Shared latent representations of speech production for cross-patient speech decoding

Spalding, Z.[1], Duraivel S.[1], Rahimpour S.[2,7], Wang C.[1], Barth K.[1], Schmitz C.[1], Lad S. P.[2], Friedman A. H.[2], Southwell, D. G.[2,3,4], Viventi J.[1,2,3,4,#], & Cogan G. B.[1,2,4,5,6,#]

[#]Corresponding authors

Author affiliations:

1. Department of Biomedical Engineering, Duke University, Durham, NC
2. Department of Neurosurgery, Duke School of Medicine, Durham, NC
3. Department of Neurobiology, Duke School of Medicine, Durham, NC
4. Duke Comprehensive Epilepsy Center, Duke University, Durham, NC
5. Department of Neurology, Duke School of Medicine, Durham, NC
6. Center for Cognitive Neuroscience, Duke University, Durham, NC
7. Department of Neurosurgery, University of Utah, Salt Lake City, UT

Abstract

Speech brain-computer interfaces (BCIs) can restore communication in individuals with neuromotor disorders who are unable to speak. However, current speech BCIs limit patient usability and successful deployment by requiring large volumes of patient-specific data collected over long periods of time. A promising solution to facilitate usability and accelerate their successful deployment is to combine data from multiple patients. This has proven difficult, however, due to differences in user neuroanatomy, varied placement of electrode arrays, and sparse sampling of targeted anatomy. Here, by aligning patient-specific neural data to a shared latent space, we show that speech BCIs can be trained on data combined across patients. Using canonical correlation analysis and high-density micro-electrocorticography (μECoG), we uncovered shared neural latent dynamics with preserved micro-scale speech information. This approach enabled cross-patient decoding models to achieve improved accuracies relative to patient-specific models facilitated by the high resolution and broad coverage of μECoG. Our findings support future speech BCIs that are more accurate and rapidly deployable, ultimately improving the quality of life for people with impaired communication from neuromotor disorders.

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: Cross-Patient Speech Decoding

Keywords: Speech, IEEG, BCI, decoding

Main

Speech is an intricate coordination of muscle movements that shape sounds into words[1]. Paralysis of these muscles by neuromotor disorders such as amyotrophic lateral sclerosis (ALS) and brainstem stroke can disrupt speech, resulting in debilitating losses to communicative ability and a reduced quality of life[2–6]. Speech brain-computer interfaces (BCIs) have shown promise in restoring communication to individuals with neuromotor disorders by using machine learning models to accurately decode intended speech units directly from brain activity[7–10].

Current speech BCIs, however, face challenges due to their reliance on patient-specific approaches that require resource-intensive, individualized data acquisition from each patient – typically spanning multiple weeks – before their associated machine learning models can be trained to achieve maximal accuracy[8]. This approach is taken to mitigate the effects of varied sensor placement, sparse sampling of the targeted anatomy, and neuroanatomical variability across patients[11]. These long training and deployment times make patient use cumbersome and hinder their adoption as assistive communication devices. Future speech BCIs that successfully combine data across multiple patients may accelerate training by requiring minimal data collection from new patients. Such cross-patient speech BCIs could be rapidly deployable and demonstrate greater robustness across a broader patient population.

For successful cross-patient speech decoding, there must be similarity in the neural representations of speech production across patient recordings. Speech decoding studies typically record from the sensorimotor cortex (SMC)[12–14] with intracranial electroencephalographic devices, such as electrocorticography (ECoG)[15–18] or intracortical microelectrode arrays[19,20], from which sampled representations of speech production may differ from patient to patient due to variability in electrode array placement or individual neuroanatomy. However, recent work has demonstrated that for motor activity, there is a low-dimensional structure in the time-varying neural population activity of the motor cortex[21–23], known as latent dynamics, that is shared across individuals. For example, in non-human primates, the latent dynamics of individual subjects

performing limb movement can be aligned and pooled across animals to train a cross-animal movement decoding model[24]. We hypothesize that this phenomenon extends to human speech production. That is, the latent dynamics underlying the motor control of articulator muscles in the vocal tract to produce speech have a shared representation across separate patients despite variability in electrode placement and individual neuroanatomy (Fig. 1a). These latent dynamics can be aligned and pooled across patients to enable training of cross-patient speech BCIs.

We additionally hypothesize that the elucidation of shared latent dynamics requires an accurate sampling of speech-relevant neural population activity. We have previously shown that accessing the rich spatiotemporal representations of speech production in the SMC requires precise neural interfaces with high spatial-specificity and broad coverage. This dense, broad sampling of the cortex has been achieved using micro-electrocorticographic (μECoG) arrays to record from a large surface of the brain with a high density of micro-contacts[25], which has allowed resolution of micro-scale features for both epilepsy[26] and speech production[27]. These results suggest that μECoG has sufficient spatial-specificity and coverage to enable successful cross-patient alignment for speech BCIs.

Here, we investigate cross-patient representations of speech production by performing acute neural recordings with high-density μECoG arrays over the SMC of eight patients during a speech repetition task. We use dimensionality reduction techniques to reduce high channel-count μECoG recordings into low-dimensional latent dynamics, and we subsequently learn linear transformations that align these latent dynamics across patients. We show that cross-patient latent dynamics encode speech information relevant to the control of articulator muscles in the vocal tract. These shared speech encodings enable the training of cross-patient speech decoding models, which in turn outperform patient-specific decoding models. Lastly, we show that micro-scale sampling of the SMC with broad coverage is critical for learning the proper alignment between patients that preserves speech information.

Our results demonstrate improved speech decoding for BCIs facilitated by a shared, cross-patient representation of speech production encoded in the latent dynamics representing SMC population activity. These findings enable future speech BCIs with accelerated deployment times and increased generalizability, resulting in improved quality of life for individuals who have lost the ability to speak due to neuromotor disorders.

Results

We recorded speech-relevant neural activity from the surface of the SMC using high-density, high channel-count μECoG arrays in eight patients undergoing awake neurosurgical procedures (four with 128 channels – 1.33 mm spacing, four with 256 channels – 1.72 mm spacing, see Table S1 for summary of patient information). Fig. 1b shows the electrode array locations of all patients projected to an average brain (see Fig. S1 for array locations on patient-specific brains). Recordings were performed intraoperatively while patients were instructed to repeat non-word sequences of three phonemes, or sound units (e.g. */b/* in bat), immediately after hearing an

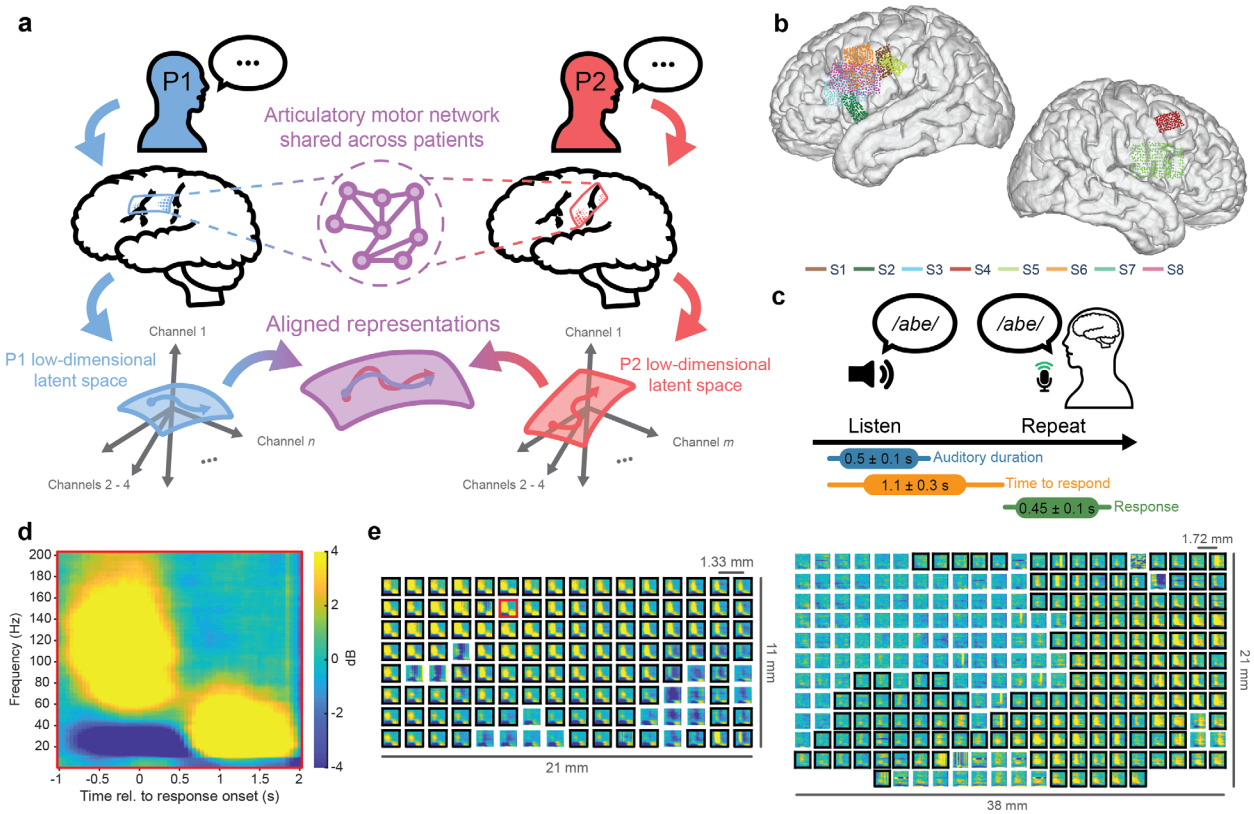

Figure 1. Shared latent dynamics underlying speech production. a. Despite differences in electrode placement and individual neuroanatomy, we propose that the neural population activity that controls articulatory muscles for speech production is shared across separate patients and can be elucidated by aligning their latent dynamics to a common space. **b.** Neuroanatomical location of μECoG arrays for all patients projected onto a common brain (MNI-152, see Fig. S1 for patient-specific array locations). **c.** Speech repetition task. Patients are presented with an auditory non-word stimulus that is composed of three phonemes (e.g. */abe/*) and are instructed to immediately repeat this phoneme sequence. **d.** Example spectrogram showing high-gamma (HG; 70-150 Hz) power centered around the response onset ($t = 0$). **e.** Spectrograms across 128-channel (left, S1) and 256-channel (right, S3) μECoG arrays. Black borders indicate channels with significant HG power during patient response relative to a pre-stimulus baseline (FDR-corrected non-parametric permutation tests). The red border indicates the example spectrogram from **d**.

auditory presentation of the non-word (Fig. 1c, see Methods and Table S2). Using spectrotemporal analysis of the recorded neural signals, we observed neural activations around patient response onset (Fig. 1d). Specifically, we found strong increases in power in the high-gamma (HG; 70-150 Hz) band, which has been shown to be a robust marker of speech production in the SMC[8,17,28] that correlates well with multi-unit neural firing[29–31]. Significant HG activations during patient responses (relative to a pre-stimulus baseline, FDR-corrected non-parametric permutation test – see Methods) were observed in many electrodes across the entirety of each μECoG array (significant electrode amounts in Table S3), exhibiting unique spatial activation patterns (Fig. 1e, all patients in Fig. S2).

Alignment of cross-patient latent dynamics

Speech trials for each patient were grouped into four categories representing types of articulation corresponding to the phonemes presented in our stimuli (low: */a/, /e/*, high: */i/, /u/*, labial: */p/, /b/, /v/,* dorsal: */k/, /g/*). There were distinct spatiotemporal patterns of HG activity specific to each type

