search was conducted on the training set. The nested cross-validation was conducted on every training fold. The overall decoding accuracy was calculated by calculating the proportion of correctly predicted labels in the test set. We trained decoding models to decode phonemes for each of the three positions in the nonword (P1, P2, and P3) and by collapsing phonemes across all positions.

To characterize decoding error, we examined values in the confusion matrix to test for confusion between phonemes that share similar phonological features. To characterize each phoneme, we created a 17-bit categorical vector based on work by Chomsky and Halle to represent phonological

features[46]. We then computed the Hamming distance (bits) for each phoneme pair to determine the pairwise phonological distance. Phonemes with similar phonological features (e.g., /p/ vs. /b/) differed by 1-bit, whereas phonemes with dissimilar features differed by a maximum of 9 bits (e.g., /p/ vs. /i/). Finally, we mapped the percentage misclassification onto phoneme distance to evaluate phonological confusion of the neural signals and our classifier. The decoding error in terms of phonological distance was calculated using the equation:

$$e = \frac{1}{9}\sum_{i,j}\frac{1}{2}C_{i,j}D_{i,j} \tag{5}$$

e is phoneme error (bits), C is normalized confusion matrix, D is phonological distance (bits).

In addition, we used the decoding strategy to infer the spatial and sequential representation of phonological features. To infer the spatial encoding of features, we first grouped nine phonemes into four articulator features (low: /a/, /ae/, high: /i/, /u/, labials: /b/, /p/, /v/, and dorsal: /g/, /k/). We then trained an SVD-LDA decoder model (four-way) on each electrode to decode these articulatory features. The Area Under the Receiver Operating Characteristic Curve (ROC-AUC or AUC) was calculated for each electrode to quantify the specificity of articulatory feature encoding.

To examine the sequential representation of phonemes, we used temporal decoder models to train and test at every time-segment t, with respect to the speech utterance[57]. With this temporal modeling strategy, each SVD-LDA model was trained on sequential time-segments with respect to utterance onset, and the accuracy values were evaluated on the same time-segments. Through this strategy, we could assess the presence of temporal phoneme information present in the raw time signal. In both training and testing, the time-segments consisted of 200 ms windows (10 ms step size) ranging from -1000 ms to 1000 ms aligned to the utterance. The resultant time-series of accuracy values was used to determine the temporal activation of phoneme representations during speech production.

4.6.5 RNN based seq2seq model for phoneme sequence decoding

We constructed a seq2seq encoder-decoder network to simultaneously decode all phonemes in a nonword from HG activations[20]. The neural-network contained three layers to translate HG activity to phoneme sequences: 1) Temporal convolutional layer, 2) Encoder recurrent neural network (RNN), and 3) Decoder RNN.

1) *Temporal convolutional layer*: The first layer contained a set of sequential kernels to transform HG into downsampled features. This convolutional layer to model temporal features was essential to reduce memory loading on the encoder RNN. We fixed both the kernel width and the hop length to

50 ms and no non-linearity was applied to the convolution kernels. The convolution layer resulted in linear temporal decomposition of HG.

2) *Encoder RNN*: The second part of the network contained a single layer of RNN units to model the output of convolutional layer at each time step. We used bi-directional LSTM (Long-Short-Term-Memory)[58] cells to represent each RNN unit, and the memory states of both forward (right arrows) and backward (left arrows) layers were averaged to return the final memory state as the output.
3) *Decoder RNN*: The final layer contained a sequence of a unidirectional RNN to predict one-hot encoded phoneme units at each position. The decoder RNN was instantiated by the final output state of the encoder RNN and was configured to return both the memory state and output sequences. The output values from each time-bin were then converted to phoneme probabilities through a dense neural net with SoftMax activation.

The architecture was implemented in Keras-Tensorflow[59]. The network was trained using the Adam[47] gradient descent optimization (learning rate = 1e-3; 800 epochs) by minimizing the categorical cross-entropy between true and predicted phoneme probabilities[47]. We included an L2 regularization term at both convolutional and recurrent layers to prevent overfitting. Overall, the architecture contained the following hyperparameters: 1) number of convolutional kernels, 2) number of RNN units, and 3) L2 penalty. For each subject, we held 20% of trials for testing and we performed training and hyperparameter optimization on the remaining 80% of trials.

During training (80% trials), the decoder RNN evaluated the phoneme probability at every position and the RNN input was set to one-hot encoded vector of phonemes at the previous step. For faster convergence and stable modeling, we utilized the “teacher – forcing” approach where we used the ground truth (true phonemes) as input to the decoder RNN[60].

We identified the optimal hyperparameters on the training-set (80% trials) for the neural-network using the keras-tuner platform[61]. For each subject, we randomly explored the hyperparameter space through 200 hyperparameter combinations. We then calculated the validation accuracy (10-fold cross-validation, three repetitions) for each combination, and identified the hyperparameter set with the maximum-validation accuracy. The hyperparameters for each subject are summarized in supplementary table 5.

During testing (20% trials), we used the trained decoder model (using optimal hyperparameters) to generate the most probable phoneme sequence by identifying the most likely phoneme at each position and used it as the input to the next position.

Overall, we evaluated the model by calculating its phoneme decoding accuracy between the tested and predicted phonemes. To validate the seq2seq model against the linear approach, we trained an SVD-

LDA model (5-fold cross-validation; analogous to 20% held-out) and calculated the accuracy and phoneme error in decoding phonemes across all positions. Further, to quantify the influence of spatial resolution, we evaluated the decoding performance at different spatial resolutions through subsampling and training the encoder-decoder models on subsampled versions of the μECoG recordings.

4.6.6 Statistical analysis

We used multiple statistical methods to validate μECoG recordings for speech decoding. To identify electrodes with significant speech-neural activation, we compared HG power between the response and the baseline period, using a 10,000-iteration one-sided permutation test (corrected for false discovery rate). For the control subjects, we performed similar non-parametric test on SEEG/ECoG electrodes anatomically localized to SMC to identify significant speech-neural activations. For decoding comparisons to standard IEEG, we used a Mann-Whitney U test to compare the HG power (during the speech production response) between the population of μECoG electrodes and the control IEEG. Finally, we used one-way ANOVA to examine the effect of spatial subsampling, as well as post hoc *t-tests* to identify subsampling groups with significant differences in means.

