Accurate speech decoding requires high-resolution neural interfaces

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[5,6], Jonathan Viventi[1,2,4,6, #], Gregory Cogan[2,5,6,7,8, #]

[#]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 Department of Psychology and Neuroscience, Duke University, Durham, NC, USA

8 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 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 speech production. We obtained neural signals with 56× higher spatial resolution and 1.48× higher signal-to-noise ratio compared to standard invasive recordings. This increased signal quality nearly doubled decoding accuracy 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 standard techniques. We show for the first time that micro-scale neural interfaces can enable high-quality speech decoding for neural speech prostheses.

Abstract Word Count: 150/150

1 Main

Amyotrophic lateral sclerosis (ALS) and clinical locked-in syndromes are motor neuron disorders that greatly reduce or eliminate the patients' ability to verbally communicate, dramatically affecting 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 neural signals. Advancement of high-resolution neural interfaces would enable highly accurate decoding which is paramount to developing successful neural speech prostheses.

Previous attempts to accurately decode speech utilized invasive macro electrocorticography (ECoG, 4 – 10 mm inter-electrode distance and 1.5 mm exposed diameter) that targeted speech motor cortex (SMC) during speech production[6–9]. These studies demonstrated that SMC encodes articulatory properties of speech motor sequences, which 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]. Accurate resolution of these features is therefore a crucial component to enable neural speech prostheses.

The rich spatio-temporal structure of human brain signals occurs at multiple spatial scales. Recordings of brain signals have previously been limited by macro ECoG recordings, which are typically measured from up to 64 contacts spaced 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 computations from the surface of the brain. HG has a high correlation with multi-unit firing and shows high spatial specificity[21–23]. Further, HG has been previously shown to accurately estimate neural-firing patterns that are stable over longer periods when compared to estimating single units using threshold based detection techniques[24]. Technology that obtains high-resolution, high-fidelity recordings of HG would therefore enable better neural speech prostheses.

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 distinctive at millimeter scale resolutions[25–27]. 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 high-resolution neural interfaces to accurately resolve HG to enable more accurate speech decoding.

Work in other domains have shown the ability to resolve micro-scale neural features using high-density micro-electrocorticography (μECoG) arrays in both rodents and non-human primates[28–31]. This high-resolution electrode technology also enabled the identification of micro-scale epileptic signatures of seizure-onset-zones in epileptic patients[32–36] and have shown improvements in motor neural prostheses by resolving the motor cortex at millimeter level resolution[37–40]. These results show that high-resolution neural interfaces could enable recording of micro-scale articulatory features from SMC, leading to high-performance neural speech prosthesis.

In this study, we recorded intra-operatively from patients using liquid crystal polymer thin-film (LCP-TF) μECoG arrays (1.33 – 1.72 mm inter-electrode distance, 200 μm exposed diameter) placed over SMC during a speech production task. We demonstrated the strength of our micro recordings against standard macro recordings. We decoded speech with μECoG arrays and compared their prediction performance to macro intracranial electroencephaographic (IEEG) recordings. Further, we examined the impact of spatial and sequential features of neural activation patterns that enable stronger decoding. We demonstrate that micro-scale recordings enable better speech decoding through increased spatio-

temporal sampling of neural signals. These results demonstrate the benefits of high-resolution brain interfaces for neural speech prostheses.

2 Results

2.1 High resolution speech-neural activations

We examined micro-scale speech neural activity from four subjects recorded using LCP-TF μECoG electrodes in the intraoperative setting. 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 subdural array (8 x 16 array; inter-electrode distance: 1.33 mm). Each of the subjects completed three task blocks (52 trials per block). Subject S4 (F: 26 years old) underwent surgery for a tumor resection and was implanted with a 256-channel subdural array (12 x 24 array; inter-electrode distance: 1.72 mm). This subject completed one block of the speech repetition task (**Fig. 1A & 1B**, Supplementary Table 1). 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. 1C**). Subjects correctly repeated the nonwords on more than 95% of the trials (**Fig. 1D,** S1: 96%, S2: 98%, S3: 98%, S4: 100%), indicating that subjects understood and could complete the task in the intraoperative setting.

We observed uniform *in vivo* impedance throughout the array (Supplementary Fig. 1), and discarded electrodes with higher impedance (> 1 MOhm) for neural analysis. We observed significant modulation of spectro-temporal activation patterns during speech articulation and significant increases in HG power. Example arrays demonstrate distinct spatial patterns of activation on the micro-scale and are shown for both 128 channel (S1) and 256 channel (S4) grids (**Fig. 2A**). Each individual channel resolved strong and distinct HG modulation in SMC aligned to the speech utterance (**Fig. 2B,** and Supplementary Fig. 2). Electrodes with significant HG power were identified using a non-parametric permutation test with an FDR-corrected alpha threshold of $p < 0.05$ (see Methods) and 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 than standard methods. To assess this benefit, we examined the evoked-signal-to-noise ratio (ESNR) of HG power from μECoG neural recordings and compared them to recordings from standard IEEG (see Methods). We pooled IEEG electrodes (ECoG and stereo-electroencephalography - SEEG electrodes implanted during clinical pre-operative epileptic monitoring on separate patients) anatomically localized to SMC that exhibited significant HG power during speech articulation. Our μECoG recordings demonstrated a 1.48x (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) when compared against standard IEEG (median ESNR = 1.2 dB, **Fig. 2C**).

To quantify the uniqueness of each μECoG micro-electrode, we measured the distinctiveness of neural signals between neighboring micro-electrodes by computing the Pearson correlation coefficient of corresponding HG envelopes. The HG correlation increased with proximity from 10 mm to 1.33 mm, and remained largely uncorrelated at 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 showed higher correlations, but were also largely uncorrelated ($r = 0.6$), indicating the presence of spatially distinct information in speech activations at scales below 2 mm.

To investigate this micro-scale neural information, we examined spatio-temporal activations specific to some key units of speech production: the phoneme and articulatory features. 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.

Previous studies examining speech motor articulation identified functional organization patterns of phonemes based on their composite articulators in SMC[41,42]. We sought to identify similar population-driven articulator patterns for μECoG HG 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 HG activity into low-dimensional features (80% of variance explained), and used t - distributed Stochastic Neighbor Embedding (tSNE)[43] to visualize the transformed eigenvalues in two-dimensional vector space. **Fig. 3B** shows clear separation of speech trials in tSNE state-space for both articulators as well as phonemes. We examined the contribution of high spatial-resolution by repeating the state-space analysis at subsampled μECoG HG (Poisson disc sampling). We observed decreased distinction for both articulators and phonemes with reduced spatial sampling (**Fig. 3C,** Supplementary Fig. 4). This effect is further quantified by a significant decrease in silhouette scores for sub-sampled versions in both articulator and phoneme space (**Fig. 3D**). 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 ($F(3, 196) = 114$, $p = 9.6e-43$) and phoneme space ($F(3, 196) = 47.8$, $p = 3e-23$. Post hoc t-tests 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 show that high-resolution neural interfaces enable more accurate sampling of motor features in SMC.