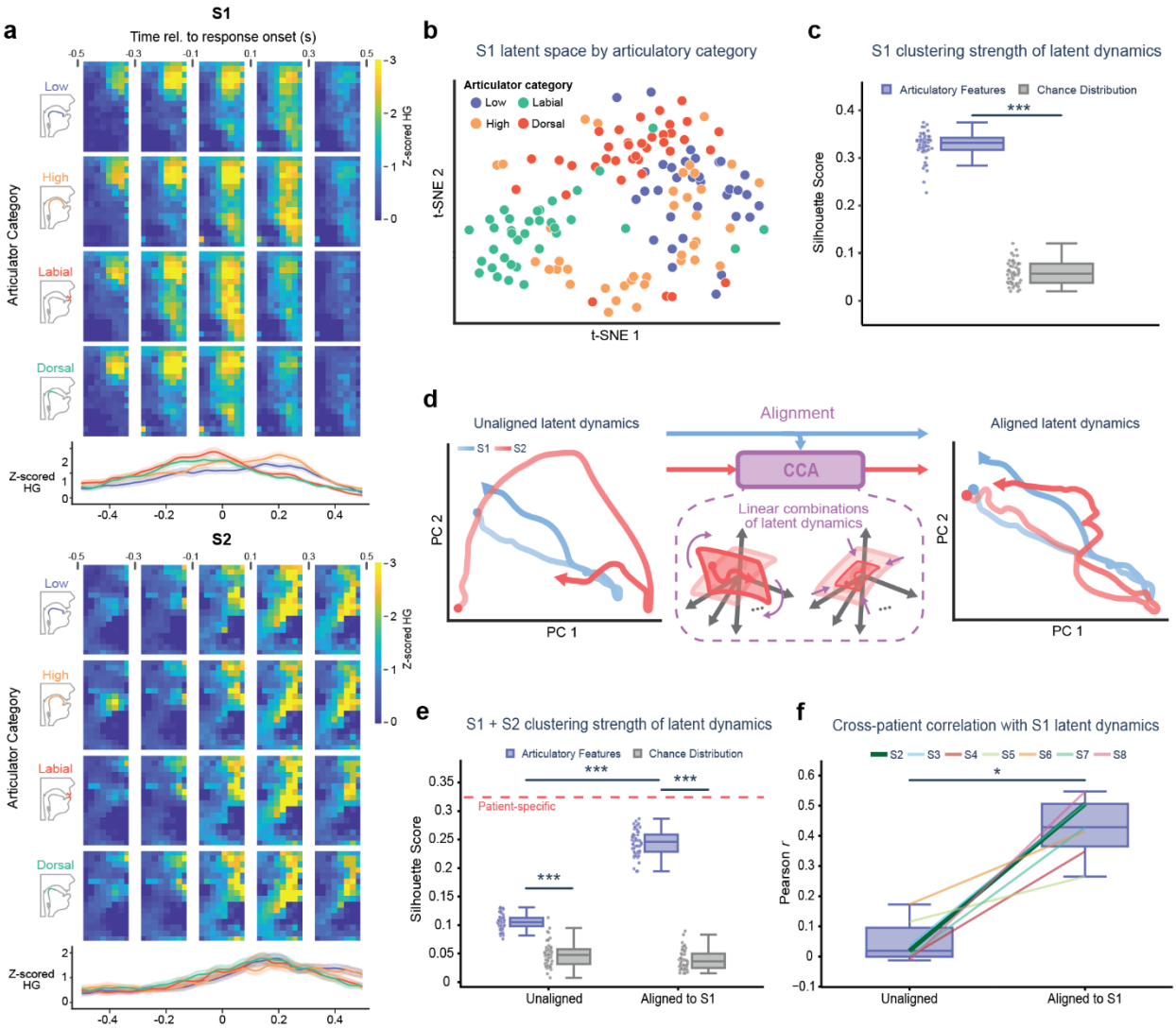

Figure 2. Alignment of latent dynamics across patients. a. Spatiotemporal activation patterns of HG neural activity (z-scored relative to pre-stimulus baseline) across the μECoG arrays for S1 and S2 grouped by different articulatory gestures. Vocal tract diagrams show visual representations of articulatory category (low, high, labial, dorsal). HG traces below the heatmaps show average activity across all channels for each articulator category. The spatiotemporal HG patterns of these articulatory gestures are distinct across patients, motivating a need for alignment for cross-patient grouping. **b.** Latent dynamics are extracted from HG activity using principal component analysis (PCA). Projection of S1 latent dynamics to a 2D space as shown by t-distributed stochastic neighbor embedding (t-SNE) which qualitatively demonstrates the clustering of latent dynamics by articulator type. **c.** Quantification of articulatory clustering strength in **b** by the silhouette score of points in t-SNE space. The silhouette score for latent dynamics is significantly higher than that of a chance distribution generated by shuffling the ground truth articulator labels (***$p < 0.001$, Mann-Whitney U test). **d.** Latent dynamical trajectories for S1 and S2, as visualized in 2D PC space. The trajectories appear dissimilar. Alignment of S2 latent dynamics to S1 is performed with canonical correlation analysis (CCA). Resultant aligned latent dynamics appear more visually similar. **e.** As in **c**, but with data from both S1 and S2. Without alignment, the latent dynamics articulatory clustering strength is above chance ($p < 0.001$, Mann-Whitney U test) but much lower than the patient-specific S1 mean. With alignment, the articulatory clustering strength is significantly higher than both the unaligned case ($p < 0.001$, Mann-Whitney U test) and chance ($p < 0.001$, Mann-Whitney U test), and much closer to the patient-specific S1 mean. **f,** Across all patients, CCA alignment to S1 improves the Pearson correlation between the latent dynamics of each patient and S1 relative to no alignment (*$p < 0.05$, Wilcoxon signed-rank test).

of articulation (Fig. 2a, all patients in Fig. S3). While patterns differ across articulators, they also differ across patients. This can be seen across all patients, but we illustrate this difference for two representative patients (S1 and S2 – Fig. 2a). These patient-specific differences in spatiotemporal

speech representations demonstrates the necessity of specialized alignment techniques in uncovering a shared representation of speech across patients.

To extract latent dynamics from these high-dimensional data, we use principal component analysis (PCA) as a linear dimensionality reduction tool to decompose HG activity (± 500 ms around speech onset) across the significant channels into a low-dimensional set of principal components (PCs) while preserving timing information. Articulatory information encoded in the latent dynamics can be visualized by further projecting the latent dynamics into a two-dimensional space using t-distributed Stochastic Neighbor Embedding (t-SNE, see Methods), where separation between clusters for different articulator types was qualitatively apparent (shown here for S1 – Fig. 2b). Clustering strength of the latent dynamics in this space was quantified using the silhouette score (for additional clustering strength metrics, see Fig. S4), which showed significantly better separation of articulatory feature clusters than in a chance distribution where ground-truth cluster identities were shuffled (Fig. 2c; Mann-Whitney U test, $U = 2500.0$, $n_1 = n_2 = 50$, $p = 7.07e^{-18}$). This shows that PCA successfully extracts latent dynamics that are informative to speech.

The latent dynamics showed distinct trajectories through their respective latent spaces, as visualized in two-dimensional PC space (S1 and S2: Fig. 2d, left). We sought to align these latent dynamics to a similar space to train speech decoding models that learn a similar representation of speech production from multiple patient sources. To model this alignment, we used canonical correlation analysis (CCA) to identify linear combinations of patient-specific latent dynamics such that they are maximally correlated across patients. This resulted in patient-aligned latent dynamics that show more similar trajectories through a common latent space (shown for S1 and S2: Fig. 2d, right). Repeating our previous technique of clustering strength on latent dynamics, we found weaker across-patient clustering when these latent dynamics were unaligned relative to patient-specific clustering, indicating differences in articulator representations across patients (Fig. 2e). With alignment, clustering strength was significantly higher than when unaligned (Mann-Whitney U test, $U = 2500$, $n_1 = n_2 = 50$, FDR-corrected $p = 1.06e^{-17}$) and much closer to patient-specific values, showing that CCA alignment can successfully map latent dynamical representations of speech across different patients. We found that alignment of all other patients to a representative patient (S1) significantly improved the similarity between them (Wilcoxon signed-rank test, $W = 0$, $n = 8$, $p = 0.02$), as measured by the Pearson correlation between their latent dynamics (Fig. 2f), confirming that shared latent dynamics could be elucidated across all patients. We again found that speech information was important to this alignment, as shuffling the ground-truth speech labels prior to alignment (see Methods) resulted in significantly lower similarity than standard alignment (Supp Fig. 5; Wilcoxon signed-rank test, $W = 0$, $n = 8$, $p = 0.02$). These results demonstrate the success of CCA in aligning latent dynamics across patients to elucidate a shared representation of speech production.

Cross-patient alignment preserves spatial articulator maps

Previous work leveraging the high spatial resolution of μECoG showed micro-scale patterns of spatial tuning to different articulator types[27]. As the alignment of latent dynamics showed that articulator representations could be co-aligned temporally across patients, we hypothesized that

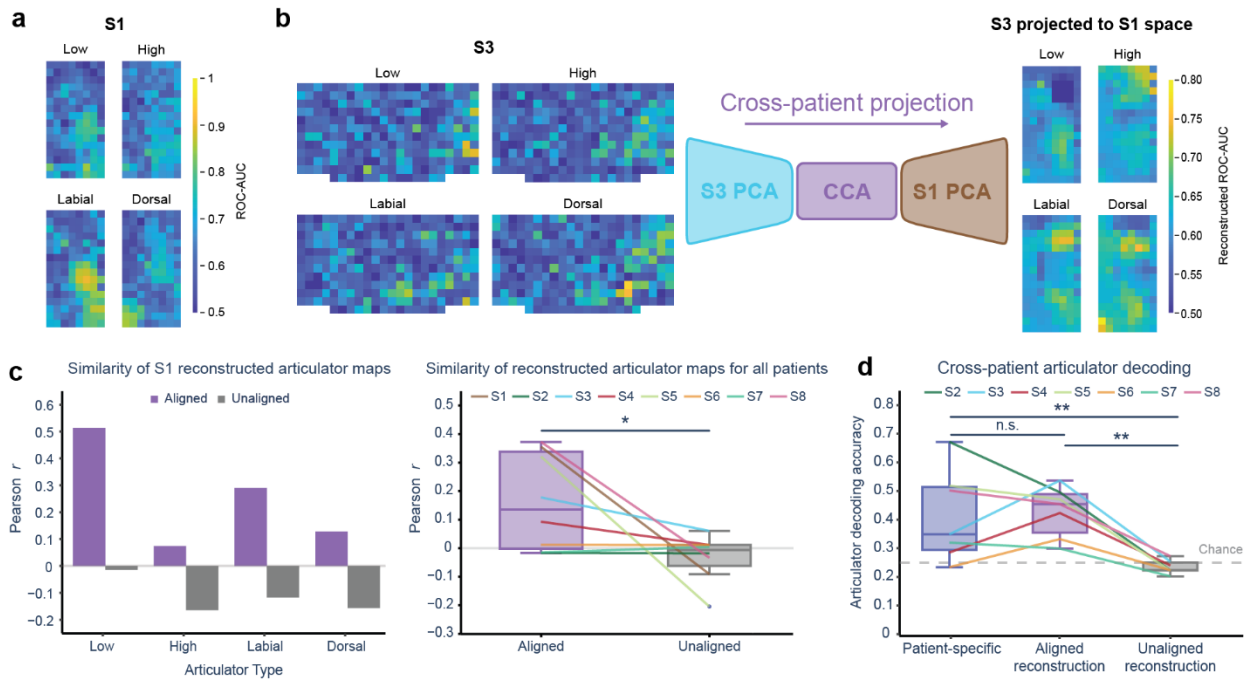

Figure 3. Cross-patient alignment preserves spatial representations of articulatory features. a. Spatial map of electrode tuning to different articulator categories for S1 (128 channels), as quantified by the ROC-AUC score from univariate electrode decoding. Different articulators have distinct spatial representations over the array. **b.** Same as in **a**, but for patient S3 (256 channels). Cross-patient projection enabled reconstruction of a source patient's (S3) articulator tuning in a target patient's (S1) space. **c.** (Left) Quantification of Pearson correlation between reconstructed articulator maps from source patient S3 and ground-truth articulator maps for target patient S1. Reconstructed maps are more similar when using CCA alignment than without alignment across all articulator types. (Right) Similarity in articulator map reconstructions for all target patients (n = 8, each dot is a target patient). For each target patient, reconstructed maps were averaged across all source patients within articulator type to produce a resultant map to compare against the ground-truth articulator map. Reported similarity values for each target patient are the average similarity across all four articulator types. Similarity in reconstructed maps from aligned latent dynamics are significantly higher than those without alignment (***$p < 0.001$, Wilcoxon signed-rank test). **d.** Articulator decoding from cross-patient data. Support vector machine (SVM) decoding models are trained to classify articulator type from latent dynamics. There is no significant difference in accuracy between patient-specific SVMs and SVMs trained on S1's data then tested on data from other patients that have been aligned and reconstructed to S1's space ($p = 0.79$, exact two-sided paired permutation test). Without alignment, cross-patient decoding accuracy is significantly lower (**$p < 0.01$, exact two-sided paired permutation test).

spatial tuning to articulatory features would also be preserved following alignment. To investigate this, we first calculated single-electrode tuning to the four articulator categories. For each electrode, we trained a linear discriminant analysis decoding model to classify articulator type. We quantified the tuning of an electrode to an articulator by its receiver-operator-characteristic area-under-the-curve (ROC-AUC) value. Fig. 3a shows the spatial arrangement of electrode-articulator tunings for S1, where distinct spatial maps for each articulator type are visible.

To investigate the cross-patient preservation of these maps, we used CCA alignment to project spatial articulator information from one patient's space to another. We first used PCA to extract latent dynamics from two patients and learned a CCA transformation to align them. The transformations learned in this process enable the projection of data from a source patient's "electrode-space" to a target patient's electrode-space, which we call cross-patient projection. Specifically, electrode-level data can be projected to the source patient's latent space with the source patient's PCA transform, (optionally) aligned to the target patient's latent space, and then

projected to the target patient's electrode-space by applying the inverse of target patient's PCA transform. We illustrate this analysis with S3 as the source patient and S1 as the target patient (Fig. 3b). After learning the PCA and CCA transformations, the electrode-articulator tunings for the source patient (Fig. 3b, left) were projected to the target patient's space. We denote these projected spatial maps as reconstructed articulator maps (Fig. 3b, right). The Pearson correlation between ground-truth articulator maps and reconstructed articulator maps was calculated to quantify similarity in the reconstructed spatial tunings. We found that the similarity between reconstructed and target articulator maps was higher when cross-patient projection was performed with alignment than without (Fig. 3c, left). Additionally, we found that the target patient's articulator maps could be reconstructed with high similarity by averaging reconstructed maps across all source patients (Fig. S6 – illustrated for target patient S1), showing that articulator information in one patient is well-represented across multiple patient sources. Across all target patients, we found significantly higher similarity between reconstructed and ground-truth maps with alignment (Fig. 3c, right) than without (Wilcoxon signed-rank test, $W = 3$, $n = 8$, $p = 0.04$), indicating that cross-patient alignment results in improved preservation of spatial articulator tuning.