To quantify signal sharing between μECoG electrodes, a cross-channel Pearson correlation coefficient (ρ) was calculated between HG time-series (during the speech production response) and was linked to corresponding Euclidean distance (x). The mean correlation coefficient was identified at each unique spatial distance to assess the spatial relationship of HG between electrode pairs.

For phoneme decoding, the prediction accuracies were tested against chance using a binomial cumulative distribution test[45]. All population data are indicated with mean and standard errors, and we estimated the significance of model fits using linear regression.

Grant Support: J.V., G.C., S.L, D.S., A.F., S.S., S.D. were supported by NIH R01DC019498 and DoD W81XWH-21-0538, J.V., G.C., C.W., D.S., S.S., S.D. were supported by NIH CTSA UL1TR002553, and G.C., J.V., D.S, S.L, S.S, S.D. were supported by an internal Duke Institute for Brain Sciences Incubator Award. D.S. was supported by the NINDS Neurosurgery Career Development Program and the Klingenstein-Simons Foundation.

Acknowledgements: We would like to thank Anna Thirakul for help with consenting and IRB compliance, Palee Abeyta for help with data collection for the in-unit epilepsy patients, Seth Foster for helping with visualization software, as well as Stephen Harward and James Carter for help with intraoperative data collection. We would also like to thank the patients who graciously volunteered their time for this study.

References

1. Mitchell, J. D. & Borasio, G. D. Amyotrophic lateral sclerosis. *Lancet (London, England)* **369**, 2031–41 (2007).

2. Doble, J. E., Haig, A. J., Anderson, C. & Katz, R. T. Impairment, activity, participation, life satisfaction, and survival in persons with locked-in syndrome for over a decade: follow-up on a previously reported cohort. *J. Head Trauma Rehabil.* **18**, 435–44 (2003).

3. Rousseau, M.-C., Pietra, S., Nadji, M. & Billette de Villemeur, T. Evaluation of quality of life in complete locked-in syndrome patients. *J. Palliat. Med.* **16**, 1455–8 (2013).

4. Lugo, Z. R. *et al.* Beyond the gaze: Communicating in chronic locked-in syndrome. *Brain Inj.* **29**, 1056–1061 (2015).

5. Sellers, E. W., Kübler, A. & Donchin, E. Brain-computer interface research at the University of South Florida cognitive psychophysiology laboratory: The P300 speller. *IEEE Trans. Neural Syst. Rehabil. Eng.* **14**, 221–224 (2006).

6. Mugler, E. M. *et al.* Differential Representation of Articulatory Gestures and Phonemes in Motor, Premotor, and Inferior Frontal Cortices. *J. Neurosci.* **4653**, 1–23 (2017).

7. Jiang, W., Pailla, T., Dichter, B., Chang, E. F. & Gilja, V. Decoding speech using the timing of neural signal modulation. *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS* **2016-Octob**, 1532–1535 (2016).

8. Cogan, G. B. *et al.* Manipulating stored phonological input during verbal working memory. *Nat. Neurosci.* **20**, 279–286 (2017).

9. Cogan, G. B. *et al.* Sensory-motor transformations for speech occur bilaterally. *Nature* **507**, 94–8 (2014).

10. Anumanchipalli, G. K., Chartier, J. & Chang, E. F. Intelligible speech synthesis from neural decoding of spoken sentences. *Nature* **568**, 1–32 (2019).

11. Angrick, M. *et al.* Speech synthesis from ECoG using densely connected 3D convolutionalneural networks. *J. Neural Eng.* (2019).

12. Akbari, H., Khalighinejad, B., Herrero, J., Mehta, A. & Mesgarani, N. Reconstructing intelligible speech from the human auditory cortex. *Sci. Rep.* 350124 (2019) doi:10.1101/350124.

13. Kellis, S. *et al.* Decoding spoken words using local field potentials recorded from the cortical surface. *J. Neural Eng.* **7**, 056007 (2010).

14. Mugler, E. M. *et al.* Direct classification of all American English phonemes using signals from functional speech motor cortex. *J. Neural Eng.* **11**, 035015 (2014).

15. Ramsey, N. F. *et al.* Decoding spoken phonemes from sensorimotor cortex with high-density ECoG grids. *Neuroimage* **180**, 301–311 (2018).

16. Wilson, G. H. *et al.* Decoding spoken English from intracortical electrode arrays in dorsal precentral gyrus. *J. Neural Eng.* **17**, 66007 (2020).

17. Livezey, J. A., Bouchard, K. E. & Chang, E. F. Deep learning as a tool for neural data analysis: Speech classification and cross-frequency coupling in human sensorimotor cortex. *PLOS Comput. Biol.* **15**, e1007091 (2019).

18. Herff, C. *et al.* Brain-to-text: decoding spoken phrases from phone representations in the brain.

Front. Neurosci. **9**, 217 (2015).

19. Moses, D. A., Leonard, M. K., Makin, J. G. & Chang, E. F. Real-time decoding of question-and-answer speech dialogue using human cortical activity. *Nat. Commun.* **10**, (2019).

20. Makin, J. G., Moses, D. A. & Chang, E. F. Machine translation of cortical activity to text with an encoder–decoder framework. *Nat. Neurosci.* **23**, 575–582 (2020).

21. Moses, D. A. *et al.* Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *N. Engl. J. Med.* **385**, 217–227 (2021).

22. Ray, S., Crone, N. E., Niebur, E., Franaszczuk, P. J. & Hsiao, S. S. Neural correlates of high-gamma oscillations (60-200 Hz) in macaque local field potentials and their potential implications in electrocorticography. *J. Neurosci.* **28**, 11526–11536 (2008).

23. Dubey, A. & Ray, S. Cortical Electrocorticogram (ECoG) is a local signal. *J. Neurosci.* (2019).

24. Duraivel, S. *et al.* Comparison of signal decomposition techniques for analysis of human cortical signals. *J. Neural Eng.* **17**, (2020).