2.2 Decoding phonemes using μECoG spatio-temporal features

To assess the capability of micro-recordings for a neural speech prosthesis, we evaluated speech decoding accuracy by decoding manually-aligned individual phonemes in spoken nonwords from μECoG neural signals. We performed a 20-fold nested cross-validated decoding model based on the 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 70-80% of neural variance for phoneme prediction. Benefiting from high spatial sampling, we observed strong decoding performance in all subjects in predicting phonemes in all positions (**Fig. 4A & 4 B**, $p < 0.01$, Binomial test[44]). All subjects had higher accuracies in decoding first position phonemes (average P1 across subjects: 49.9%, chance = 11.11%) when compared to phonemes in the second (P2: 45%) and third positions (P3: 38.8%). S1 and S2 exhibited higher decoding performances across all positions attributed to higher HG-SNR and increased utterance duration (**Fig. 4B**). S1-P1 had a maximum prediction score of 57% with the best consonant (/g/) decoding at 84.2% and best vowel decoding (/i/) at 75.3% (**Fig. 4A**). Similarly, S2-P1 had similar maximum prediction scores with the best consonant (/g/) decoding at 72.7% and the best vowel (/u/) decoding at 83.3%. The 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 both that a linear decoder can be affected by the manual-alignment of phonemic onsets within an utterance and that there may be unique position-based information present in the signal Together, this indicates a need for position-dependent decoding. The lower prediction quality of S3 was attributed to low HG-SNR and lower spatial coverage. 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 highly-varying decoding performances across phoneme positions (data not shown).

To assess the utility of recording micro-scale neural signals, we compared μECoG decoding to standard IEEG recordings (SEEG/ECoG). We performed the same phoneme decoding analysis using neural signals recorded from ten epilepsy patients (4 Male, mean age = 30) who were implanted with standard IEEG electrodes during pre-operative epilepsy monitoring. 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 μECoG outperformed standard clinical recordings by a factor of nearly two (**Fig. 4C**: SEEG/ECoG: 22.8%, μECoG: 49.9%, chance 11.11%, $p = 0.0059$, Mann-Whitney U-test).

Overall, these strong phoneme decoding results highlight the high-resolution recording ability of μECoG for accurate neural speech prostheses.

Since high-density μECoG demonstrated both a greater ability to resolve higher ESNR and superior decoding performance as compared to standard clinical recording methods, we next sought to determine the relationship between decoding performance and ESNR. To quantify this relationship, we examined the phoneme tuning properties of μECoG by calculating the decoding performance of 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 stronger 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 resolve high fidelity signals using high-density μECoG and the 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 errors would not be uniform across phonemes, but rather reflect similar articulatory properties of the resolved signals recorded. For example, in **Fig. 4A**, we see increased confusion between phonemes with similar phonological features (/p/ vs. /b/), as previously reported in phoneme decoding studies[6,14,16]. To systemically characterize these types of errors, we calculated each phonemes' phonological features by assigning a 17-bit categorical vector to each phoneme. These features were taken from classic work on phonological features[45] (see Methods). We used Hamming distance to calculate phonological distance for each phoneme pair (Supplementary Fig. 5), and observed that phonemes with similar phonologies were arranged at the 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**, $p < 0.01$, F-statistic test against constant model). We observed similar inverse relationship for all subjects and the resultant phoneme errors were well lower than chance, (**Fig. 4F** chance error: 2.42 bits; uniform confusion matrix). S1 and S2 had the lowest phoneme distance error across all phoneme positions (less than 1 bit; except S1-P3), indicating phoneme misclassifications were predominantly biased with phonological articulation. This analysis shows high-density μECoG exhibited decoding error that is largely due to articulatory structure, and not because of random misclassifications.

2.3 Decoding is improved with increased resolution, coverage, and micro-scale electrodes

Having shown an increased HG-ESNR, and superior decoding for S1 and S2, we next sought to directly study the effect of spatial resolution, spatial coverage, and micro-scale sampling on phoneme decoding. First, we examined the effect of spatial resolution on decoding by sampling electrodes (n ≥ 30 samples) across the array, using Poisson disc sampling. For each sampling, the resultant pitch 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. We then grouped the accuracy values by the resultant pitch (rounded to nearest integer) and compared every distribution 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, and the subset of electrodes with spacing less than 1.5 mm reached the 95% threshold of the full array (**Fig. 5A** (left), & Supplementary Fig. 6), indicating that accurate speech decoding requires at least millimeter-scale cortical sampling. Further, the subsampling distributions from the 4 to 10 mm range (electrode distance range for standard IEEG system) had median accuracy values of 20-35%, values that were close to values obtained from standard IEEG recordings (**Fig. 4C**), which validates our sub-sampling approach. Next, to demonstrate the importance of spatial coverage we subsampled electrodes from rectangular subgrids at fixed spatial resolution. Each of these subgrids were two dimensional windows with dimensions from 2 x 4 to 8 x 16 electrodes. The decoding performance improved with increased coverage (**Fig. 5A** (middle) & Supplementary Fig. 6). Finally, to investigate the impact of micro-scale contact-size on speech decoding, we used rectangular subgrids with dimensions from 1 x 1 (200 μm) to 8 x 8 to spatially average raw neural recordings. HG signals extracted from spatially smoothed neural recordings were used to decode phonemes. The decoding accuracy decreased with increased contact size, and the highest performance was held at the lowest contact size (**Fig. 5A** (right) & Supplementary Fig. 6). Together, these results demonstrate the necessity of high-spatial resolution, large spatial coverage, and micro-scale electrodes to obtain high decoding accuracy for neural speech prostheses.

We next used our recordings from micro-scale signals to identify articulatory maps in SMC. Previous studies have identified spatial characteristics in human SMC examined articulator encoding[42], however, the fine-scale spatial properties on a subject-specific scale are not well characterized. To quantify

micro-scale neural articulatory maps, we trained an SVD-LDA model for each electrode to decode four articulators (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 the 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 a high level of overlap of informative electrodes across articulators. Despite this overlap, high-density μECoG arrays exhibited distinct clustering of electrodes for articulators across all subjects, indicated by red contours (90 percentile of AUC) and corresponding centroids (**Fig. 5B &** Supplementary Fig. 7). 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 3). 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. 7 and Supplementary Table 3). These subject-specific high-resolution articulator maps further show the need for micro-scale cortical sampling using high-density μECoG for human speech decoding.