We validated the preservation of articulatory information across patients by training support vector machine (SVM) decoding models to classify articulatory features from patients that had not been used for training. Specifically, we trained a singular SVM on HG signals (± 500 ms around speech onset) from one patient and evaluated its accuracy when tested on all other patients projected to the single patient's space, both with alignment and without (Fig. 3d – illustrated for S1). As a baseline, we compared cross-patient accuracies to the accuracy of patient-specific SVMs for each patient. We found both patient-specific and cross-patient decoding on aligned reconstructed data yielded above chance articulator decoding accuracies (chance = 0.25; mean: patient-specific = 0.39, Wilcoxon signed-rank test, W = 1, $n = 8$, $p = 0.047$; mean: aligned reconstruction = 0.43, Wilcoxon signed-rank test, W = 0, $n = 8$, $p = 0.047$). Additionally, we found no significant difference between patient-specific and aligned cross-patient decoding accuracies ($n = 8$, $p = 0.79$, exact two-sided paired permutation test), indicating that the aligned cross-patient SVM performs just as well as the patient-specific SVM despite the separate patient sources. Without alignment, there was a significant decrease in cross-patient decoding accuracy relative to the aligned context (mean unaligned reconstruction = 0.23; $n = 8$, $p = 0.002$, exact two-sided paired permutation test), which was expected due to the dissimilarity found between reconstructed articulator maps without alignment and ground-truth articulator maps. There was no significant difference between unaligned cross-patient decoding accuracies and chance accuracy (chance = 0.25; Wilcoxon signed-rank test, W = 4, $n = 8$, $p = 0.11$), demonstrating the necessity of alignment in elucidating shared neural representations of speech. Overall, we have demonstrated that cross-patient alignment preserves important spatial characteristics of articulatory features in the SMC, enabling successful cross-patient speech decoding.

Expanded cross-patient datasets improve speech decoding accuracy

Given the preservation of speech information in aligned, cross-patient latent dynamics, we investigated if cross-patient datasets could be used to address data limitations in speech BCI training and improve their performance. With a selected target patient, additional data from other

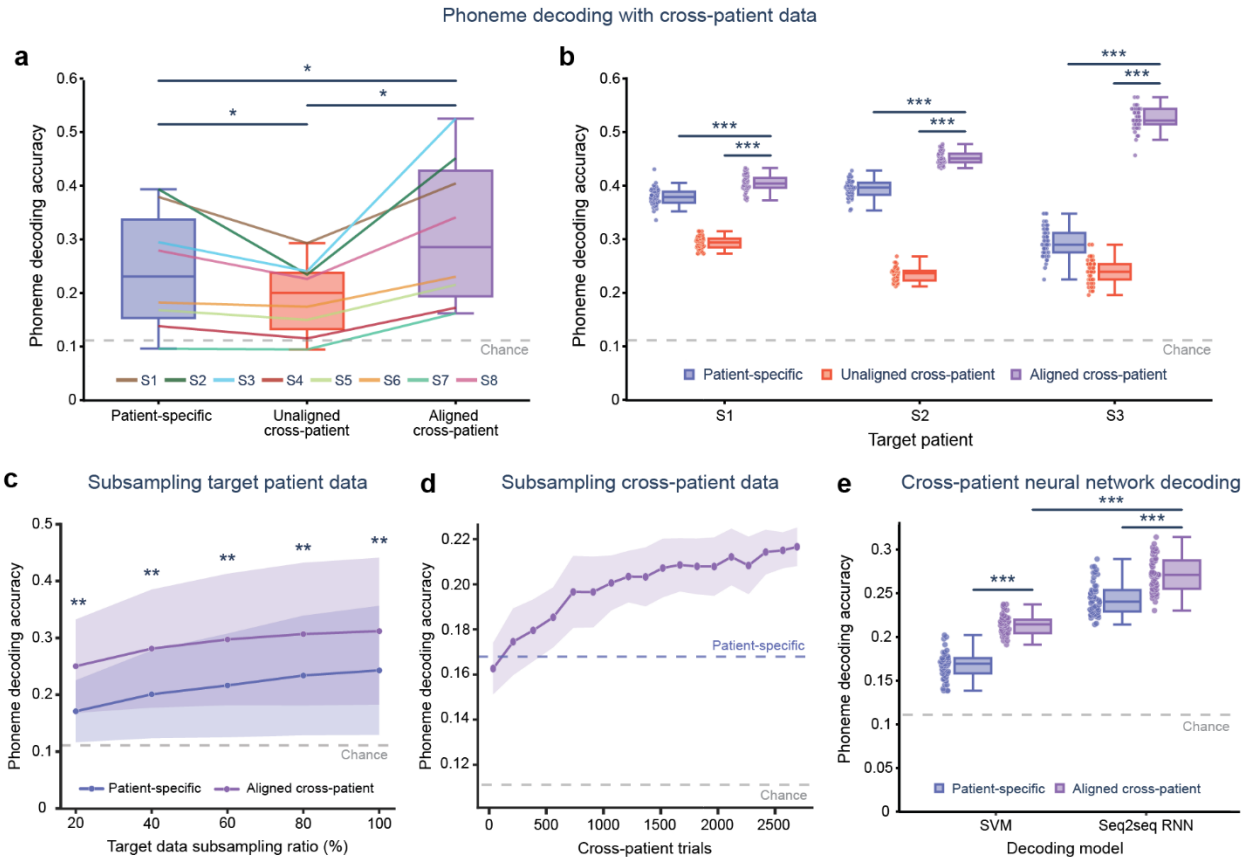

Figure 4. Expanded cross-patient datasets improve speech decoding accuracy. a, Cross-patient phoneme decoding across all eight patients. Each line represents cross-patient decoding aligned to a specific target patient's space. In blue, SVM models are trained and tested on patient-specific data. In red, SVM models are trained on latent dynamics from all eight patients without alignment and tested on held-out patient-specific data from the target patient, resulting in a significant decrease in decoding accuracy (*$p < 0.05$, repeated measures ANOVA and follow-up FDR-corrected paired t-tests). In purple, SVM models are trained on latent dynamics from all patients with CCA alignment and tested on held-out patient specific data from the target patient, improving phoneme decoding accuracy relative to unaligned and patient-specific cases ($p < 0.05$, repeated measures ANOVA and follow-up FDR-corrected paired t-tests). The gray dashed line indicates chance levels of decoding. **b,** Cross-patient decoding for target patients S1, S2, and S3. Aligned cross-patient accuracies were significantly higher than both unaligned and patient-specific decoding contexts in each target patient (***$p < 0.001$, post-hoc Tukey HSD test following one-way ANOVA). **c,** Patient-specific and aligned cross-patient decoding repeated as the percentage of data (trials) in the target patient dataset is subsampled. Even with limited target patient data to learn latent dynamics alignment, the cross-patient decoding outperformed patient-specific decoding (**$p < 0.01$ at all subsampling ratios, exact two-sided permutation test). **d,** Aligned cross-patient decoding as the amount of additional cross-patient data (trials) is varied for patient S5. Little additional data was needed to increase the phoneme decoding accuracy beyond patient-specific levels. Diminishing gains in decoding accuracy are observed as additional cross-patient data is added. **e,** Cross-patient decoding with a seq2seq RNN model for S5. As with the SVMs, using aligned cross-patient data resulted in a significant increase in phoneme decoding accuracy ($p < 0.001$, post-hoc Tukey HSD test following two-way ANOVA,). The cross-patient RNN significantly outperformed the cross-patient SVM ($p < 0.001$, post-hoc Tukey HSD test following two-way ANOVA).

patients can be aligned to this target patient and appended to the target datasets, enabling vastly expanded training datasets relative to patient-specific contexts. To test this, we trained SVM models on aligned cross-patient latent dynamics during speech (total combined dataset size across all patients: ~8.2 minutes utterance duration, ~68.5 minutes experiment duration, see Table S3 for amount of speech data per patient) and we evaluated their accuracy on a held-out set from the target patient using 20-fold cross-validation. We then compared these results to patient-specific latent dynamics and unaligned cross-patient latent dynamics. While thus far, our

analyses focused on decoding articulatory features, here we additionally evaluated the accuracy of classifying distinct phoneme units present within these articulator categories (e.g. /b/, /a/, /k/), to mimic translational BCI use cases where phonemes are rapidly decoded and strung together into words and sentences.

We found significant differences in decoding accuracy across patient-specific, unaligned cross-patient, and aligned cross-patient decoding contexts (Fig. 4a; 9-way phoneme decoder, chance = 0.11; repeated measures ANOVA, $F_{2,14}$ = 12.59, p = 0.007). In decoding models trained on unaligned cross-patient latent dynamics, individual phoneme decoding accuracy was significantly lower relative to patient-specific decoding (mean: unaligned cross-patient = 0.19, patient-specific = 0.24, follow-up FDR-corrected paired t-test, corrected p = 0.03). Models trained on aligned cross-patient latent dynamics, however, exhibited strong phoneme decoding accuracy – even outperforming patient-specific models (mean aligned cross-patient = 0.31; follow-up FDR-corrected paired t-test, corrected p = 0.01). Fig. 4b shows the cross-patient decoding accuracies separated by target patient for a few sample patients (results for all target patients in Fig. S7). One patient (S3) in particular received an immense benefit from this additional cross-patient data (mean: aligned cross-patient = 0.53, patient-specific = 0.29), which we attribute to the initial lack of patient-specific data available (S3 had approximately one-third the amount of patient-specific data as the other patients), so the additional data provided critical repetitions to enable sufficient training. To validate that these increased cross-patient decoding accuracies were not due to the alignment procedure artificially inducing structure in cross-patient data that was not originally present, we generated surrogate data that preserved the first and second order statistical structure of our neural data (equal mean and covariance across trials, channels, and time) but was otherwise random (see Methods) and aligned this data to the target patient instead. We hypothesized that this surrogate neural data contained similar statistical structure but lacked speech-relevant latent dynamics and thus would fail to decode speech information from the target patient following alignment to the target patient's space. Models trained on only this aligned surrogate data yielded accuracies with no significant difference from chance across all target patients (Fig. S8; Wilcoxon signed-rank test, W = 8, n = 8, p = 0.20), indicating alignment itself does not artificially yield improved phoneme decoding accuracy. These results show that speech decoding models can learn a shared representation of phoneme information from separate patient sources, substantiating the success of cross-patient speech BCIs.

When aligning the latent dynamics of one patient to those of a new target patient, it is unclear how much new data from the target patient must be collected to enable sufficient alignment to the target patient's latent space. We expect that this alignment can be learned with minimal amounts of data – much less than needed to sufficiently train a complex patient-specific decoding model – as CCA is a linear alignment method. To test this, we repeated our cross-patient decoding analysis while subsampling the number of trials from the target patient (excluding S3 due to a previously mentioned lack of trials) used for alignment and decoder training (Fig. 4c, Fig. S9 for all patients separately). We found that CCA yielded successful alignment that enabled cross-patient speech decoding with significantly higher accuracies than patient-specific contexts as the subsampling ratio decreased (n = 8, p = 0.008 at all subsampling ratios with FDR correction, exact