25. Nason, S. R. *et al.* A low-power band of neuronal spiking activity dominated by local single units improves the performance of brain–machine interfaces. *Nat. Biomed. Eng.* **4**, 973–983 (2020).

26. Muller, L., Hamilton, L. S., Edwards, E., Bouchard, K. E. & Chang, E. F. Spatial resolution dependence on spectral frequency in human speech cortex electrocorticography. *J. Neural Eng.* **13**, (2016).

27. Chang, E. F. Towards large-scale, human-based, mesoscopic neurotechnologies. *Neuron* **86**, 68–78 (2015).

28. Cheung, C., Hamilton, L. S., Johnson, K. & Chang, E. F. The auditory representation of speech sounds in human motor cortex. *Elife* **5**, e12577 (2016).

29. Viventi, J. *et al.* Flexible, foldable, actively multiplexed, high-density electrode array for mapping brain activity in vivo. *Nat. Neurosci.* **14**, 1599–605 (2011).

30. Woods, V. *et al.* Long-term recording reliability of liquid crystal polymer µECoG arrays. *J. Neural Eng.* (2018) doi:10.1088/1741-2552/aae39d.

31. Chiang, C.-H. *et al.* Development of a neural interface for high-definition, long-term recording in rodents and nonhuman primates. *Sci. Transl. Med.* **12**, eaay4682 (2020).

32. Chiang, C. H. *et al.* A modular high-density µeCoG system on macaque vlPFC for auditory cognitive decoding. *J. Neural Eng.* **17**, 046008 (2020).

33. Stead, M. *et al.* Microseizures and the spatiotemporal scales of human partial epilepsy. *Brain* **133**, 2789–97 (2010).

34. Schevon, C. A. *et al.* Evidence of an inhibitory restraint of seizure activity in humans. *Nat. Commun.* **3**, 1060 (2012).

35. Blanco, J. A. *et al.* Data mining neocortical high-frequency oscillations in epilepsy and controls. *Brain* **134**, 2948–59 (2011).

36. Schevon, C. A. *et al.* Microphysiology of epileptiform activity in human neocortex. *J. Clin. Neurophysiol.* **25**, 321–30 (2008).

37. Wetjen, N. M. *et al.* Intracranial electroencephalography seizure onset patterns and surgical outcomes in nonlesional extratemporal epilepsy. *J. Neurosurg.* **110**, 1147–52 (2009).

38. Wang, W. *et al.* Human motor cortical activity recorded with micro-ECoG electrodes during individual finger movements. in *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009* 586–589 (IEEE, 2009). doi:10.1109/IEMBS.2009.5333704.

39. Mehring, C. *et al.* Inference of hand movements from local field potentials in monkey motor cortex. *Nat. Neurosci.* **6**, 1253–4 (2003).

40. Chiang, C.-H. *et al.* Flexible, high-resolution thin-film electrodes for human and animal neural research. *J. Neural Eng.* **18**, (2021).

41. Trumpis, M. *et al.* Sufficient sampling for kriging prediction of cortical potential in rat, monkey, and human $μ$ECoG. *Submiss.* (2020).

42. Bouchard, K. E., Mesgarani, N., Johnson, K. & Chang, E. F. Functional organization of human sensorimotor cortex for speech articulation. *Nature* **495**, 327–32 (2013).

43. Chartier, J., Anumanchipalli, G. K., Johnson, K. & Chang, E. F. Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex. *Neuron* **98**, 1042–1054 (2018).

44. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, (2008).

45. Combrisson, E. & Jerbi, K. Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *J. Neurosci. Methods* **250**, 126–136 (2015).

46. Chomsky, N. & Halle, M. The sound pattern of English. (1968).

47. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv Prepr. arXiv1412.6980* (2014).

48. Schevon, C. A. *et al.* Spatial characterization of interictal high frequency oscillations in epileptic neocortex. *Brain* **132**, 3047–3059 (2009).

49. Sun, J. *et al.* Intraoperative microseizure detection using a high-density micro-electrocorticography electrode array. *medRxiv* 2021.09.13.21263449 (2021) doi:10.1101/2021.09.13.21263449.

50. Williamson, J. H., Quek, M., Popescu, I., Ramsay, A. & Murray-Smith, R. Efficient human-machine control with asymmetric marginal reliability input devices. *PLoS One* **15**, e0233603 (2020).

51. Conant, D. F., Bouchard, K. E., Leonard, M. K. & Chang, E. F. Human sensorimotor cortex control of directly-measured vocal tract movements during vowel production. *J. Neurosci.* **38**, 2382–17 (2018).

52. Sun, P., Anumanchipalli, G. K. & Chang, E. F. Brain2Char: a deep architecture for decoding text from brain recordings. *J. Neural Eng.* **17**, 66015 (2020).

53. Marchetti, M. & Priftis, K. Brain–computer interfaces in amyotrophic lateral sclerosis: A metanalysis. *Clin. Neurophysiol.* **126**, 1255–1263 (2015).

54. Mesgarani, N., Cheung, C., Johnson, K. & Chang, E. F. Phonetic feature encoding in human superior temporal gyrus. *Science* **343**, 1006–10 (2014).

55. Insanally, M. *et al.* A low-cost, multiplexed μ ECoG system for high-density recordings in freely moving rodents. *J. Neural Eng.* **13**, 026030 (2016).

56. Kaufman, L. & Rousseeuw, P. J. *Finding groups in data: an introduction to cluster analysis*. (John

Wiley & Sons, 2009).

57. Gwilliams, L., King, J.-R., Marantz, A. & Poeppel, D. Neural dynamics of phoneme sequencing in real speech jointly encode order and invariant content. *BioRxiv* (2020).

58. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **9**, 1735–1780 (1997).

59. Abadi, M. *et al.* Tensorflow: A system for large-scale machine learning. in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)* 265–283 (2016).

60. Lamb, A. M. *et al.* Professor forcing: A new algorithm for training recurrent networks. in *Advances in neural information processing systems* 4601–4609 (2016).