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. These signals however reflect underlying complex neural dynamics, which also evolve in time. We therefore sought to leverage our ability to spatio-temporally resolve high resolution neural signals by modeling sequential phoneme dynamics. To assess the phoneme sequential decoding ability of HG neural information, we constructed an SVD-LDA model trained on phoneme-aligned neural data for each position that was then tested on utterance aligned data. Since the training data was manually aligned to phoneme onsets, sequencing should only be present in the utterance aligned testing set if phonemes were positionally encoded in HG activations.

For both S1 and S2 that obtained superior decoding results, we obtained stable linear decoding contours (90th percentile of accuracy values) for phonemes across all positions. The mean optimal training times ranged from 5 – 175 ms post phoneme utterances for all three positions, but testing times were optimal at the sequential ordering of phonemes (**Fig. 6A**: S1: -0.03, 0.15, and 0.48 s for Position 1, 2, and 3 respectively, and for S2: 0.05, 0.225, and 0.4 s for utterance aligned centroid of decoding positions – dotted lines). These results indicate that phoneme identities across positions can be sequentially represented from neural data aligned to speech utterance onset. Overall, we elucidate sequential coding of phonemes in an utterance from micro-scale neural recordings that capture rich spatio-temporal neural features. This ability is fundamental to reconstruct motor sequences of speech for neural speech prostheses.

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

Having demonstrated the presence of sequential phoneme information in the micro-scale neural signal, we sought to extract this information more directly. We reasoned that a neural speech prosthesis would require the ability to decode speech sequences without explicit phoneme onset information. We therefore developed a phoneme generation technique to sequentially decode positional phonemes within the utterance. Here, we used a seq2seq RNN model to translate HG neural activity to the entire phoneme sequence directly[18]. For each spoken utterance, HG activations 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, **Fig. 6B**) contained a set of optimal learned filters. These time-domain filters downsampled HG activations resulting in low dimension temporal features with length of 50 ms.
2) The bidirectional encoder RNN (blue boxes, **Fig. 6B**) captured the downsampled features at every time-step, and used a hidden state (memory unit) to encode neural features along the time-axis. The hidden state from the final encoder RNN (average of forward and backward states) acted as a neural feature representation for all the phonemes in the utterance.
3) The decoder RNN (orange boxes, **Fig. 6B**) was initialized by the final hidden state of encoder RNN, to transform neural features to phoneme sequences. At each phoneme step, the decoder received input of its own state and its predicted phoneme at the previous state.

We trained the entire network using Adam[46] 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 4).

We quantified the decoding performance by calculating the percentage of correctly predicted phonemes for each position. The resultant phoneme decoding accuracies and phoneme errors were considerably better for seq2seq models for both S1 and S2 as compared to equivalent linear decoding models that used neural information aligned to individual phoneme onsets (**Fig. 6C** and supplementary table 5). This result demonstrates that non-linear sequential decoding can more accurately reconstruct spoken phoneme-sequences than linear methods that require explicit phoneme onset times. This sequential decoding method provides an algorithmic framework for a neural speech prosthesis from high-resolution neural interfaces.

3 Discussion

We elucidated the importance of using micro-scale neural recordings for the development of neural speech prostheses. Our high-density µECoG arrays recorded neural signals in SMC that produced significantly higher fidelity HG-ESNR as compared to standard intracranial methods (ECoG/SEEG). This result is 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[47,48]. Importantly, microseizures were observed occurring on just two electrodes spaced 750 µm apart, and were not visible on neighboring electrodes. In the present work, we measured HG activation which has previously been estimated to have a resolution of 0.25 mm to several millimeters[49–51]. We empirically validated these previous estimations by demonstrating HG inter-electrode correlation of r ~= 0.6 at <2mm spacing during the speech utterance[25,26]. This higher fidelity recording ability combined with evidence for distinct neural patterns of activations at the micro-scale, argues for the advantage of using micro-recording technologies for neural speech prostheses.

To further characterize 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[41]. We found that clustering however, was dependent on micro-scale spatial resolution as a decrease in sampling produced significantly worse clustering results. This inverse relationship further shows the necessity of high-resolution neural recordings to accurately model speech cortical dynamics.

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. S1 and S2 achieved the highest decoding performances. Subjects S3 and S4 obtained lower decoding values likely due to low SNR and reduced number of trials respectively. Previous speech decoding studies with phoneme classification results employed electrode technologies ranging from standard ECoG[13–15,17] to penetrating intracortical arrays[16]. These studies involved subjects performing speech production tasks while in the monitoring unit, with total utterance durations (actual speaking time) of >~10 minutes. In our study, we report stronger accuracy metrics from intra-operative data with total spoken duration of only ~1.1 minutes (156 total trials x 98% response rate x 450 ms mean response duration). In addition, when directly comparing our decoding results to work using electrode arrays with 1 mm diameter electrodes, spaced 3-4 mm apart[15], our micro-scale (0.2 mm diameter electrodes), high-density (1.4-1.7 mm spacing) arrays performed better: 85% vs 75% (best phoneme: 91.7% vs. 82.2%, 5-way decoder). Finally, from our own control study, we observed superior decoding ability of µECoG electrodes when compared to standard IEEG electrodes targeted to SMC. These results show that even when using short time periods, micro-scale neural recordings using high-density µECoG

outperform current technology. We hypothesize that with longer recordings of spoken utterances, our decoding performance will show even better performance.

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. This means of evaluating speech decoding models with phonological distance could lead to error correction that comparatively weighs similar articulated phonemes with a lower cost versus phonemes with dissimilar features. This phonological error metric could be used in addition to classification accuracy to optimize cost-functions for neural speech prostheses.

In this study, we characterized the importance of spatial characteristics of micro-scale neural dynamics for speech decoding. We observed superior phoneme decoding performance that increased with spatial resolution, spatial coverage and micro-scale sampling. For the first time, we observed high-resolution, subject-specific articulator maps, analogous to previously established group-level somatotopic organization[42]. 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. Future work will systemically identify and characterize these maps both within and across subjects. Our ability to measure millimeter-scale maps at the individual subject level necessitates high-resolution recordings for the development of neural speech prostheses.

The benefits of using high-density μECoG for speech decoding was not only restricted to the spatial domain. Speech production involves the temporal sequencing of articulatory movements[41,42], and we argued that neural properties of phoneme articulators should be sequentially encoded. We demonstrated sequential dynamics of phoneme articulation by successfully decoding individual phonemes in time. This sequential decoding reflected the correct order of the serial phoneme execution during the speech utterance. Our cross-modeling strategy of training and testing resulted in linear and stable decoding activation contours across the utterance duration period for each phoneme position. These decoding activation contours overlapped in time for each successive phoneme providing evidence for potential co-articulatory dynamics during the speech utterance. This finding provides a way to sequentially represent all phonemes within a spoken utterance. We then leveraged this ability to decode the temporal dynamics of the speech utterance by training a deep-neural-net based sequential model to decode all phonemes within an utterance. This model did not require any pre-defined phoneme onset labels and yet showed a significant improvement when compared to our linear

modeling that required phoneme onset information. Thus, micro-scale neural recordings using μECoG arrays combined with nonlinear models can more accurately capture high-resolution spatio-temporal neural dynamics for successful neural speech prostheses.