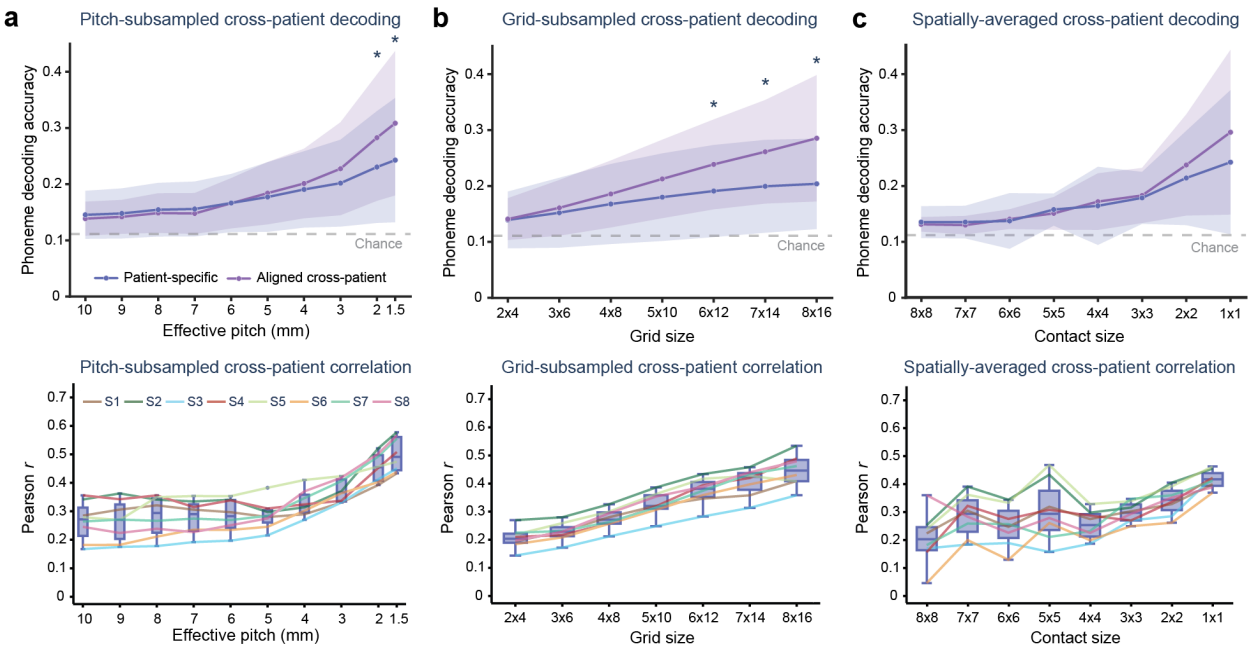

Figure 5. High-resolution sampling with broad coverage is required for successful cross-patient alignment. a. Effect of electrode density. (Top) Cross-patient phoneme decoding with varied pitch (distance between electrodes). Accuracy increases across patient-specific and aligned cross-patient decoding contexts as pitch decreases. Aligned cross-patient accuracies were only significantly higher than patient-specific accuracies for effective pitch less than 3 mm ($p < 0.05$ for pitch ϵ [2, 1.5] mm, exact two-sided paired permutation test, *p<*0.05*). (Bottom) Cross-patient similarity, as measured by the Pearson correlation between latent dynamics increases as the pitch decreases. Each line represents a single target patient, where correlations are averaged across all non-target patients aligned to the target patient. **b.** Effect of array coverage. As in **a,** but for subsampled grids of varying size. (Top) Phoneme decoding accuracy similarly increases as grid size increases. Aligned cross-patient accuracies were only significantly higher than patient-specific accuracies for grid sizes greater than 6x12 (~8 mm x 17 mm; $p < 0.05$ for grid sizes ϵ [6x12, 7x14, 8x16], exact two-sided paired permutation test). (Bottom) Cross-patient similarity increases as grid size increases. **c.** Effect of contact size. As in **a,** but with varying contact sizes obtained by spatially-averaging different amount of neighboring micro-contacts. (Top) Phoneme decoding accuracy increases as contact size decreases. The mean phoneme decoding accuracy of aligned cross-patient models is higher than patient-specific models for contact sizes under 3x3 (~3 mm diameter) however there is no significant difference after correction for multiple comparisons ($p > 0.05$ for all contact sizes, exact two-sided paired permutation test). (Bottom) Cross-patient similarity generally increases as contact size decreases.

two-sided permutation tests), even for target patient datasets with only 20% of their original trials. This demonstrates the feasibility of cross-patient alignment with minimal data collection from new patients.

To quantify the relationship between the amount of additional cross-patient data and decoding performance, we repeatedly trained cross-patient SVMs while varying the amount of additional cross-patient data trials available. For each patient, we randomly subsampled a fixed number of trials equally from each other patient. We then aligned the subsampled trials to the target patient, trained a cross-patient SVM model, and evaluated its performance on a held-out test set from the target patient with 20-fold cross-validation. The results from this analysis for an illustrative patient (S5) are shown in Fig. 4d (all patients in Fig. S10). We found patient-specific performance (mean accuracy = 0.17) was exceeded after only 210 trials (~1.6 minutes utterance duration, ~13.1 minutes experiment duration), with cross-patient decoding accuracy increasing to a mean of ~0.22 (chance = 0.11) with all cross-patient data used. We observed diminishing gains in decoding accuracy as further cross-patient data was added for S5, as shown by a log-fit to the

data in Fig. 4d (r^2 = 0.98; Fig. S11). We hypothesized that a more complex RNN model may be able to better utilize this additional data to achieve higher performance than the simpler SVM models. We therefore trained a sequence-to-sequence (seq2seq) RNN model to sequentially decode spoken phonemes (see Methods, hyperparameters in Table S4) with the same alignment procedure as in previous SVM evaluations (Fig. 4e – illustrated for S5, all patients in Fig. S12). As expected, we found a significant main effect of model type, indicating RNN models outperformed SVM models in phoneme decoding accuracy (two-way ANOVA, $F_{1,196}$ = 830.89, p = $2.05e^{-72}$). Similar to our previous results, we also found models trained on aligned, cross-patient data outperformed patient-specific models ($F_{1,196}$ = 258.51, p = $1.20e^{-37}$). Post-hoc comparisons showed that the seq2seq RNN achieved higher phoneme decoding accuracies when trained on aligned cross-patient data than on patient-specific data (Mean: aligned cross-patient = 0.27, patient-specific = 0.24; p = $4.67e^{-14}$, post-hoc Tukey HSD test), showing the extra cross-patient data is additionally beneficial to complex decoding models used in state-of-the-art speech decoding.

High-resolution and broad coverage are critical to alignment

To fully assess the benefits of high spatial sampling and coverage for cross-patient decoding, we quantified the spatial sampling of the SMC required to sufficiently elucidate speech-relevant latent dynamics that can be aligned across patients. We subsampled various spatial characteristics of our µECoG arrays and calculated both the cross-patient decoding accuracy of SVM models trained on aligned latent dynamics and the cross-patient similarity (Pearson correlation) in the extracted latent dynamics at each subsampled point. We separately investigated the influence of subsampling electrode density, array coverage, and electrode contact size.

To investigate the effects of electrode density, as represented by the distance between electrodes (pitch), we used Poisson disk sampling to randomly sample individual electrodes across the array while constraining the minimum distance between sampled electrodes to approximate the cortical sampling of lower-density ECoG arrays. Simulated pitch ranged from 10 mm, the pitch of standard clinical ECoG devices, to 1.5 mm. To quantify the speech information encoded in the pitch-subsampled data, we repeated our cross-patient phoneme decoding analysis from Fig. 4 at each effective pitch value. We observed increases in phoneme decoding accuracy as pitch decreased (an increase in spatial sampling) in both patient-specific and aligned cross-patient contexts (Fig. 5a, top; Fig. S13 for all patients separately). Crucially, we found that aligned cross-patient models significantly outperformed patient-specific models for only pitch values under 3 mm (n = 8, FDR-corrected p = [0.04, 0.04] for pitch ϵ [2, 1.5] mm, respectively, $p > 0.05$ for all other pitch values, exact two-sided paired permutation test). We additionally found that the correlation between latent dynamics of separate patients increased as the effective pitch decreased across all patients, with more rapid increases for smaller pitch values (Fig. 5a, bottom). These results indicate that sufficient sampling of cross-patient latent dynamics necessitates high-density neural interfaces.

To evaluate the necessity of neural interfaces with broad coverage of the SMC, we selected sub-grids of electrodes from our full electrode arrays and repeated the analyses in Fig. 5a. Grid sizes ranged from 2x4 (~2 mm x 5 mm) to 8x16 electrodes (~11 mm x 23 mm). Grid-subsampled phoneme decoding gradually increased as coverage increased for both decoding contexts (Fig.

5b, top; Fig. S14 for all patients separately). Aligned cross-patient models had significantly higher accuracies than patient-specific contexts for only grid sizes 6x12 and larger (n = 8, FDR-corrected p = [0.04, 0.04, 0.04] for grid sizes ϵ [6x12, 7x14, 8x16], respectively, $p > 0.05$ for all other grid sizes, exact two-sided paired permutation test), which corresponds to coverages greater than ~8 mm x 17 mm, showing that broad coverage of the SMC is critical for successful sampling of cross-patient latent dynamics. Similarly, we found that the correlation between cross-patient latent dynamics increased as the size of the subsampled grids increased across all patients, showing better similarity with broader coverage (Fig. 5b, bottom).

Finally, we quantified the effects of contact size by extracting HG information from recorded signals averaged together across neighboring electrode contacts to approximate larger contacts. Contact sizes ranged from 8x8 micro-contacts (~11 mm diameter) to a single micro-contact (200 µm). We found that decoding accuracy increases as contact size decreases for both decoding contexts (Fig. 5c, top; Fig. S15 for all patients separately). While the decoding accuracy of cross-patient SVMs was higher than the patient-specific decoding accuracy for contact sizes under 3x3 (~3 mm), there was no significant difference after correction for false discovery rate (n = 8, $p > 0.05$ for all contact sizes, exact two-sided paired permutation test). Correlation in aligned latent dynamics generally increased as contact size decreased (Fig. 5c, bottom). These results suggest that electrode density and array coverage play a more important role in properly sampling SMC cross-patient latent dynamics during speech.

Discussion

Speech BCIs have recently demonstrated impressive performance in rapidly and accurately translating neural signals to words in patients with an inability to speak due to neuromotor disorders. However, these BCIs typically require long periods of patient-specific data collection that limit their deployability and effectiveness. In this work, we investigated shared, cross-patient representations of speech production in the SMC to reduce the deployment time of speech BCIs by enabling cross-patient decoding models. While cross-patient speech decoding has been previously investigated using transfer learning[12], as of this writing, this is the first study, to our knowledge, to use neural activity aligned across human patients for speech decoding. Our results support the idea that cross-patient models can reach sufficient data requirements for advanced decoding methodologies with minimal data collection from any individual patient and can be easily transferred across patients, resulting in robust and rapidly deployable speech BCIs that are more accessible to a broader range of patients.

A novel aspect of this study is the investigation of shared latent dynamics in human speech. Previously, the alignment of latent dynamics across individuals has been constrained to non-human primates and rodents performing limb movement tasks[24]. Here, we show this phenomenon of shared neural activity across individuals extends to the human motor speech system, where the implications for translational BCI contexts stand to provide substantial benefits for training machine learning models and accelerating deployment following device implantation. Despite differences between limb movement and speech production, the presence of shared speech latent dynamics is a logical extension as both movement contexts pertain to the alignment of low-level neural motor representations underlying highly-coordinated muscle movements.

We hypothesized that neural population activity underlying speech articulator movement is well-represented in patients' latent dynamics and also shared across patients. To sample this neural population activity, we analyzed the time-varying HG activity in neural signals recorded with µECoG arrays intraoperatively from eight patients. Previous studies have instead investigated motor-related latent dynamics via multi-unit activity recorded from intracortical microelectrode arrays[21,32], which directly sample single neuron activity in comparison to neural population activity in the HG band of neural local field potentials (LFPs) recorded from surface electrodes[33]. However, strong correlations between neuronal spiking and HG signals are well-documented[31,34], especially in speech production[29], suggesting that latent dynamics extracted from both signals should be similar. More directly, similarity in the latent dynamics extracted from the LFP and multi-unit activity in the same individual has previously been assessed, showing HG latent dynamics are well-correlated with multi-unit latent dynamics[35]. As such, the latent dynamics analyzed in this study represent valid decompositions of neural population activity underlying speech production.

To extract latent dynamics from HG activity across many channels, we used PCA as a dimensionality reduction method to approximate latent dynamics. Further, to perform alignment between latent dynamics from different patients, we used CCA to learn linear transformations between latent dynamics. We chose these methods for their simple and interpretable transformations, where aligned latent dynamics are the result of spatial filters applied to the original electrode space. However, PCA and CCA are limited because they treat data from different time points as independent observations, which discards potential temporal information that could be useful in learning a better alignment across patients. Recent dynamical systems frameworks have been developed that use deep representation learning to elucidate latent dynamics while explicitly modeling temporal information, such as Latent Factor Analysis for Dynamical Systems (LFADS)[36] and MAnifold Representation Basis LEarning (MARBLE)[37]. Future studies should investigate cross-patient representations of speech production extracted from these advanced dynamical systems models. However, study designs must be wary of limited interpretability, relative to linear decomposition methods, with deep learning models. Additionally, the amount of data per patient required to successfully learn latent dynamics should be considered in these deep models, as steep data requirements to learn and align latent dynamics is antithetical to the desired rapid deployability for translational BCI contexts.

We found that alignment of latent dynamics by CCA preserved spatial properties of speech information which enabled successful cross-patient decoding. Specifically, decoding models trained on additional cross-patient data achieved higher phoneme decoding accuracies than models trained on solely patient-specific data, with particularly impressive improvements in some patients. However, we only combined data across eight patients, with limited data per patient. We expect that a larger dataset with a more expansive library of patients to align to a given target patient will result in further increases in decoding accuracy, especially in RNN (and other deep learning) decoding models.

By subsampling various spatial characteristics of our µECoG arrays, we found that neural interfaces with both high resolution and broad coverage of the SMC are best suited to align latent dynamics and train cross-patient speech decoding models. Neural interfaces in current speech BCIs[7,8] typically prioritize only one of these characteristics, as intracortical microelectrode arrays

offer high electrode density but limited coverage and ECoG grids offer wider coverage but with limited spatial sampling. However, since cross-patient alignment is only calculated on dimensionality-reduced latent dynamics, alignment could still be successful with a heterogeneity of neural interfaces across patients, which would increase data availability for cross-patient alignment. Nonetheless, our current findings motivate the use of μECoG for cross-patient speech decoding as it enables both dense and broad sampling.