61. O'Malley, T. *et al.* Keras Tuner. *Github.[(accessed 31 January 2021)]* (2019).

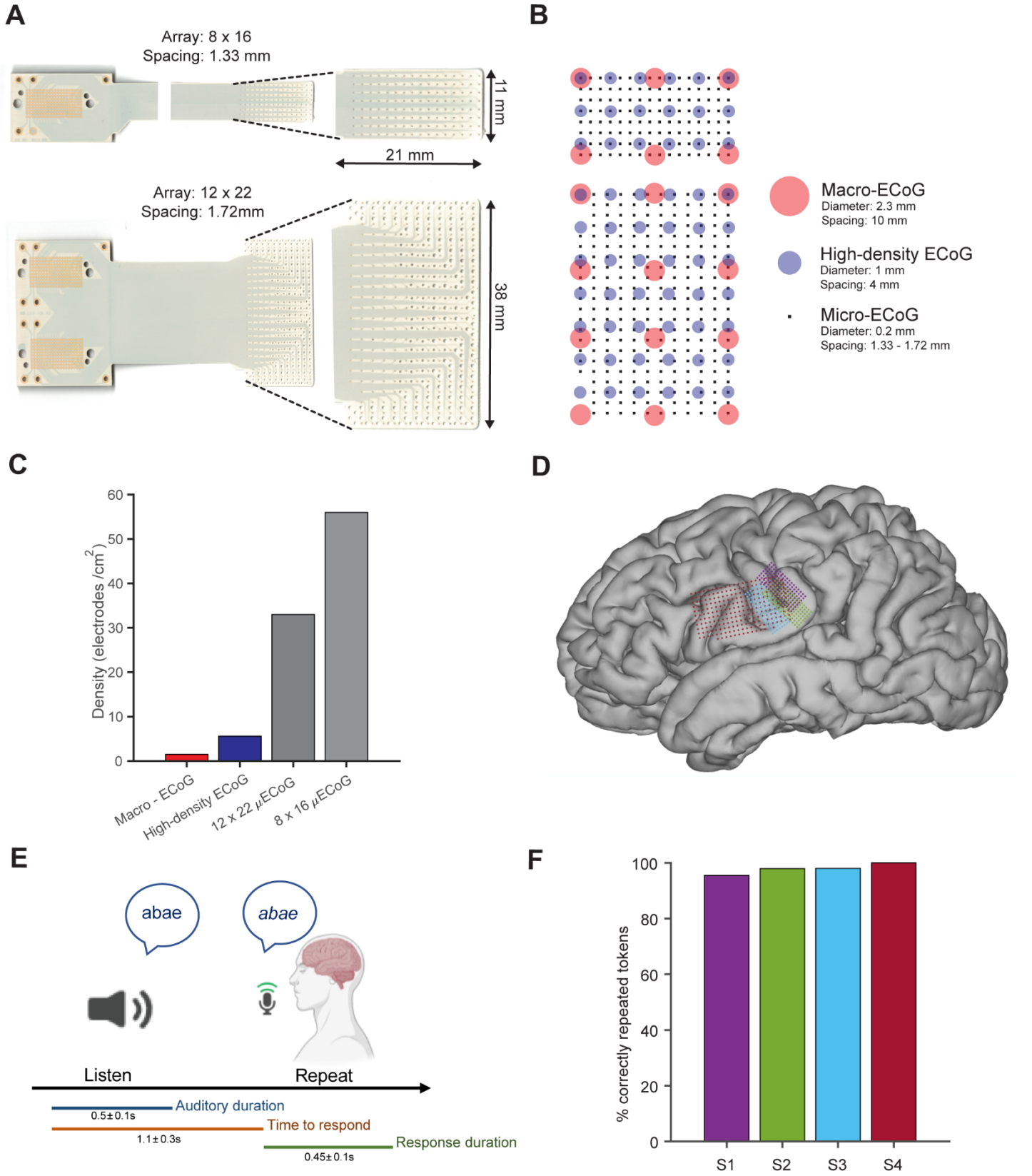

Figure 1. **Recording from high-density micro-electrocorticographic (μECoG) electrodes on the human brain during a speech production task. A)** Flexible liquid crystal polymer (LCP) electrodes with 128 (top) and 256 (bottom) recording electrodes (200-um diameter). **B)** Visual comparison of high-density micro-electrodes (black square markers) with macro-ECoG (red circles) and high-density ECoG (blue circles). **C)** μECoG electrodes exhibit higher spatial density (34 – 57 electrodes/cm^2) compared to the existing macro-ECoG (1 electrode/cm^2) and high-density ECoG arrays (6 electrodes/cm^2). **D)** Electrode arrays were implanted over speech motor cortex (SMC) in four awake patients (projected onto an average MNI brain). Electrode arrays had either 128 channels with 1.33 mm center to center spacing (pitch) for S1 (violet), S2 (green), and S3 (blue), or 256 channels with 1.72 mm pitch for S4 (red). **E)** A schematic of the intra-operative speech production task. Color bars indicate the duration of the auditory stimulus (blue), time-to-response (orange), and response duration (green). **F)** Subjects performed the speech production task with behavioral results above 95% accuracy for correctly repeated nonwords.