Our results identified the benefits of high-density, high-channel count LCP-TF μECoG arrays for human speech decoding and the development of neural speech prostheses. Our micro-scale recordings enabled more accurate phoneme decoding compared to standard IEEG systems and showed a more effective characterization of the spatio-temporal dynamics of speech production. Our findings motivate the use of high-density μECoG for reliable neural interfaces, thereby providing effective means of increasing the quality of life for patients suffering from debilitating motor disorders who have lost verbal communication abilities[52].

4 Methods

4.1 Participants

We recorded neural activity from speech motor cortex (SMC) of four awake patients using LCP-TF μECoG arrays, during intraoperative procedures at Duke University Medical Center (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 DBS implantation surgery. The fourth subject (S4) underwent a brain tumor resection and neural data were recorded during the awake craniotomy. As a control, we recorded from eleven epileptic patients implanted with conventional electrodes (ECoG or SEEG) and recorded with a clinical IEEG system (Natus Quantum), during preoperative intracranial monitoring at the Duke University Medical Center. All subjects were fluent in English and we obtained written informed consent in accordance with the Institutional Review Board at the Duke University Health System.

4.2 Intra-operative μECoG array implantation

We performed intraoperative recordings using custom LCP-TF μECoG arrays 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 diameter) arranged in an 8 x 16 fashion with 1.33 mm center-to-center spacing/pitch (**Fig. 1A,** top). The 128-channel array design is narrow and flexible enough to be implanted through the burr-hole opened during DBS implantation. The 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 pitch) and was used to target anatomy directly over eloquent cortex exposed during an awake craniotomy (**Fig. 1A,** 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 Duke Medical Center prior to surgical implantation.

In order to co-register the arrays between subjects, the array locations were assessed using Brainlab neuronavigation software. The array templates (for both 128 & 256 channels) were 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. 1B** 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. For the control subjects implanted with the standard IEEG system, the intracranial data were recorded using a Natus Quantum LTM amplifier (Natus Medical, Inc.), with an analog filter between 0.5 and 500 Hz and digitized at 2,048 samples per second.

4.4 Speech task

The subjects performed a speech production task during the intra-operative procedure. In each trial, they listened 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, and which included a total of nine different phonemes (four vowels: /a/, /ae/, /i/, /u/ and five consonants: /b/, /p/, /v/, /g/, /k/). These nonwords were auditorily presented and subsequently repeated by the subject during an awake period

within the surgical procedure. Each trial lasted between 3.5 – 4.5 s including the auditory stimulus (mean stimulus duration: 0.5 second) followed by three seconds 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 performed a similar speech production task during preoperative clinical monitoring in the epilepsy monitoring unit at Duke University Medical Center. This slightly modified task contained a delay period between the auditory stimulus presentation and the instructed speech response.

4.5 Audio recording

The speech production task included both listening and speaking components. For the listening component, the stimulus waveforms 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 and 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 activations. The spectral information was extracted for every 500 ms analysis window (50 ms step-size) with 10 Hz frequency smoothing using a multi-taper analysis[8]. The resultant spectrograms were normalized with respect to baseline period (500 ms window pre-stimulus across trials) to remove *1/f* broadband activity. The normalized spectrograms were then averaged across trials and the log-power was computed 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 a good correlate of multi-unit firing[21,22]. 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[53]. The envelope of each frequency band was extracted using the Hilbert transform and was averaged across the bands. The resultant envelope was then downsampled to 200 Hz using an anti-aliasing filters and mean-normalized with respect to the baseline period. For each trial, we used a one second time-window of normalized HG centered at phonemic onset (marked offline) 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[39,54]. SVD was carried out on the covariance matrix of the HG tensor (dimension: *U* X *ET*; U – number of utterance trials, E – number of significant HG μECoG electrodes, T – time duration of HG) to decompose highly covarying space-time activations across the array.

$$H: U\ XET \quad (3)$$

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 an a principal component (PC) score corresponding to the eigen-vector. 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)[43], a non-linear affine transformation technique for visualizing high-dimensional data. Finally, the trial embeddings were coded to represent either phoneme or articulator identities.

To evaluate the clustering of trials with respect to either articulatory or phoneme group, we performed a silhouette analysis (Euclidean distance)[55] 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 number of trials with positive silhouette score to total number of trials, and 2) silhouette score, 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[30]. For each distribution, we performed 50 sampling iterations, and computed silhouette-scores on the sub-sampled cortical state-space.

4.6.4 Linear decoding model using SVD-LDA

We constructed a linear-classification model to decode phonemes units from normalized HG neural activations. First, we used SVD to compress the HG matrix to a low-dimensional subspace as explained in the previous section (Equations 3 & 4). Then, we utilized the resultant eigenvalues to classify phonological units using a supervised linear-discriminant (LDA) model. The optimal number of eigenvectors was identified using a nested cross-validation procedure and the overall decoding accuracy was calculated by calculating the proportion of correctly predicted labels. We trained decoding models to decode phoneme for each of the three positions in the nonword (P1, P2, and P3) and collapsed across all positions.

To characterize decoding error, we utilized values in the confusion matrix to test for confusions 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 (1968), to represent phonological features[45]. 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 unlike 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 spatial and sequential representation of phonological features. To infer 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 articulators. 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 trained at time-segment *t*, to test at time *t'*[56]. With this cross-modeling strategy, each SVD-LDA model was trained on sequential time-segments with respect to phoneme onset, and the accuracy values were evaluated on time-segments corresponding to the utterance onset. 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 250 ms windows (25 ms step size) ranging from -1000 ms to 1000 ms. The resultant square-matrix (training time x testing time) of accuracy values was used to determine the sequential dynamics of phoneme representation 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 activations to phoneme sequences: 1) Temporal convolutional layer, 2) Encoder recurrent neural network (RNN), 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 a bi-directional LSTM (Long-Short-Term-Memory)[57] 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 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[58]. The network was trained using the 'Adam' gradient descent optimization by minimizing the categorical cross-entropy between true and predicted phoneme probabilities[46]. 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, 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[59]. During testing, we generated the most probable phoneme sequence by identifying the most likely phoneme at each position and used it as the input to the next position.

We identified the optimal hyperparameters for the neural-network using the keras-tuner platform[60]. For each subject, we randomly explored the hyperparameter space through 200 trials (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 4.

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 phonological error in decoding phonemes across all positions.

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 comparison 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. Finally, we used one-way ANOVA to examine the effect of spatial subsampling on state-space clustering, 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[44]. All population data were 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.