It is important to note that the current study is limited to the analysis of cross-patient latent dynamics during overt speech in able-bodied speakers. These patients differ functionally from individuals with impaired speech function due to neuromotor disorders like ALS or brainstem stroke. It remains unknown if the preservation of speech latent neural dynamics that we observed here (in individuals without speech dysfunction) will be present across individuals with motor system impairments. However, recent speech BCI studies have found that articulatory representations of speech may be well preserved in patients with impaired speech many years after diagnosis or injury with ALS[7] or brainstem stroke[8], respectively, suggesting that the shared articulatory latent dynamics we observed here may indeed be relevant to the translational development of speech BCIs.

Together, our results demonstrate that speech-relevant neural population dynamics are shared across human patients, providing a foundation for more efficient and scalable speech BCI systems. With cross-patient alignment based on latent dynamics, future speech BCIs may achieve faster adaptation with minimal patient-specific data, resulting in improved patient usability and performance. Most importantly, cross-patient speech BCIs will improve communicative ability and quality of life for individuals with paralysis from neuromotor disorders.

Methods

All our research studies were approved by the Institutional Review Board of the Duke University Health System under the following protocol ID:

Pro00072892: Studying human cognition and neurological disorders using μECoG electrodes.

Participants

This study included neural data collected intraoperatively from eight awake patients (mean age = 57.4, one female) with fully intact speech at the Duke University Medical Center (DUMC). Intraoperative recordings were performed while patients S1, S2, S4, and S5 underwent surgical interventions for movement disorders by deep brain stimulator (DBS) implantation and patients S3, S6, S7, and S8 were undergoing brain tumor resection. All subjects were native English speakers. Written informed consent was obtained prior to surgery in accordance with the Institutional Review Board of the Duke University Health System.

Intraoperative neural recording

Intraoperative neural recordings were performed using custom-designed micro-electrocorticographic (μECoG) arrays that were fabricated by Dyconex Micro Systems

Technology (Basserdorf, Switzerland). Recordings were performed during awake neurosurgical procedures for movement disorders (deep brain stimulator implantation – DBS) and tumor resection. Two different μECoG designs were used. First, for DBS implantation surgeries, a narrow (11 mm x 21 mm) μECoG array with 128 micro-contacts (200 μm contact diameter, 1.33 mm center-to-center spacing between contacts) was inserted through a burr hole and placed in subdural space over the target sensorimotor cortex (SMC). For tumor resection surgeries, larger (38 mm x 21 mm) μECoG arrays with 256 micro-contacts (200 μm contact diameter, 1.72 mm center-to-center spacing between contacts) were placed directly over the SMC through a craniotomy window. Arrays were sterilized with ethylene oxide gas prior to surgical implantation. μECoG arrays were connected to custom headstages using Intan Technologies RHD chips to amplify and digitize recorded neural signals. These headstages were then connected to an Intan Technologies RHD recording controller via micro high-definition multimedia interface (μHDMI) cables. Recorded neural data were processed by the Intan RHX v3.0 data acquisition software (Intan Technologies, Inc.). Recordings were analog filtered between either 0.1 Hz and 7,500 Hz or 0.1 Hz and 10,000 Hz. Recordings were digitized at 20 ksps.

To visualize μECoG array location for each patient, electrode locations were co-registered with patient cortical reconstructions that were generated from preoperative MRI images using Freesurfer's *recon-all-clinical* function[38]. For DBS implantation surgeries, array location was confirmed with an intraoperative CT scan following recording that was aligned to the reconstructed preoperative MRI volume. Electrodes were localized from the CT scan using BioImage Suite[39]. For tumor resection surgeries, array location was registered to the preoperative MRI with Brainlab Neuronavigation (Munich, Germany) coordinates marked at array corners. Brainlab coordinates were imported into BioImage Suite, where intermediary electrode locations were interpolated to match the known electrode spacing of the electrode array (due to missing Brainlab coordinates for subject S8, electrode location was manually determined based on intraoperative images pre- and post-array placement). In both DBS and tumor resection cases, electrode locations were subsequently projected to the leptomeningeal surface of the brain. Electrode arrays of all patients were visualized together by projection to a common brain (MNI-152, Fig. 1) and separately on patient-specific brains (Fig. S1).

Speech task

Patients performed a speech repetition task in the acute intraoperative neural recording setting. In a single trial, patients are presented with an auditory stimulus representing one of 52 unique non-words via a computer speaker and asked to verbally repeat the sounds they hear. Non-words were sequences of three phonemes designed to sample the phonotactics of American English. Sequences were composed of three of nine possible phonemes (four vowels: */a/, /e/, /i/, /u/* and five consonants: */b/, /p/, /v/, /g/, /k/*) arranged in consonant-vowel-consonant (e.g. */b/, /a/, /k/*) or vowel-consonant-vowel (e.g. */a/, /b/, /e/)* format (See Table S2 for all stimuli). Trials lasted were composed of stimulus presentation (stimulus duration: ~0.5 s) followed by a three second response window with 250 ms jitter in between trials. Trials were presented across three consecutive blocks (52 trials per block for each unique non-word token). Each block corresponded to ~0.5 min of utterance duration. Stimuli order was randomly shuffled within blocks. Trial start times were marked using flashing lights recorded by a photodiode on the task presentation laptop

(Thorlabs, Inc.). Patient responses were recorded with a lavalier microphone (Movophoto, Inc.) connected to a pre-amplifier (Behringer). Both photodiode and microphone data were sent via a BNC cable/connector to auxiliary channels on the Intan RHD recording controller for synchronization with recorded neural data at 20 ksps. The speech task was designed using Psychtoolbox in MATLAB R2014a. Phoneme labels (identity, start time, stop time) for both auditory stimuli and patients responses were first automatically generated using the Montreal Forced Aligner[40] and then manually corrected using the Audacity audio-editing software[41].

Neural data preprocessing

Recorded neural data was decimated to 2 ksps using a finite impulse response anti-aliasing filter (Kaiser window, MATLAB's *resample* function) and common-averaged-referenced (CAR) to remove common noise across channels. Electrodes with large impedance values (>1 Mohm) were excluded from CAR. After CAR, trials with large transient activity (>50 μV between timepoints) were discarded. All preprocessing was done in MATLAB R2021a.

Spectrogram visualization

Spectrograms were calculated for each electrode to visualize spectrotemporal neural activity during patient responses (Fig. 1d-e). Spectral information was extracted using 500 ms windows with a 50 ms step-size and 10 Hz frequency smoothing using a multi-taper time-frequency analysis method[42]. Spectrograms during patient responses were normalized by a baseline period 500 ms prior to auditory stimulus presentation by dividing response spectrograms by the average baseline power-spectral density to remove *1/f* broadband activity. Normalized response spectrograms were averaged across trials for each electrode and the log-power was calculated around patient response onset (-1 to 2 s) to yield neural activity in units of decibels.

High-gamma extraction

The high-gamma band (HG: 70-150 Hz) of the neural signal was extracted to estimate neural population dynamics in the SMC. HG signals have previously been shown to encode rich speech production information[8,17,28] and strong correlations with neuron spiking activity[29–31], suggesting a valid estimation of speech-relevant SMC population dynamics. Following preprocessing, neural signals were bandpass filtered using frequency-domain Gaussian filters centered at eight logarithmically spaced center frequencies (with logarithmically scaled bandwidths around each center frequency) between 70 and 150 Hz[43]. Specifically, the neural timeseries were decomposed into their corresponding frequency representations with the fast Fourier transform (FFT) and multiplied in frequency space with a Gaussian kernel centered around each center frequency to select and smooth information from the frequencies in the desired band. The inverse FFT was then performed to convert bandpass filtered signals at each band back to the time domain. Envelopes of signals in each band were calculated as the absolute value of the Hilbert transform and subsequently averaged across the eight bands. The resultant timeseries was then downsampled to 200 Hz with a finite impulse response anti-aliasing filter (Kaiser window, MATLAB's *resample* function). For visualization of HG activity across arrays (Fig. 2a), HG activity was z-score normalized to the baseline period (500 ms prior to auditory stimulus presentation) for

each trial. For all other analyses (latent dynamics extraction, decoding), HG activity was normalized by subtracting the average HG envelope of the baseline period. Analyses in this work used HG activity in a one second window (-500 ms to 500 ms) around patient response onset. Channels with significant speech HG activity were determined by statistical comparison between response and baseline periods (see *Statistical Analysis*). HG extraction was performed in MATLAB R2021a.

Latent dynamics extraction

We extracted latent dynamical activity from estimated SMC population activity as measured in HG activity around response onset for each electrode and trial (data array of the dimensions – *s, t, n*), where *s* is the number of trials, *t* is the number of timepoints, and *n* is the number of significant electrodes/channels). We used principal component analysis (PCA, scikit-learn[44]) to reduce the high-dimensional channel space into a low-dimensional space composed of *k* principal components (PCs), where each PC represented a linear combination of activity across the *n* channels. *k* was determined as the number of components explaining 90% of the variance in the original data. This yielded patient-specific latent dynamics (data array of the dimensions – *s*, *t, k*) as the projection of the original HG activity on the *k*-dimensional latent space. Extraction of latent dynamics was performed in Python 3.10.

Clustering of articulatory features in the latent space was visualized by further projecting latent dynamics into a two-dimensional subspace (Fig. 2b) using t-distributed Stochastic Neighbor Embedding (t-SNE, scikit-learn, perplexity = 30), a widely used technique for visualization of high-dimensional data[45]. t-SNE decomposition of latent dynamics yielded a data array of dimensions (*s*, 2). Visualization was performed by plotting each trial as a point in 2D t-SNE space and coloring each trial point by its articulatory category. Quantification of clustering strength was performed by calculating the Silhouette Coefficient (scikit-learn, distance = Euclidean) for each trial in t-SNE space, which measures the difference between inter- and intra-cluster distances[46]. The score is more positive when the inter-cluster distance is greater than intra-cluster distance. The reported silhouette score was calculated as the average of all positive-valued Silhouette Coefficient trials[27]. A distribution was generated by calculating silhouette scores across many iterations (n = 50) of projection to 2D t-SNE space. A chance distribution was generated by repeating the silhouette score analysis for latent dynamics with randomly permuted articulator labels. Clustering with additional metrics is shown in Fig. S4.

Alignment of latent dynamics

Single unaligned latent dynamic trajectories were visualized for S1 and S2 by plotting the first two PC values at each timepoint during the speech utterance window, averaged across all trials (Fig. 2c). Alignment of these latent dynamics was performed using canonical correlation analysis (CCA) to find linear combinations of the latent dynamics PCs that resulted in maximal correlation between the latent dynamics of the two patients. CCA alignment is performed in a similar fashion as in[23], which is described below and generally follows MATLAB's *canoncorr* function. Alignment of latent dynamics was performed in Python 3.10 with numpy.

Starting with the latent dynamics of patients S1 (s_1, t, k_1) and S2 (s_2, t, k_2), we require correspondence between trials s_1 and s_2 across these patients. As patients may be missing different trials due to experimental noise or outlier removal, we enforced uniformity in trials by condition averaging the trials into categories representing the unique phoneme sequences (s_{ca}, at most 52 unique sequences) presented during the task and sub-selecting only the phoneme sequences present in both patients (i.e. if S1 has no trials for the sequences */abe/*, trials for */abe/* in S2 will not be considered for alignment). We chose to condition average by phoneme sequence to enforce that the same phoneme sequences should have similar latent dynamics trajectories.

Condition-averaged trials were folded together with the time dimension to create latent dynamics matrices L_1, L_2 of dimensions ($s_{ca}*t$, k_1) and ($s_{ca}*t$, k_2), respectively. Let d be the minimum rank of L_1 and L_2. CCA alignment first centers the L_1 and L_2 by mean subtraction. Following, QR decompositions are calculated for both L_1 and L_2 as

$$L_1 = Q_1 R_1,$$
$$L_2 = Q_2 R_2\ ,$$

where Q is an orthonormal basis for the corresponding latent dynamics. Correlations between the orthonormal bases are represented by the inner product matrix $Q_1^T Q_2$. Singular value decomposition on this matrix yields

$$Q_1^T Q_2 = USV^T,$$

where U and V are the left and right singular vectors, respectively, which define the transformations of Q_1 and Q_2 to a shared, cross-patient space with maximal pairwise correlation. The U and V matrices (truncated to sizes (k_1, d) and (k_2, d), respectively) can be better understood in the context

$$M_1 = R_1^{-1} U,$$
$$M_2 = R_2^{-1} V,$$

where M_1 and M_2 are projection matrices mapping L_1 and L_2 to the shared, cross-patient space, respectively. Transformation is performed via the matrix multiplication

$$L_1^a = L_1 M_1,$$
$$L_2^a = L_2 M_2,$$
$$L_{2\to 1}^a = L_2 M_2 M_1^{-1},$$

where L_1^a *and* L_2^a are the aligned latent dynamics of S1 and S2, respectively, where the canonical correlations between these are the diagonal elements of S above. L_1^a *and* L_2^a are aligned in a cross-patient space that requires transformations of both input patients. In this work, we instead choose to align patients to the space of a desired target patient to enable alignment from any number of target patients to a single common space. The alignment of S2's latent dynamics to S1's space, denoted $L_{2\to 1}^a$, is calculated by first aligning S2 to the common cross-patient space and applying the inverse of S1's projection matrix (M_1) to map to S1's space. As such, multiple

patients can be aligned to S1's space by first calculating pairwise CCA alignment and projecting all patients to S1's target space.