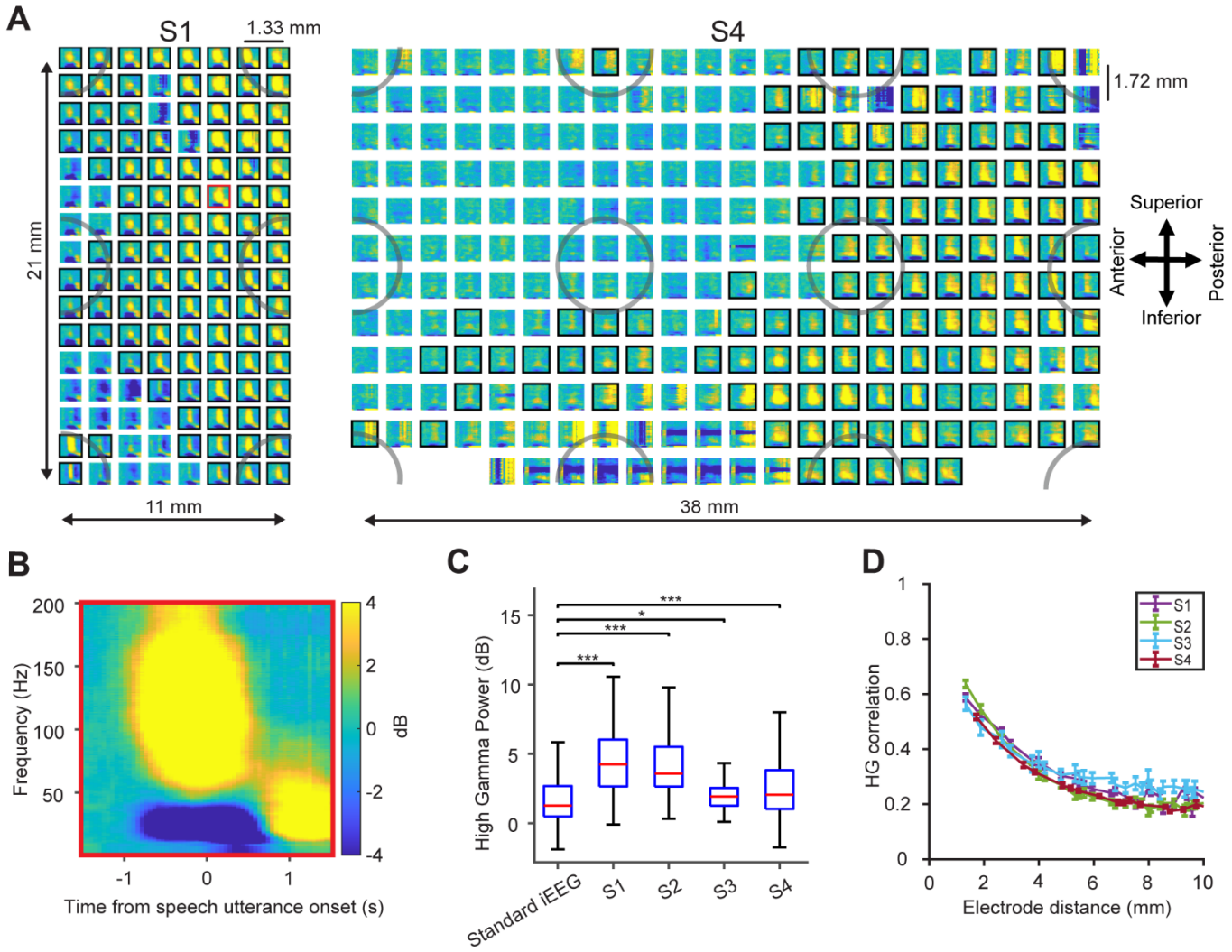

Figure 2. Highly resolved neural features as demonstrated by high-density μECoG during speech production. **A)** Spectrograms for each electrode from each array for two example subjects: S1, 128 electrodes (1.33 mm spacing) & S4, 256 electrodes (1.72 mm spacing). Data for all subjects are available in the supplemental materials. Electrodes with significant high-gamma modulation (HG: 70 – 150 Hz, FDR-corrected non-parametric permutation tests) during the speech utterance are indicated by black borders. Shaded grey circles represent superimposed simulated macro-ECoG electrode resolution with 10 mm interelectrode distance, illustrating the increased density of μECoG as compared to standard IEEG. **B)** A spectrogram of an example electrode from the 128-channel array (red border) demonstrated increase power in HG band which was time-locked to the speech utterance. **C)** μECoG arrays captured HG activity at significantly higher power than standard IEEG (ECoG and SEEG), $*p < 0.05$, $***p < 0.001$, Mann-Whitney U test, showing the increased ability of micro-electrodes to measure neural signals with a higher signal-to-noise ratio as compared to standard methods. The red lines and the blue boxes indicate the median and 25/75th percentile and dotted lines represent the full-range of the distribution. **D)** Inter-electrode HG correlation decreased with increased electrode spacing for all subjects. The correlation values at each electrode spacing are represented by mean and standard error. The correlation values at spatial resolutions less than 2 mm (r = 0.6), reveals evidence for spatially discriminative neural signals at the micro-scale during speech onset.