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**, 35015 (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. Dubey, A. & Ray, S. Cortical Electrocorticogram (ECoG) is a local signal. *J. Neurosci.* (2019).

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

24. 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).

25. 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).

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

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

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

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

30. 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).

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

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

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

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

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

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

37. Wang, W. *et al.* Human motor cortical activity recorded with Micro-ECoG electrodes, during individual finger movements. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2009**, 586–9 (2009).

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

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

40. Trumpis, M. *et al.* Sufficient sampling for kriging prediction of cortical potential in rat, monkey, and human μECoG. *J. Neural Eng.* **18**, 036011 (2021).

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

42. 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).

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

44. 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).

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

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

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

48. 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.

49. Lindén, H. *et al.* Modeling the spatial reach of the LFP. *Neuron* **72**, 859–872 (2011).

50. Kajikawa, Y. & Schroeder, C. E. How local is the local field potential? *Neuron* **72**, 847–858 (2011).

51. Katzner, S. *et al.* Local Origin of Field Potentials in Visual Cortex. *Neuron* **61**, 35–41 (2009).

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

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

54. 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).

55. Kaufman, L. & Rousseeuw, P. J. *Finding groups in data: an introduction to cluster analysis*. (John Wiley & Sons, 2009).

56. 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).

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

58. 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).

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

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

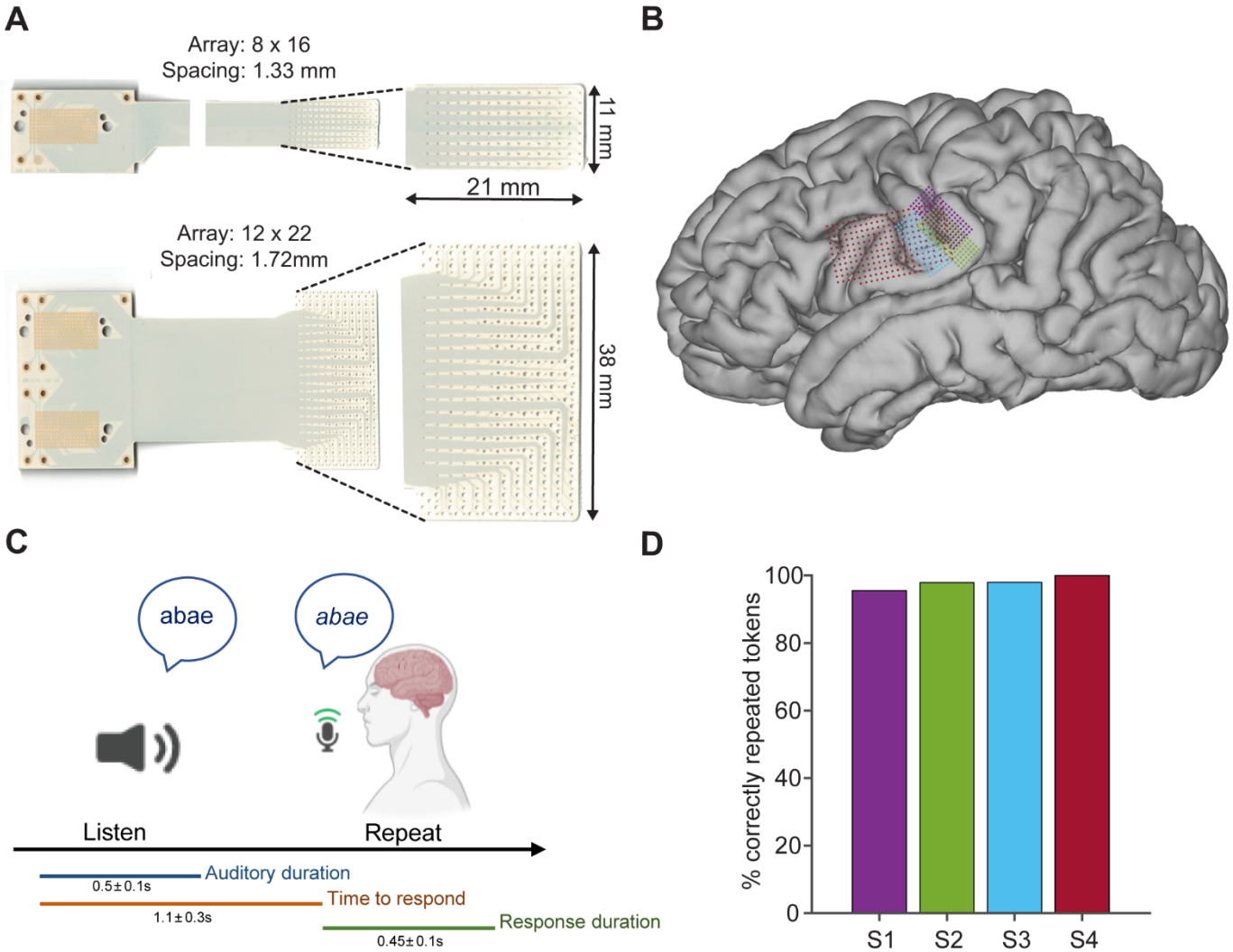

Figure 1. **Recording from high-density micro-electrocorticographic (μECoG) electrodes on the human brain during a speech production task. A)** Flexible liquid crystal polymer (LCP) electrodes with 128 (top) and 256 (bottom) recording electrodes (200-um diameter). **B)** Electrode arrays were implanted over speech motor cortex from 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). **C)** A schematic of the intra-operative speech production task. Color bars indicate the durations of the auditory stimulus (blue), time-to-response (orange), and response durations (green). **D)** Subjects performed the speech production task well with behavioral results above 95% accuracy for correctly repeated nonwords.