The clustering analysis with silhouette score quantification mentioned above was repeated on a dataset composed of both S1's and S2's data (Fig. 2d). Without alignment, latent dynamics were separately extracted from S1 and S2, concatenated (truncating the number of PCs in the patient with higher-dimensional latent dynamics to ensure consistency in dimensions for concatenation), and reduced to a 2D space with t-SNE. With alignment, S2 latent dynamics were aligned to S1's space, aligned latent dynamics were concatenated, and t-SNE decomposition to a 2D space was applied. In both cases, silhouette score quantification was identical to the patient-specific case.

Alignment of multiple patients to S1 was performed using the method described above to determine how the Pearson correlation between S1's and each other patient's latent dynamics differs with alignment (Fig. 2e). To calculate correlation between the latent dynamics of two patients, we first condition-averaged the latent dynamics into groups by articulator type at the beginning of the utterance (low, high, labial, dorsal). Then, for each articulator type and each of the top 10 PCs, we calculated the Pearson correlation coefficient (scipy) between the selected PC across patients. The average correlation across articulator type and PC was reported as single value per patient. Pearson correlation was calculated between both unaligned and aligned latent dynamics. Additionally, to show that grouping by speech information is important, correlation was calculated as above when the articulatory labels were shuffled prior to alignment (Fig. S5, Note: labels were shuffled for alignment, but the proper labels were used when calculating correlation as described above). Labels were shuffled by random permutation. Latent dynamics correlation was calculated in Python 3.10.

Single-electrode articulator tuning

Univariate tuning of electrodes to different articulator types was performed similar to Duraivel et al. (2023)[27] (Fig. 3a). Briefly, for each patient and each electrode in that patient's μECoG array, we trained a PCA-LDA (linear discriminant analysis) decoding model to predict articulator type from the timeseries HG data for the current electrode. This model first projected input data to a low-dimensional subspace explaining some percentage of the variance with PCA, then learned to classify articulator type from low-dimensional features. Evaluation and training were performed with 5-fold nested cross-validation. The outer folds, splitting training and testing data, were used to measure performance as the receiver-operator-characteristic-area-under-the-curve (ROC-AUC) for the binary classification of each articulator type. ROC-AUC scores were average across all 5-test folds within articulator type. The inner folds, splitting training and validation data, were used to optimize the PCA variance percentage hyperparameter. Grid search was performed on a one-dimensional vector of PCA variance percentages (10% to 100%, 5% step size) and the best performing percentage across all validation folds was used in the current test set fold. Maps for each articulator type and patient were generated by repeating this process for all individual electrodes. Single-electrode articulator tuning was performed in Python 3.10 with the scikit-learn library.

Cross-patient projection

To project single-electrode articulator maps from a source patient's space to a target patient's space (Fig. 3b), we first learned decompositions from HG signals to latent dynamics with PCA and pairwise alignment with CCA for each patient as described above (*Alignment of latent dynamics*), but using HG signals from all channels instead of only significant channels. We then applied these transformations to spatially decompose, optionally align, and spatially reconstruct articulator maps across patients. Decomposition was performed by passing flattened articulator maps through the source patient's PCA. For compatibility between this PCA, which expects data centered around zero, and ROC-AUC values in the range [0.5, 1], we subtracted the mean ROC-AUC value across all channels to de-bias the ROC-AUC values. This enabled articulator maps to be scaled to proper feature ranges for transformation by the previously learned PCA. When aligned reconstructions were performed, alignment transformation matrices were applied to align the latent source patient articulator maps to the target patient's latent space. With alignment, transformation to a different sized latent space is handled automatically. Without alignment, we enforced that each PCA used the same number of components (n = 30) instead of the number of components explaining 90% of the original variance for dimensionality consistency. Finally, the PCA transform from the target patient was inversely applied to map (optionally aligned) latent articulator maps up to the target patient's electrode space. The mean ROC-AUC across all source patient channels (subtracted prior to transformation via source patient PCA) was then added back in to return ROC-AUC values to the proper domain. Fig. 3b (right) visualizes the reconstructions of S1 articulator maps by S3. Fig. S6 shows the reconstructed S1 articulator maps averaged across all source patients.

To compute the similarity between ground-truth and reconstructed articulator maps (Fig. 3c), the Pearson correlation was computed between ground-truth and reconstructed articulator maps. This is analogous to computing the cosine similarity between mean-subtracted ground truth and reconstructed articulator maps. Fig. 3c (left) quantifies the similarity for reconstructions of S1 articulator maps by only S3. In Fig. 3c (right), for each target patient, articulator maps from all source patients were considered by separately reconstructing these maps and averaging the maps across source patients as a single reconstructed map per articulator type prior to calculating similarity.

Additionally, we quantified the articulator decoding performance of models tested on aligned and unaligned reconstructed data (Fig. 3d). All decoding models used were PCA-SVM (support vector machine) models. This combined a PCA layer for feature size reduction (as in PCA-LDA models described above) with a bagged SVM classifier (scikit-learn, default parameters) using a radial basis function kernel. SVM PCA explained variance was set at a fixed 80% value. Articulator decoding performance (4 classes: low, high, labial, dorsal; chance = 0.25) was quantified as the balanced accuracy across predictions from all test folds with 20-fold cross-validation. The articulator identity for the first phoneme of the repeated three-phoneme sequence was predicted by the models. First, we trained baseline patient-specific decoding models for each patient, where both training and testing data were composed of data from a single patient. We then projected HG data from each source patient's (S2-S8) to S1's electrode space by inversely applying S1's PCA transformation to the source patient's latent dynamics, either unaligned or aligned to S1's space. Cross-patient decoding was performed by applying the decoding model trained on S1's data to each source patient's reconstructed HG data as test data.

Cross-patient phoneme decoding

We trained PCA-SVM decoding models (as described above in *Cross-patient projection*) to predict phoneme (9 classes: */a/, /e/, /i/, /u/, /b/, /p/, /v/, /g/, /k/*; chance = 0.11) identity in a target patient from combined cross-patient data across all patients in our dataset (Fig. 4a). Parameters and evaluation methods are identical to those defined above, but with phonemes as the labels instead of articulators. Phonemes were predicted at all three positions of the phoneme sequence. Phoneme decoding accuracy was reported as the balanced accuracy across predictions from all 20 test folds across 50 iterations to generate a distribution across different training and testing splits. For results aggregated across all patients (Fig. 4b), the average balanced accuracy across all 50 iterations was reported as a single value per patient. For all decoding contexts (patient-specific, unaligned cross-patient, aligned cross-patient), training and test splits were only performed on the target patient. That is, in cross-patient contexts, cross-patient data was only aligned to the target patient's training data (when alignment was performed) and added to the training set, such that model accuracy was quantified on held-out data from a single target patient, as would be the clinical use-case for cross-patient speech BCIs. Only HG data from significant channels were used for cross-patient phoneme decoding. Cross-patient phoneme decoding was performed in Python 3.10, primarily with the scikit-learn library.

We started by calculating the phoneme decoding accuracies of patient-specific models for each target patient as a baseline. Models were trained on unaligned cross-patient data by first separately decomposing all patient HG data to latent dynamics explaining 90% of the variance of the input HG data with PCA. For the target patient, PCA decomposition was learned on the training set and applied to the test set to avoid information leakage. Unaligned latent dynamics across all patients were concatenated as additional observations/trials into a single dataset (truncated to the minimum number of PCs across all patients for consistency of latent dimensionality). A PCA-SVM decoding model was then trained to predict phoneme identity from the expanded cross-patient training dataset. As mentioned above, phoneme decoding performance of the PCA-SVM was quantified on the test set containing only data from the target patient. Models were trained on aligned cross-patient data by repeating the same process for unaligned cross-patient data and applying CCA alignment after decomposing all patient data to latent dynamics. CCA alignment was calculated between each source patient and the target patient to align all source patient data to the same target patient space (this was additionally beneficial as no additional transformation of the test data in the target patient's space was required). As with PCA, CCA alignment was only learned between the training set of the target patient and cross-patient sources to prevent information leakage between training and testing sets. Fig. 4b shows cross-patient phoneme decoding results S1, S2, S3. The results across all patients are shown in Fig. S7.

As a control condition for cross-patient decoding, we aligned surrogate data generated from the Tensor Maximum Entropy (TME) procedure[47] to each target patient instead of true data from other patients (Fig. S8). Surrogate TME data is random data with the constraint that it must preserve the first and second-order statistics of the true neural data (i.e. mean and covariance across trials, time, and channels). As such, latent dynamics extracted from TME data should not encode speech information when aligned to the target patient's space. Surrogate TME data were

generated for each of the eight patients, where seven TME patients were aligned to true target patient data for each target patient. Alignment and subsequent decoding were done the same as with all true data.

For subsampling of target patient data (Fig. 4c, Fig. S9 for all patients separately), patient-specific and aligned cross-patient phoneme decoding were performed as described above while a percentage of the target patient's training trials were subsampled prior to alignment (for cross-patient models) and decoding. Subsampling was performed by randomly and uniformly sampling trials from the target patient's training dataset until the desired number of trials were obtained. Reported phoneme decoding accuracy is the average balanced decoding accuracy for each of the eight target patients.

For subsampling of cross-patient data (Fig. 4d, Fig. S10 for all patients separately), aligned cross-patient models were trained as described above while varying the amount of cross-patient data aligned to the target patient. We equally subsampled *n* trials from each patient, where *n* ranged from 5 trials to the maximum number of trials from the current patient with a step size of 25 (i.e. 5, 30, 55, …). Cross-patient trials were sampled equally from each patient to avoid effects of patient identity during sampling. Each point reports the phoneme decoding accuracy across 50 iterations of sampling. Fig. 4d visualizes these results for target patient S5. The average patient-specific decoding accuracy for S5 is overlaid for comparison. To show diminishing gains in decoding accuracy as the amount of cross-patient data increased, we performed a log-fit to S5's decoding accuracy vs. cross-patient trials curve using scipy's *curve_fit* function (Fig. S11).

Cross-patient recurrent neural network decoding

We also trained recurrent neural networks (RNNs) to decode phoneme identity from cross-patient latent dynamics (Fig. 4e, Fig. S12 for all patients). The implementation is similar to our previous work[27]. We chose to implement a sequence-to-sequence (seq2seq) RNN[48] due to its sequential decoding capabilities that pair well with the three-length phoneme sequence structure of our speech task. Seq2seq RNN design and training was implemented in Python 3.10 via Pytorch-lightning v2.3.3 and Pytorch v2.4.0.

The seq2seq RNN first contained a 1D temporal convolution layer that we used to extract input features from HG timeseries (patient-specific) or cross-patient latent dynamics (cross-patient) and perform temporal downsampling. This layer contained 100 convolutional filters with a non-overlapping size of 50 ms (stride = 50 ms) and a dropout rate of 0.3. No activation function was applied to the convolutional filters. The output of the temporal convolution layer was passed through an encoder RNN composed of two hidden layers with 500 gated recurrent units (GRUs). Forward and backward states of encoder RNN were summed to aggregate state information. Finally, a decoder RNN mapped the output of the encoder RNN to a predicted three-length phoneme sequence. The decoder RNN contained a hidden layer composed of 500 GRUs followed by a fully connected layer to generate phoneme output probabilities. Both encoder and decoder RNNs had a dropout rate of 0.3.

We trained the seq2seq RNN to minimize the cross-entropy loss between ground-truth phoneme sequences and predicted phoneme sequence probabilities using the AdamW[49] gradient descent optimization procedure with an initial learning rate of 1e-4, linearly decayed to 1e-6 over 500 epochs, and a gradient clip value of 0.5. An L2 regularization of 1e-5 was applied to seq2seq RNN weights during training to improve generalizability. The decoder RNN produced phoneme probabilities at each of its three steps in the phoneme sequence, using the output probabilities from the previous step as input to the next step. Training was performed with a teacher-forcing ratio of 0.5, where ground-truth phoneme probabilities are provided instead of using phoneme probabilities from the previous step, to improve the training speed of the model.

Evaluation of the seq2seq RNNs was done by 20-fold cross-validation on the target patient dataset, where the balanced accuracy of phoneme predictions across all test folds was reported as the phoneme decoding accuracy. As with the SVM models, phoneme decoding accuracy was reported across 50 training iterations. For prediction on the test set, we loaded the weights of the model checkpoint with the highest accuracy on a separate validation set. The validation was comprised of a random 10% of trials from the training fold. In patient-specific models, the validation set contained only patient-specific data. In cross-patient models, the validation split was done after CCA alignment and dataset concatenation, resulting in a larger validation set composed of cross-patient trials.

Pitch subsampling

Lower electrode densities were simulated by using Poisson disk sampling[50] to sub-select electrodes from the μECoG array constrained by a minimum distance between electrodes. Given a desired pitch as this constraining distance, we calculated the number of electrodes that needed to be sampled to achieve this pitch as:

$$n = XY/p^2,$$

where n is the number of electrodes, X and Y are the length and width of the μECoG array, and p is the desired pitch. We rounded n to the nearest integer in implementation. We only considered subsampled electrodes with significant HG activity during the patient response for subsequent cross-patient analyses. In these cross-patient analyses, all patients were subsampled to the desired pitch, not just the target patient.