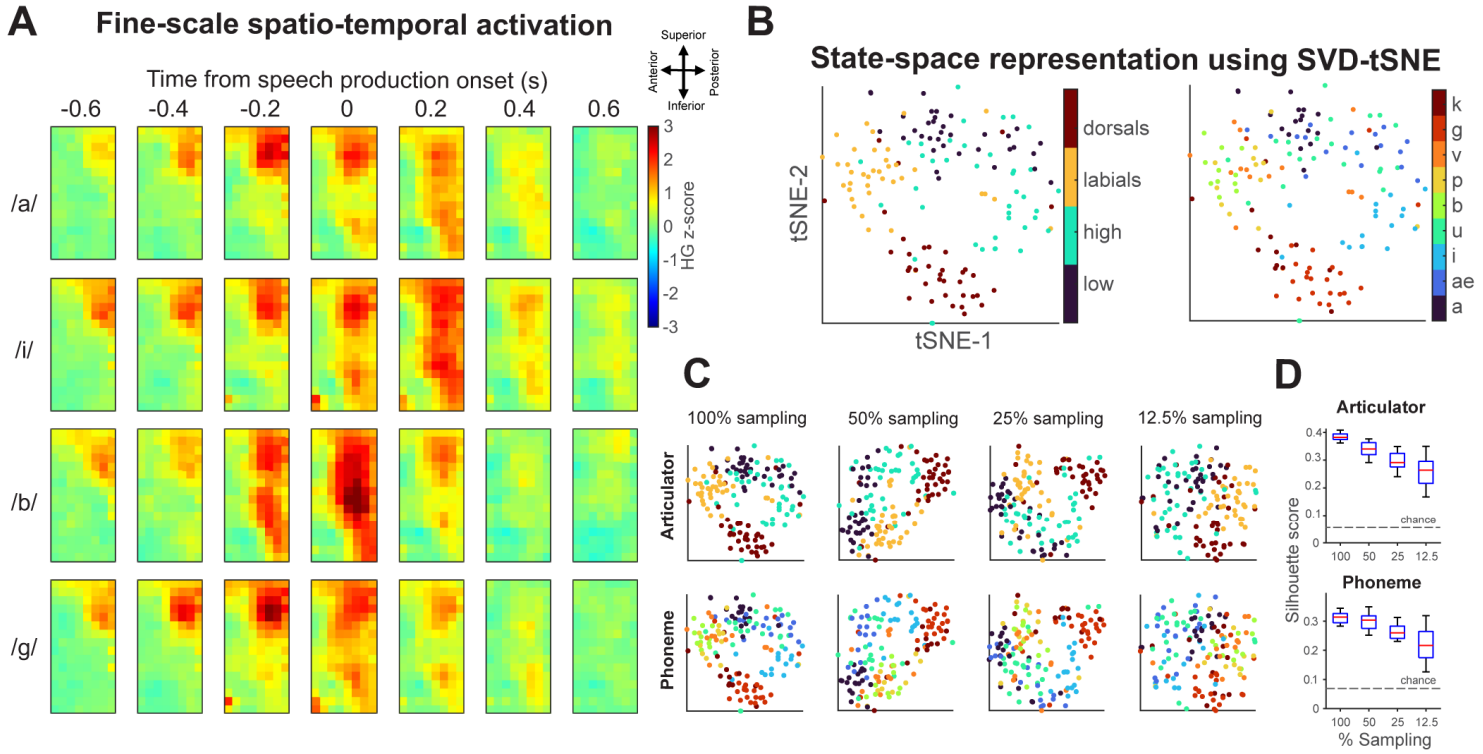

Figure 3: µECoG captures fine-scale spatio-temporal activity during speech production. A) HG activations showed fine-scale spatio-temporal patterns in SMC, shown here for an example subject (S1) during speech production. Spatial maps of HG responses showed discriminative activation patterns for each phoneme. **B)** State-space analysis using a signal value decomposition t-distributed stochastic neighbor embedding method (SVD-tSNE) depicted clear hierarchical separation of trials with respect to articulators (dorsal - /g, k/, labial (/b, p, v/), high (/i, u/), and low (/ae, a/) as well as individual phonemes. Phonemes using similar articulators were grouped together in three-dimensional state-space suggesting similarity in motor activation patterns. **C)** Electrode sub-sampling to simulate lower resolution recordings reduced the separation of clusters for both phoneme (top) and articulator (bottom) groupings in SVD-tSNE state-space (same color palette as in **B**). This reduction of grouping demonstrates the utility of high-resolution spatial sampling. **D)** To quantify this reduction, we performed a silhouette analysis that measures the relative closeness of a trial to articulator/phoneme group with respect to other groups. To show this significant decrease in cluster separation/groupings empirically, we performed a one-way ANOVA and post hoc t-tests (see main text). Each box plot depicts the distribution of silhouette scores obtained from electrode sub-samplings (50 samples using Poisson disc sampling). The red lines and the blue boxes indicate the median and 25/75th percentile and dashed lines represent the full-range of the distribution. The dotted horizonal line indicate chance silhouette metric obtained by shuffling the articulator/phoneme labels (1000 iterations). These analyses demonstrate the improved ability of high-resolution neural sampling to distinguish between articulator/phoneme motor features.