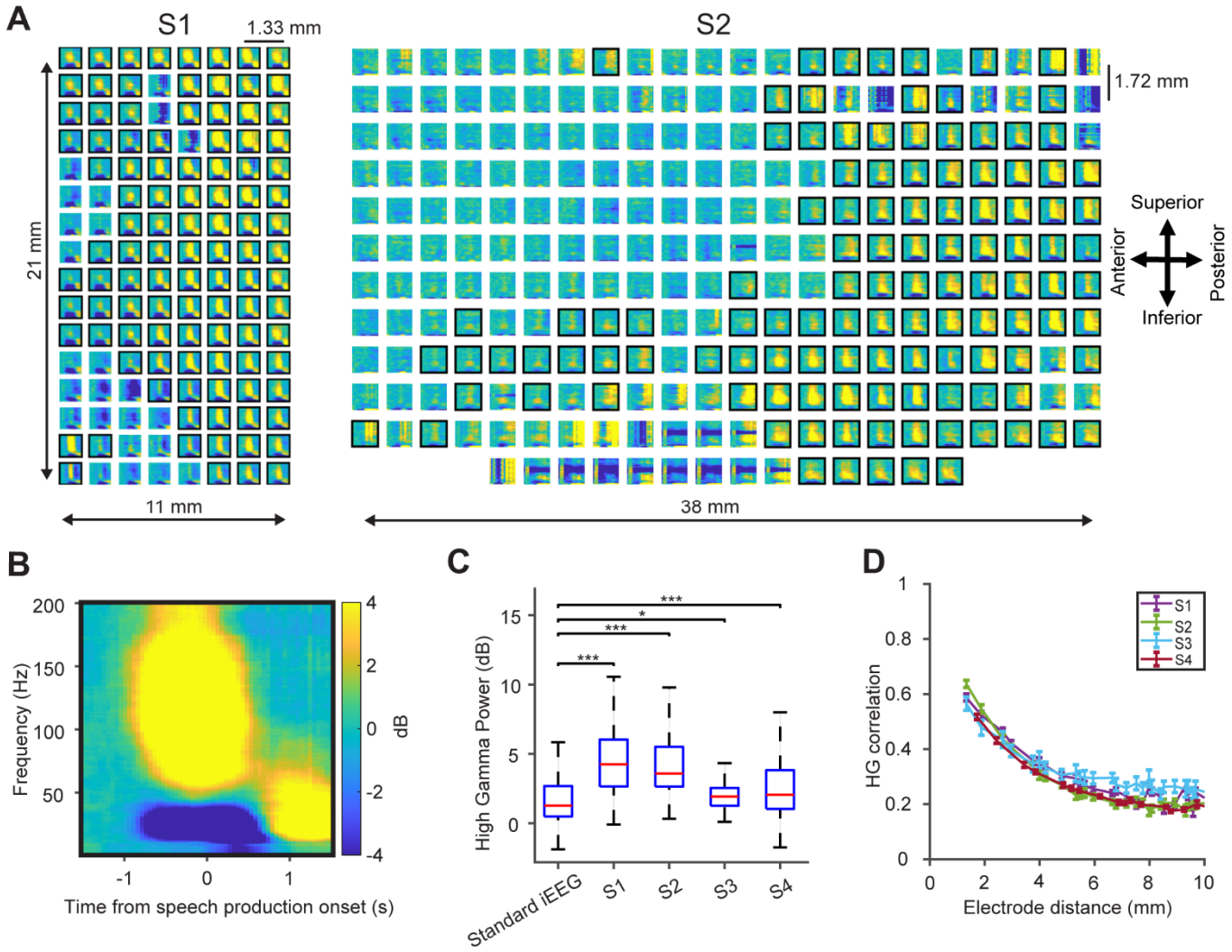

Figure 2. Highly resolved neural features as demonstrated by high-density μECoG during speech production. **A)** Spectrograms for each electrode from each array for two example subjects: S1, 128 electrodes (1.33 mm spacing) & S4, 256 electrodes (1.72 mm spacing). Data for all subjects are available in 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. **B)** A spectrogram of an example electrode from the 128-channel array 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), ***$p < 0.001$, 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 distinctive neural signals at the micro-scale.

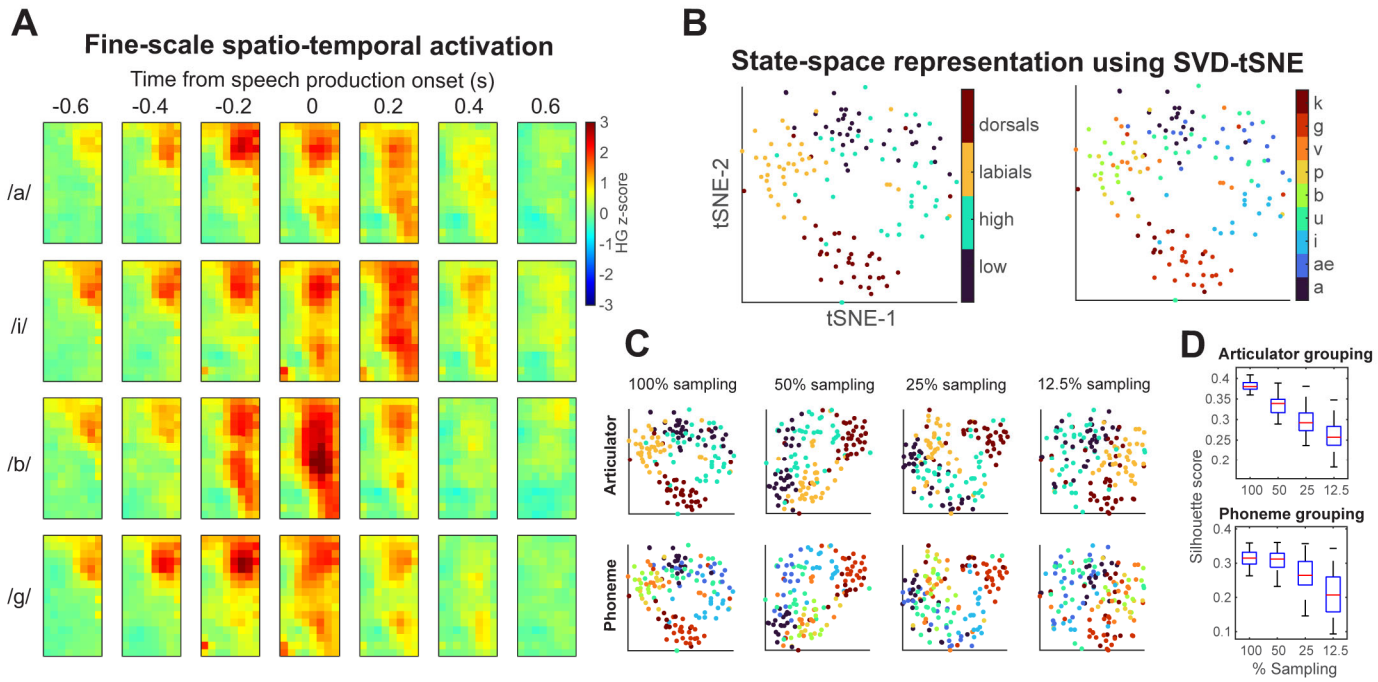

Figure 3: µECoG captures fine-scale spatio-temporal activity during speech production. A) HG activations showed fine-scale spatio-temporal patterns in speech motor cortex, shown here for an example subject (S1) during speech production. Spatial maps of HG responses showed distinct activation patterns 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 (dorsals - /g, k/, labials (/b, p, v/), high (/i, u/), and low (/ae, a/) as well as individual phonemes. Phonemes of similar articulators were grouped together in three-dimensional state-space demonstrating similarity in motor activation patterns. **C)** Electrode sub-sampling to simulate lower resolution recording 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 a phoneme/articulator 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 medianand 25/75th percentile and dashed lines represent the full-range of the distribution. These analyses demonstrate the improved ability of high-resolution neural sampling to distinguish between articulator and phoneme motor features.