Cross-patient decoding on pitch-subsampled data (Fig. 5a, top; Fig. S13 for all patients separately) was performed as described previously (*Cross-patient phoneme decoding*) with PCA-SVM models, but with pitch subsampling of the HG data prior to latent dynamics extraction, alignment, and decoding. At each of the 50 training iterations, a new subsampling of electrodes at the current pitch was selected to mitigate effects of electrode identity. Correlation between cross-patient latent dynamics from subsampled arrays (Fig. 5a, bottom) was calculated similarly to previous analyses (*Alignment of latent dynamics*) with a few minor differences: 1) at each pitch, the Pearson correlation was averaged across 50 iterations to mitigate effects of electrode identity during sampling and 2) correlation was calculated for each target patient

instead of just S1 as in Fig. 2e, so reported correlation values for a given target patient are averaged across all source patients.

Grid subsampling

Smaller electrode array coverages were simulated by sampling rectangular sub-grids of size n x $2n$ ($n \epsilon [2,8]$) from the total electrode array. While maximum electrode array sizes exceeded 8x16 for patients with arrays of 256 electrodes, we only considered sub-grids up to 8x16 for consistency across all patients. For a given sub-grid size, we consider all possible starting locations that fit within the total electrode array. That is, for an 8x16 electrode array with a sub-grid size of 6x12, there would be a total of 15 possible sub-grids to select from. Only electrodes with significant HG activity during the patient response were selected from each sub-grid.

Cross-patient decoding from sub-grids (Fig. 5b, top; Fig. S14 for all patients separately) was again performed as previously described with PCA-SVM models and followed a similar approach to the cross-patient correlation analysis for sub-grid selection. For a given sub-grid size, training was repeated for all possible target patient sub-grids and a random sub-grid from each source patient was selected at each iteration. At each sub-grid size, phoneme decoding accuracy was averaged across all target patient sub-grids. Correlation between cross-patient latent dynamics from sub-grids (Fig. 5b, bottom) were extracted similarly to previous analyses. For each target patient and sub-grid size, we extracted latent dynamics from HG in the sub-selected electrodes and aligned them across patients. We iterated over all possible sub-grids of the given size for the target patient. For each possible sub-grid, we selected a random sub-grid (uniformly) from the possibilities for each source patient. Following, Pearson correlation across aligned latent dynamics was calculated as previously described. For each sub-grid size, the reported correlation was averaged across all target patient sub-grids and across all source patients, yielding a single averaged Pearson r for each target patient.

Spatial-averaging

Larger contact sizes were simulated by averaging data together from neighboring contacts with varying sizes. Contact sizes ranged from 1x1 (unaveraged micro-contact used in previous analyses in this work) to 8x8 squares. To simulate neural recording with these larger contacts, raw signals were averaged together on these contacts and HG was extracted from the resultant spatially-averaged neural signal. Only spatially-averaged contacts with significant HG activity during the patient response (relative to a baseline period 500 ms prior to stimulus presentation) were considered for subsequent analyses (see *Statistical Analysis*). Note that this spatial-averaging method does not account for changes in electrode impedance for different sized contacts. For contact sizes that did not span the total electrode array (e.g. Two 8x8 contacts span an 8x16 electrode array, but 7x7 contacts do not), electrodes were centered within the electrode array.

Cross-patient decoding on spatially-averaged data (Fig. 5c, top; Fig. S15 for all patients separately) was performed as described previously (*Cross-patient phoneme decoding*) with PCA-SVM models, but with the spatially-averaged data used for latent dynamics extraction, alignment,

and decoding. Cross-patient latent dynamics correlation for spatially averaged data (Fig. 5c, bottom) was again calculated similar to previous analyses, but with spatial-averaging of data prior to latent dynamics extraction and alignment. As the spatial-averaging results in a singular, deterministic sub-selection (unlike pitch and grid subsampling), the correlation in latent dynamics as not repeated over multiple iterations. Instead, a single Pearson *r* was calculated between each target patient and source patient and averaged across all source patients for each target patient (as in *Alignment of latent dynamics*). For each contact size, correlation was reported across a distribution of all target patients.

Statistical analysis

All statistical analyses were performed in MATLAB R2021a and Python 3.10 (scipy and statsmodels). All permutation tests mentioned below use the difference in means as the test statistic. All false discovery rate (FDR) control mentioned below implements the Benjamini-Hochberg method[51].

Fig. 1: To identify electrodes with significant HG activation during patient responses, we compared the average (over time) HG power around the patient response onset (-250 ms to 250 ms) to a pre-stimulus baseline period (-500 ms to 0 ms relative to stimulus presentation) for each electrode using a FDR-corrected 10,000-iteration one-sided permutation test.

Fig. 2: For latent dynamics clustering strength (quantified by the silhouette score), comparisons between distributions (articulatory features vs chance) were performed with Mann-Whitney U (MWU) test. MWU tests were also used for cluster strength comparisons when combining data from both S1 and S2, where FDR correction was applied. Differences in paired distributions for cross-patient latent dynamics correlation was compared with a Wilcoxon signed-rank test.

Fig. 3: Reconstructed articulator map similarity was compared between aligned and unaligned contexts with a Wilcoxon signed-rank test. Cross-patient articulator decoding accuracy distributions were compared with exact two-sided, paired permutation tests and FDR correction. Comparison of cross-patient articulator decoding accuracies to chance accuracy was performed using Wilcoxon signed-rank tests corrected with FDR correction.

Fig. 4: Differences in aggregated results for all target patients were tested with a repeated measures ANOVA and follow-up paired t-tests with false discovery control. Within target patients, differences in distributions of phoneme decoding accuracy were compared across decoding contexts (patient-specific, unaligned cross-patient, aligned cross-patient) using a one-way analysis of variance (ANOVA) and post-hoc Tukey tests. For target patient data subsampling, an exact two-sided permutation test was performed at each target data subsampling ratio. FDR correction was applied across all subsampling ratio tests. Finally, for cross-patient RNN and SVM comparisons, statistical testing was performed with a two-way ANOVA across decoding context (patient-specific, aligned cross-patient) and model type (SVM, RNN) with post-hoc Tukey HSD tests.

Fig. 5: Statistical testing was performed to compare differences between aligned cross-patient and patient-specific decoding accuracies for different subsampling methods. For each subsampling method (pitch, grid, and spatial-averaging) we performed an exact two-sided, paired permutation test at each subsampled point with FDR correction across all points.

Code Availability

Code for cross-patient alignment analyses and notebooks recreating the figures in this work can be found at https://github.com/coganlab/cross_patient_speech_decoding.

Data Availability

The data from μECoG recordings that support the decoding analysis are available via the Data Archive for The Brain Initiative (DABI; accession code: <To be updated after acceptance>) under restricted access (restrictions will be exempted after appropriate modifications of the IRB protocol: Pro0072892). The access can be obtained upon written requests from the corresponding authors (gregory.cogan@duke.edu, j.viventi@duke.edu). Source data are provided with this paper.

References

1. Levelt, W. J. M. *Speaking: From Intention to Articulation*. (The MIT Press, 1989). doi:10.7551/mitpress/6393.001.0001.
2. Mitchell, J. & Borasio, G. Amyotrophic lateral sclerosis. *The Lancet* **369**, 2031–2041 (2007).
3. Gowda, S. N., Munakomi, S. & De Jesus, O. Brainstem Stroke. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2024).
4. Lulé, D. *et al.* Life can be worth living in locked-in syndrome. in *Progress in Brain Research* vol. 177 339–351 (Elsevier, 2009).
5. Haig, A. J., Katz, R. T. & Sahgal, V. Mortality and complications of the locked-in syndrome. *Arch. Phys. Med. Rehabil.* **68**, 24–27 (1987).
6. Felgoise, S. H., Zaccheo, V., Duff, J. & Simmons, Z. Verbal communication impacts quality of life in patients with amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler. Front. Degener.* **17**, 179–183 (2016).
7. Willett, F. R. *et al.* A high-performance speech neuroprosthesis. *Nature* **620**, 1031–1036 (2023).

8. Metzger, S. L. *et al.* A high-performance neuroprosthesis for speech decoding and avatar control. *Nature* **620**, 1037–1046 (2023).
9. Moses, D. A. *et al.* Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria. *N. Engl. J. Med.* **385**, 217–227 (2021).
10. Card, N. S. *et al.* An Accurate and Rapidly Calibrating Speech Neuroprosthesis. *N. Engl. J. Med.* **391**, 609–618 (2024).
11. Dabagia, M., Kording, K. P. & Dyer, E. L. Aligning latent representations of neural activity. *Nat. Biomed. Eng.* **7**, 337–343 (2022).
12. 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).
13. Chartier, J., Anumanchipalli, G. K., Johnson, K. & Chang, E. F. Encoding of Articulatory Kinematic Trajectories in Human Speech Sensorimotor Cortex. *Neuron* **98**, 1042-1054.e4 (2018).
14. Metzger, S. L. *et al.* Generalizable spelling using a speech neuroprosthesis in an individual with severe limb and vocal paralysis. *Nat. Commun.* **13**, 6510 (2022).
15. Kellis, S. *et al.* Decoding spoken words using local field potentials recorded from the cortical surface. *J. Neural Eng.* **7**, 056007 (2010).
16. Herff, C. *et al.* Brain-to-text: decoding spoken phrases from phone representations in the brain. *Front. Neurosci.* **9**, (2015).
17. Bouchard, K. E., Mesgarani, N., Johnson, K. & Chang, E. F. Functional organization of human sensorimotor cortex for speech articulation. *Nature* **495**, 327–332 (2013).
18. Pei, X. *et al.* Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition. *NeuroImage* **54**, 2960–2972 (2011).
19. Stavisky, S. D. *et al.* Neural ensemble dynamics in dorsal motor cortex during speech in people with paralysis. *eLife* **8**, e46015 (2019).

20. Wilson, G. H. *et al.* Decoding spoken English from intracortical electrode arrays in dorsal precentral gyrus. *J. Neural Eng.* **17**, 066007 (2020).
21. Churchland, M. M. *et al.* Neural population dynamics during reaching. *Nature* **487**, 51–56 (2012).
22. Gallego, J. A., Perich, M. G., Miller, L. E. & Solla, S. A. Neural Manifolds for the Control of Movement. *Neuron* **94**, 978–984 (2017).
23. Gallego, J. A., Perich, M. G., Chowdhury, R. H., Solla, S. A. & Miller, L. E. Long-term stability of cortical population dynamics underlying consistent behavior. *Nat. Neurosci.* **23**, 260–270 (2020).
24. Safaie, M. *et al.* Preserved neural dynamics across animals performing similar behaviour. *Nature* **623**, 765–771 (2023).
25. Chiang, C.-H. *et al.* Flexible, high-resolution thin-film electrodes for human and animal neural research. *J. Neural Eng.* **18**, 045009 (2021).
26. Barth, K. J. *et al.* Flexible, high-resolution cortical arrays with large coverage capture microscale high-frequency oscillations in patients with epilepsy. *Epilepsia* **64**, 1910–1924 (2023).
27. Duraivel, S. *et al.* High-resolution neural recordings improve the accuracy of speech decoding. *Nat. Commun.* **14**, 6938 (2023).
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. Leonard, M. K. *et al.* Large-scale single-neuron speech sound encoding across the depth of human cortex. *Nature* **626**, 593–602 (2024).
30. Ray, S. & Maunsell, J. H. R. Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS Biol.* **9**, (2011).

31. 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).
32. Sussillo, D., Churchland, M. M., Kaufman, M. T. & Shenoy, K. V. A neural network that finds a naturalistic solution for the production of muscle activity. *Nat. Neurosci.* **18**, 1025–1033 (2015).
33. Buzsáki, G., Anastassiou, C. A. & Koch, C. The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes. *Nat. Rev. Neurosci.* **13**, 407–420 (2012).
34. Dubey, A. & Ray, S. Cortical Electrocorticogram (ECoG) Is a Local Signal. *J. Neurosci.* **39**, 4299–4311 (2019).
35. Gallego-Carracedo, C., Perich, M. G., Chowdhury, R. H., Miller, L. E. & Gallego, J. Á. Local field potentials reflect cortical population dynamics in a region-specific and frequency-dependent manner. *eLife* **11**, e73155 (2022).
36. Pandarinath, C. *et al.* Inferring single-trial neural population dynamics using sequential auto-encoders. *Nat. Methods* **15**, 805–815 (2018).
37. Gosztolai, A., Peach, R. L., Arnaudon, A., Barahona, M. & Vandergheynst, P. MARBLE: interpretable representations of neural population dynamics using geometric deep learning. *Nat. Methods* **22**, 612–620 (2025).
38. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical Surface-Based Analysis. *NeuroImage* **9**, 179–194 (1999).
39. Papademetris, X. *et al.* BioImage Suite: An integrated medical image analysis suite: An update. *Insight J.* (2006) doi:10.54294/2g80r4.
40. McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. & Sonderegger, M. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. in *Interspeech 2017* 498–502 (ISCA, 2017). doi:10.21437/Interspeech.2017-1386.
41. Audacity Team. Audacity. Muse Group (2024).