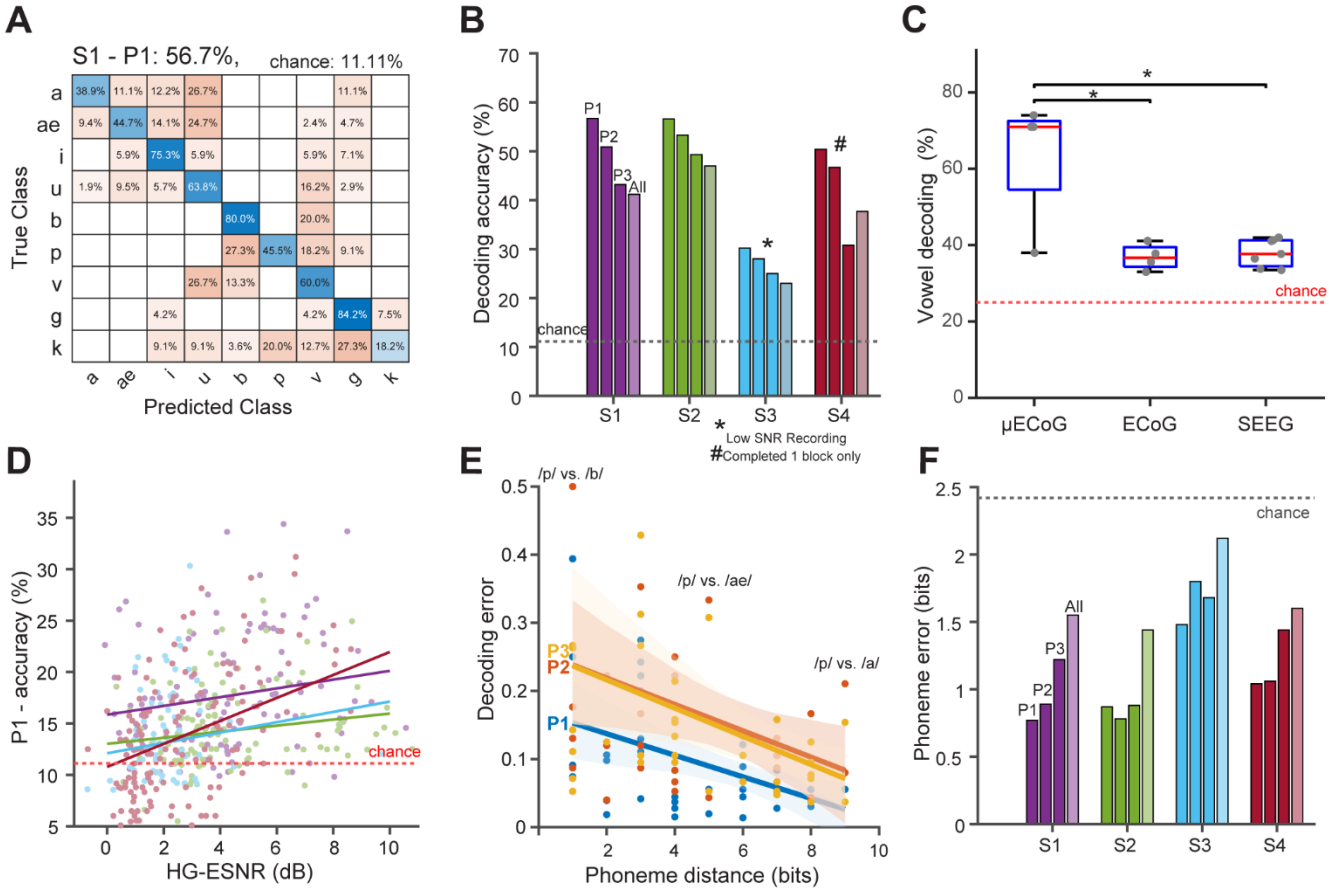

Figure 4. **High accuracy phoneme decoding from μECoG. A)** A confusion matrix for an example subject (S1) for phoneme position one (P1) using singular value decomposition linear discriminant analysis (SVD-LDA) decoding demonstrated high accuracy for each phoneme. **B)** Decoding was also highly accurate for each position within the nonword (P1, P2, P3), as well as for all phonemes combined across position (mean P1 = 49.9%, P2 = 45%, P3 = 38.8%, All = 36.6%, and chance = 11.11%). **C)** This accurate decoding outperformed standard IEEG in a 4-way vowel decoder (P1: μECoG: 71%, vs. ECoG: 36.7%, $p = 0.015$), vs. SEEG: 36.5%, $p = 0.011$; chance = 25%, , one-sided permutation test with 10,000 iterations). This nearly doubled performance increase further demonstrates the benefit of higher resolution neural recordings for speech decoding. **D)** Distribution of single-electrode decoding values were correlated with HG-ESNR for all subjects. Linear model fits indicated moderate but significant correlations for S1 (purple - Pearson r = 0.23, $p < 0.01$), S2 (green - r = 0.23, $p < 0.01$), and S3 (blue - r = 0.19, $p < 0.05$), and a stronger correlation for S4 (red - r = 0.44, $p < 0.001$). **E)** Decoding errors decreased with phoneme distance for all positions indicating higher confusion between phonemes with shared articulatory features (see methods and supplement for how error/phoneme distance was computed). Shaded regions indicate the 95% confidence interval of the linear model fit. **F)** Phoneme errors at all three positions in the nonwords and all phonemes combined across position show decoding errors (bits) well below chance (chance = 2.4 bits for a uniform confusion matrix).