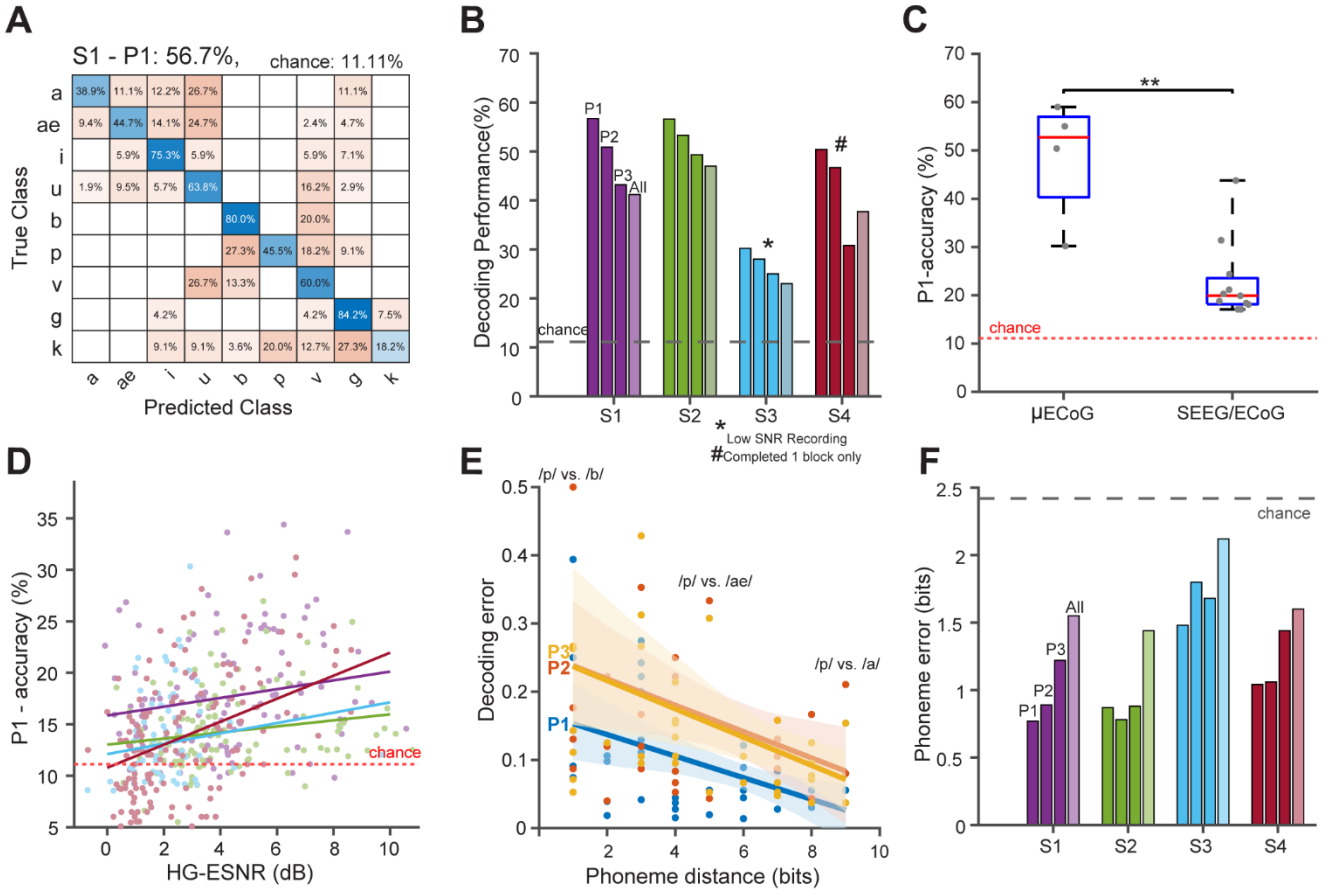

Figure 4. **High accuracy phoneme decoding from μECoG. A)** A confusion matrix for an example subject (S1) for phoneme position one (P1) using singular value decomposition linear discriminant analysis (SVD-LDA) decoding demonstrated high accuracy for each phoneme. **B)** Decoding was also highly accurate 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 highly accurately decoding outperformed standard IEEG (P1: μECoG: 49.9%, vs. SEEG/ECoG: 22.8%, chance = 11%, $p = 0.0059$, Mann-Whitney U-test). This nearly doubled performance increase further demonstrates the benefit of higher resolution neural recordings for speech decoding. **D)** Distribution of single-electrode decoding accuracies 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). **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).

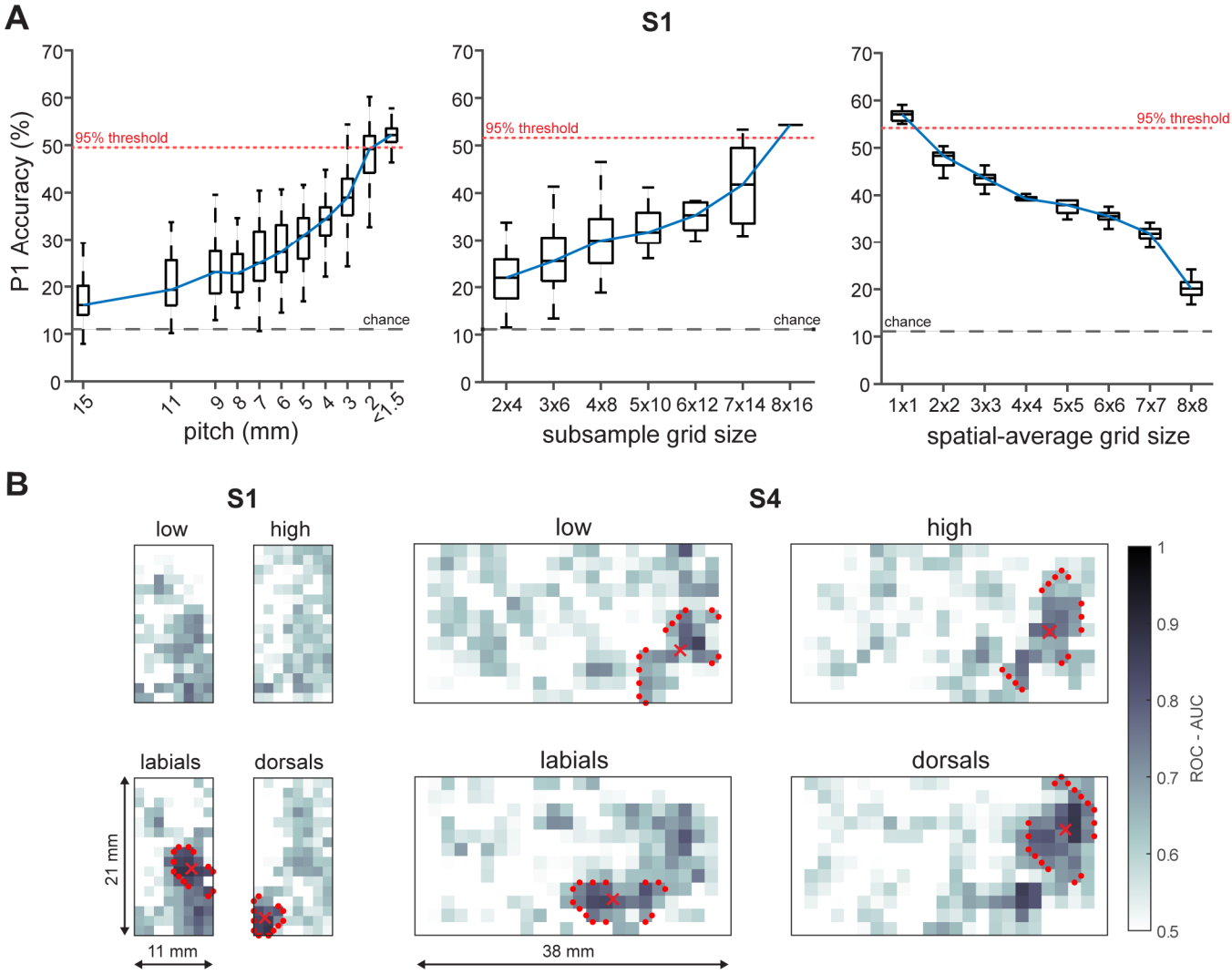

Figure 5. **Accurate decoding is dependent on the high-spatial resolution, high-channel count capability, and micro-scale recording of µECoG. A)** Accuracy of P1 decoding increased as a function of spatial resolution and coverage as demonstrated by sub-sampling electrodes, and decreased with contact 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) further demonstrating the benefits of high-resolution recordings for neural decoding. **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.

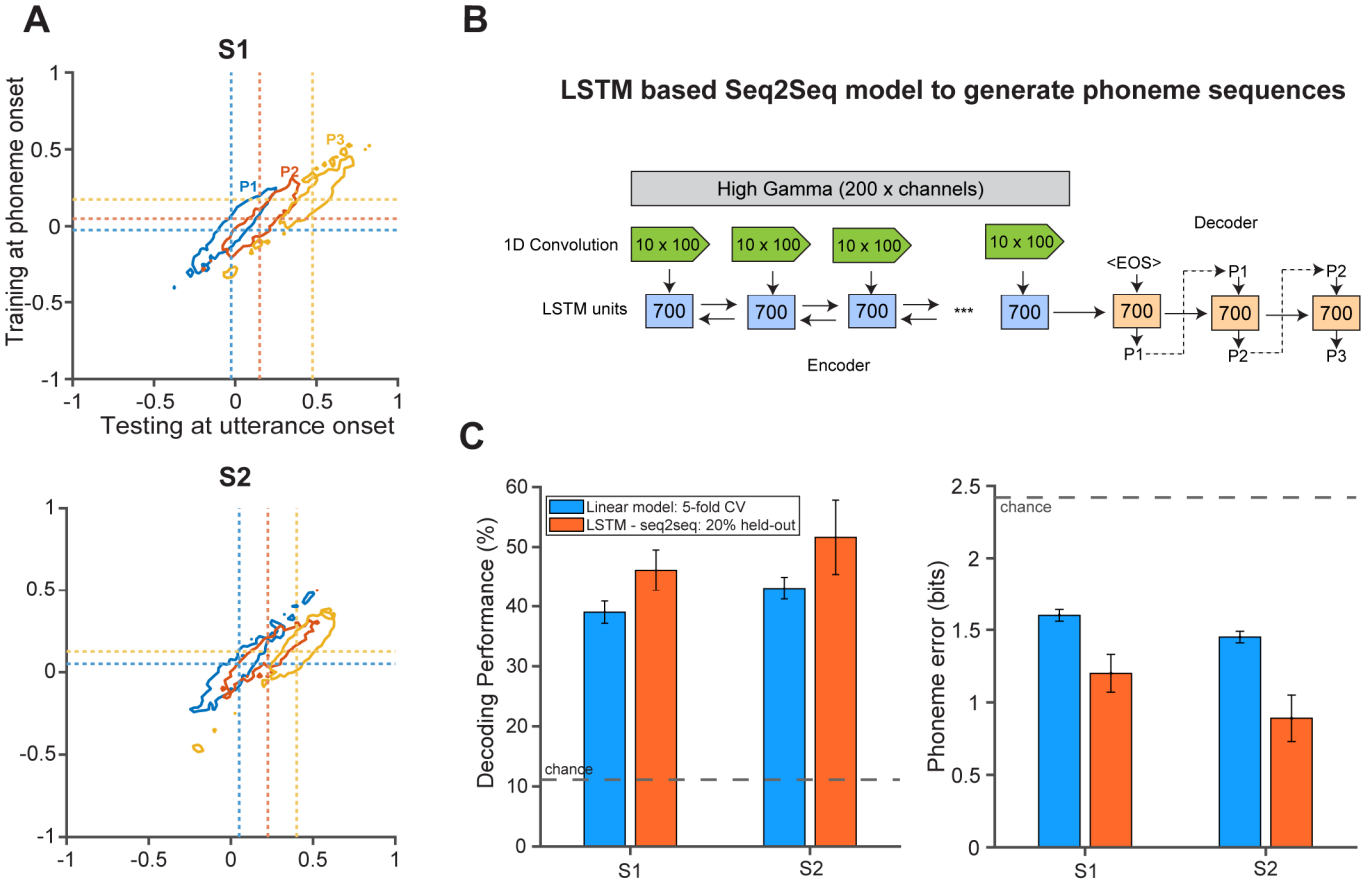

Figure 6. **µECoG enables the sequential temporal decoding of phonemes. A)** A temporal generalization model was trained on data aligned to each individual phoneme and was then tested on the whole utterance. Result are shown for two example subjects (S1, S2) revealing reconstructed temporal windows of successful decoding for each phoneme position (P1: blue, P2: orange, P3: yellow – outlines show 99th percentile). Dotted lines reflect centroid of the decoding area. These windows were aligned to the phoneme onset for our model training but demonstrated sequential temporal activation patterns from testing aligned to the whole utterance. This ability to demonstrates that the temporal sequence of phoneme positions can be reconstructed via the information present in the neural signal aligned to the utterance. **B)** To leverage this ability, we constructed a nonlinear decoding model to take advantage of the ability to resolve temporal speech sequences. 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), bi-directional encoder RNN (blue), and a decoder RNN (orange) layer. **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. 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.