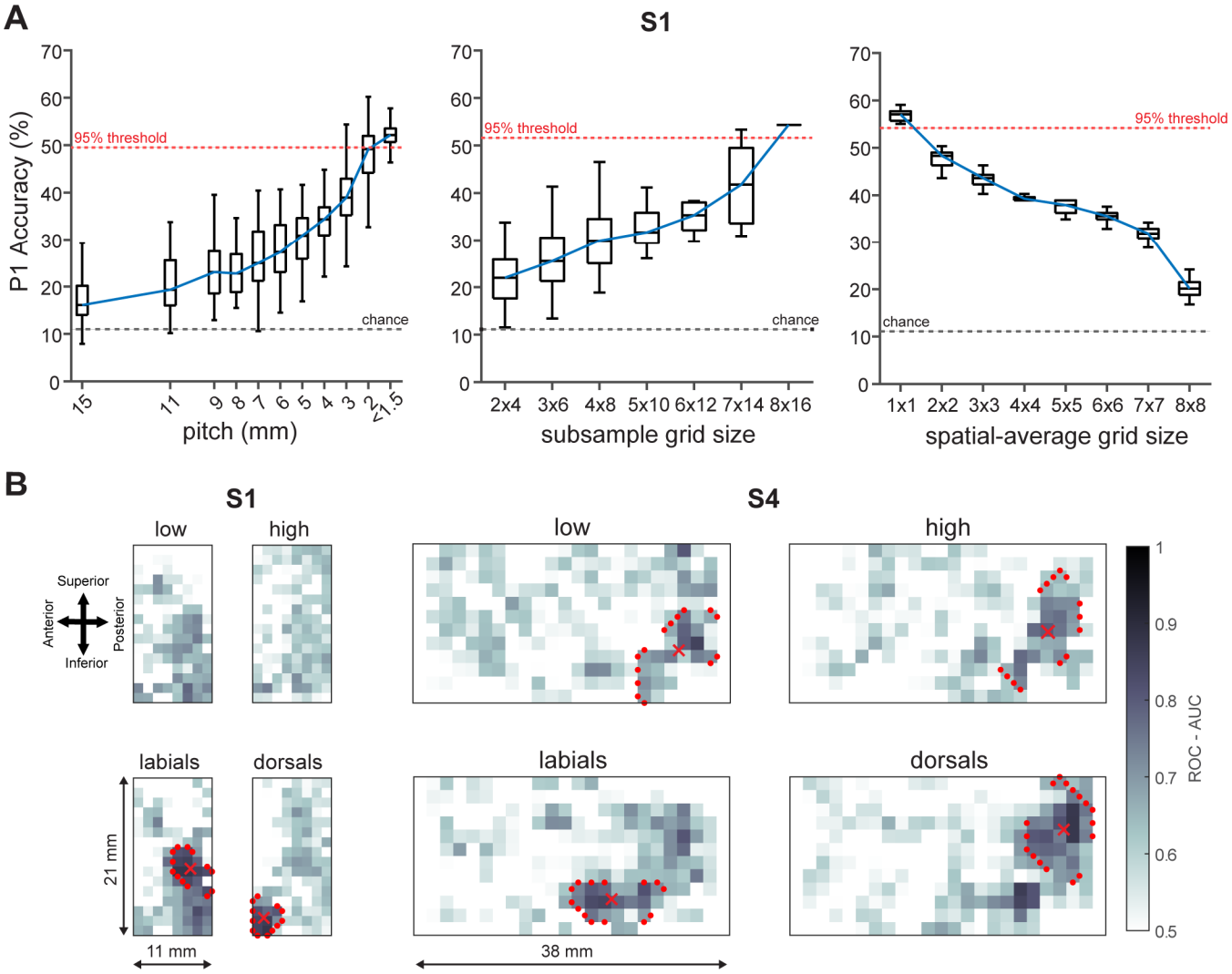

Figure 5. **Accurate decoding is dependent on the high-spatial resolution, high-channel count capability, and micro-scale recording of µECoG. A)** Accuracy of P1 decoding increased as a function of spatial resolution and coverage as demonstrated by sub-sampling electrodes, and decreased with contact size as demonstrated by spatial-averaging. Each box plot shows the distribution of accuracy metrics from the subsampled electrodes. Blue lines connect the median accuracy values from the box-plots. Electrode pitch less than 2mm along with micro-scale sampling was required to obtain decoding accuracy within 95% of the maximum decoding for the entire array (red dotted line) and demonstrated 29% improvement from the 10 mm pitch simulation and a 17% improvement from the 4 mm pitch simulation. **B)** Classification for articulators was spatially dependent, resulting in millimeter-scale spatially-specific articulator maps shown here in example subjects S1 and S4. Darker colors show higher area under the curve for receiver operating characteristic curves (ROC-AUC) for the univariate linear decoding values for each articulatory feature. Performance for vowels (top) was spatially diffuse, whereas performance for consonant (bottom) formed spatially distinct clusters. The contour points identify 90th percentile of AUC values across all articulators, and the marker indicates corresponding centroids for each articulator.

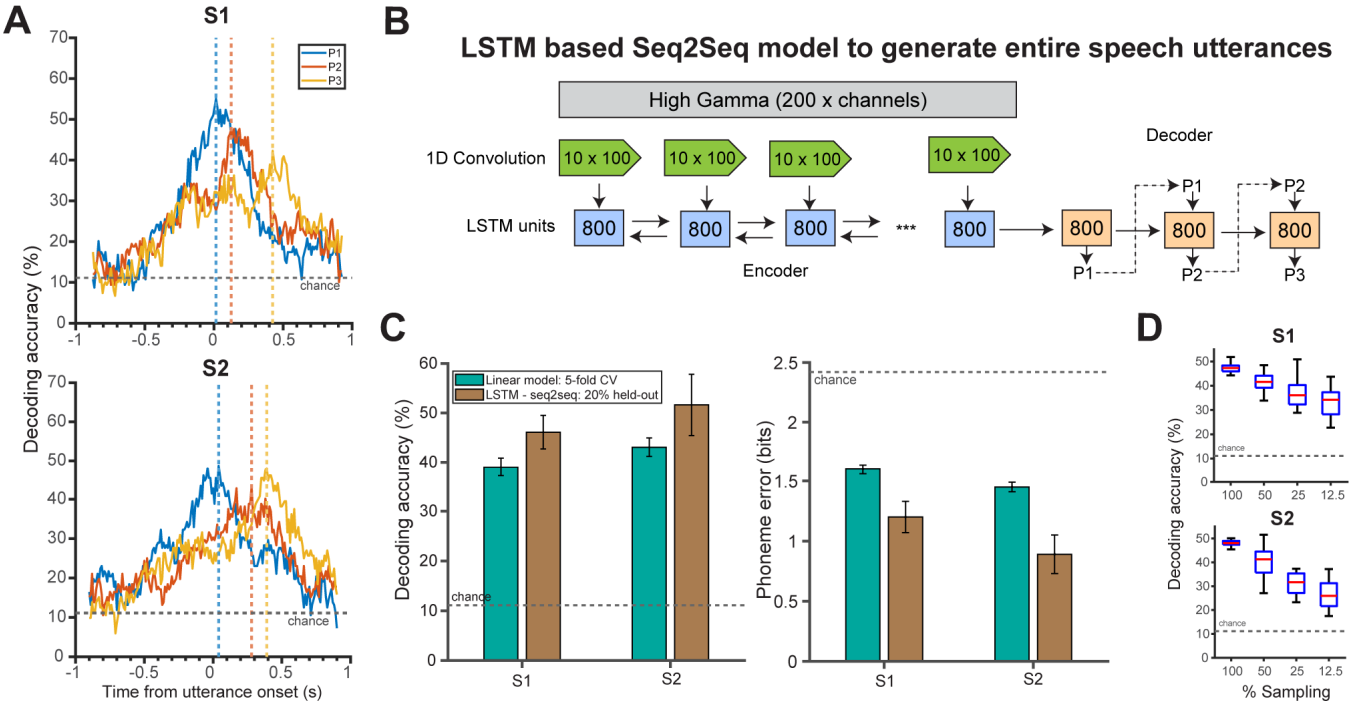

Figure 6. **µECoG enables the sequential temporal decoding of utterances. A)** A temporal decoding model was trained on data aligned to the speech utterance onset for phonemes at each position. Results are shown for two example subjects (S1, S2) revealing temporal patterns of successful decoding for each phoneme position (P1: blue, P2: orange, P3: yellow). Dotted lines reflect peak decoding. This ability demonstrates that the temporal sequence of phoneme positions can be reconstructed via the information present in the neural signal aligned to the utterance. **B)** We leveraged this ability by constructing a nonlinear decoding model to resolve temporal speech sequences of entire utterances. We used a long short-term memory (LSTM) Seq2Seq model with an encoder-decoder recurrent neural network (RNN) to translate HG spatio-temporal activation patterns (1 second window centered at utterance onset) to phoneme sequences, with three layers: a temporal convolution (green), a bi-directional encoder RNN (blue), and a decoder RNN (orange) layer. The numbers in the boxes represent the number of LSTM units **C)** Our LSTM model outperformed a linear model for decoding phonemes across all positions, shown here for two example subjects, S1 (mean accuracy: 46.1% vs. 38.9%, mean phoneme error: 1.2 bits vs. 1.6 bits) and S2 (accuracy: 51.6% vs. 43.1%, phoneme error: 0.89 bits vs. 1.4 bits). Error bars indicate standard deviations. **D)** Subsampling of electrodes reduced the LSTM performance in decoding phoneme sequences as validated by one-way ANOVA and post-hoc t-tests. This improvement in decoding demonstrated the benefit of combining nonlinear decoding models that leverage high spatio-temporal neural activation with µECoG recordings that can sample the human brain at the micro-scale.