42. Cogan, G. B. *et al.* Manipulating stored phonological input during verbal working memory. *Nat. Neurosci.* **20**, 279–286 (2017).
43. Mesgarani, N., Cheung, C., Johnson, K. & Chang, E. F. Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science* **343**, 1006–1010 (2014).
44. Fabian Pedregosa *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
45. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
46. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
47. Elsayed, G. F. & Cunningham, J. P. Structure in neural population recordings: an expected byproduct of simpler phenomena? *Nat. Neurosci.* **20**, 1310–1318 (2017).
48. Sutskever, I., Vinyals, O. & Le, Q. V. Sequence to Sequence Learning with Neural Networks. Preprint at https://doi.org/10.48550/ARXIV.1409.3215 (2014).
49. Loshchilov, I. & Hutter, F. Decoupled Weight Decay Regularization. Preprint at https://doi.org/10.48550/ARXIV.1711.05101 (2017).
50. Bridson, R. Fast Poisson disk sampling in arbitrary dimensions. in *ACM SIGGRAPH 2007 sketches* 22 (ACM, San Diego California, 2007). doi:10.1145/1278780.1278807.
51. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57**, 289–300 (1995).

Acknowledgements

J.V., G.B.C., S.P.L., A.H.F., D.G.S, S.D., Z.S. were supported by NIH R01DC019498 and DoD W81XWH-21-0538. S.R. was supported by an NRCDP NIH K12 award. J.V., G.B.C., C.W., S.D. were supported by NIH CTSA UL1TR002553. K.B. was supported by the National Science Foundation (Graduate Research Fellowship Program DGE-1644868). C.S. was supported by the National Science Foundation (Graduate Research Fellowship Program DGE-2139754) J.V. and G.B.C. were supported by NIH UG3NS120172, G.B.C, J.V., S.P.L, A.H.F, D.G.S. were supported by NIH R01NS129703, and G.B.C., J.V., S.P.L., S.D. were supported by an internal Duke Institute for Brain Sciences Incubator Award. We thank Nicole Liddle, Palee Abeyta, and Anna Thirakul for help with consenting participants, Seth Foster for help with the experimental design, Steve Harward for help with surgical preparation, and Daniel Sexton for help with surgical preparation and neuroanatomical localizations of electrode arrays. With tremendous gratitude, we thank the patients who volunteered their time for our study.

Author Contributions

Z.S. performed all analyses, performed statistical testing, and generated figures. S.D. performed significant preprocessing of the neural data and assisted in analysis design. G.B.C. and S.R. designed the experimental task. S.R., S.P.L., A.H.F, D.G.S performed intraoperative electrode implantation and neural data acquisition. C.W., K.B., C.S., and J.V. were involved in μECoG electrode design and setting up the intraoperative data acquisition process. Z.S., S.D, K.B., C.S., and J.V., and G.B.C performed intraoperative neural data collection. Z.S., J.V., and G.B.C wrote the manuscript with feedback from other co-authors. G.B.C. was involved in clinical supervision. J.V., and G.B.C. designed and supervised the human intraoperative study.

Competing Interests

The authors declare no competing interests.

1046 ## Supplementary Tables

1047 **Table S1. Clinical summary of patients**

Patient	Age	Sex	Diagnosis	Number of channels
S1	61	M	Parkinson's	128
S2	64	M	Parkinson's	128
S3	26	F	Tumor resection	256
S4	63	M	Parkinson's	128
S5	61	M	Parkinson's	128
S6	71	M	Tumor resection	256
S7	53	M	Tumor resection	256
S8	60	M	Tumor resection	256

1048 **Mean age:** 57.4

1049

1050

1051

1052

1053

1054

1055

1056

1057

1058

1059

1060

1061

1062

1063

1064 **Table S2. Task stimuli**

CVC	VCV
/bab/	*/abe/*
/bek/	*/abi/*
/bak/	*/eba/*
/bup/	*/ebi/*
/gab/	*/ebu/*
/geb/	*/ega/*
/gev/	*/eka/*
/gak/	*/epi/*
/gav/	*/aka/*
/gig/	*/aku/*
/gib/	*/ava/*
/gub/	*/ave/*
/kab/	*/ibu/*
/keg/	*/ika/*
/kug/	*/ike/*
/pek/	*/ipu/*
/pep/	*/iva/*
/pev/	*/ivu/*
/puk/	*/uba/*
/pup/	*/uga/*
/vek/	*/uge/*
/veg/	*/uke/*
/vip/	*/upi/*
/vug/	*/uve/*
/vuk/	*/uvi/*

1065

1066 **Table S3. Patient data summary**

Patient	No. of trials with patient response	Approx. utterance duration (min.)	Approx. experiment duration (min.)	No. of channels with significant HG activity (no. and % of total)
S1	144	1.08	9.00	111/128 (86.7%)
S2	148	1.11	9.25	111/128 (86.7%)
S3	46	0.35	2.88	149/256 (58.2%)
S4	151	1.13	9.44	74/128 (57.8%)
S5	151	1.13	9.44	63/128 (49.2%)
S6	137	1.03	8.56	144/256 (56.3%)
S7	141	1.06	8.81	171/256 (66.8%)
S8	178	1.34	11.13	201/256 (78.5%)

1067 **Total utterance duration across patients:** ~8.23 minutes
1068 **Total experiment duration across patients:** ~68.51 minutes

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

1080

1081

1082

Table S4. Seq2seq RNN hyperparameters

# Convolutional filters	100
Stride, kernel size	50 ms
RNN unit type	GRU
# RNN units (hidden size)	500
# RNN encoder layers	2
# RNN decoder layers	1
CNN dropout	0.3
RNN dropout	0.3
Total epochs*	500
Learning rate	1e-4
L2 Regularization	1e-5
Optimization	AdamW
Gradient clipping	0.5

* While 500 total epochs were run for training, model weights from the checkpoint with the highest accuracy on the validation set were loaded for use on the test set

Supplementary Figures

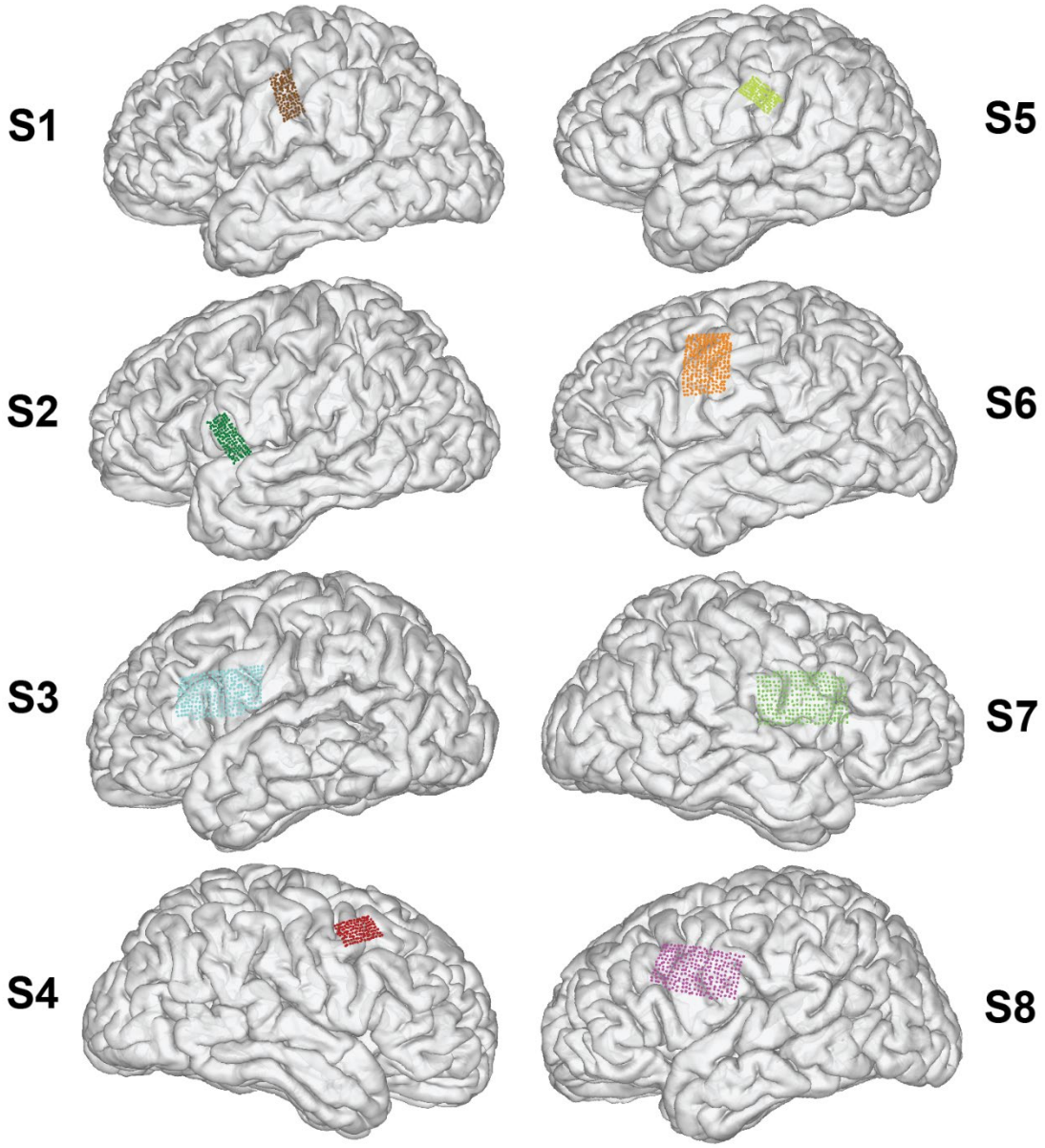

Figure S1. Patient-specific electrode array locations. Neuroanatomical locations of μECoG arrays as shown in Fig. 1, but on patient-specific brains instead of projected to a common brain.

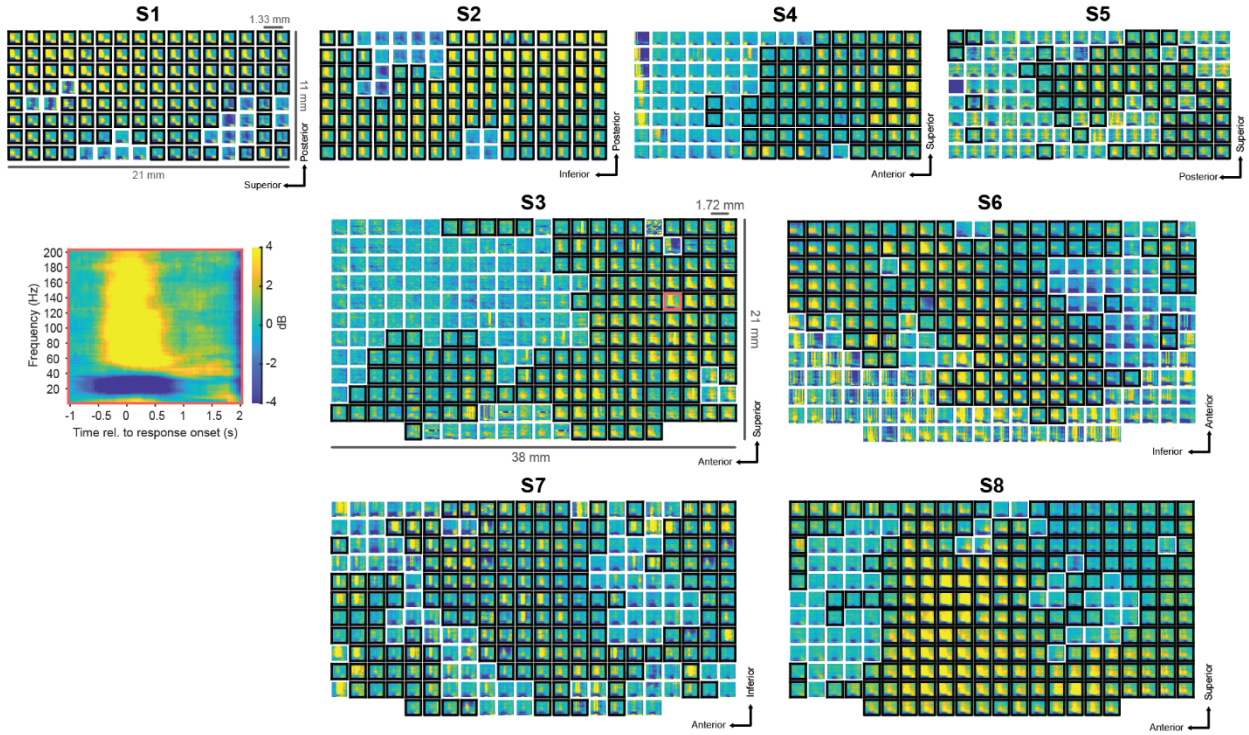

Figure S2. Spectrotemporal speech activations. Spectrotemporal activation patterns for all electrodes across the μECoG arrays of all patients (S1-S8), where each box represents a single electrode. Electrodes with black borders were found to have significant HG activity during the patient response, relative to a pre-stimulus baseline (see Methods). The spectrogram on the left shows a single channel from S3's array with relevant HG activity around speech production. μECoG array characteristics (size, pitch) are annotated on the S1 (128 channels) and S3 (256 channels) arrays. Axes on the bottom right of each array indicate the neuroanatomical orientation of the array (See Fig. S1 for full neuroanatomical locations).

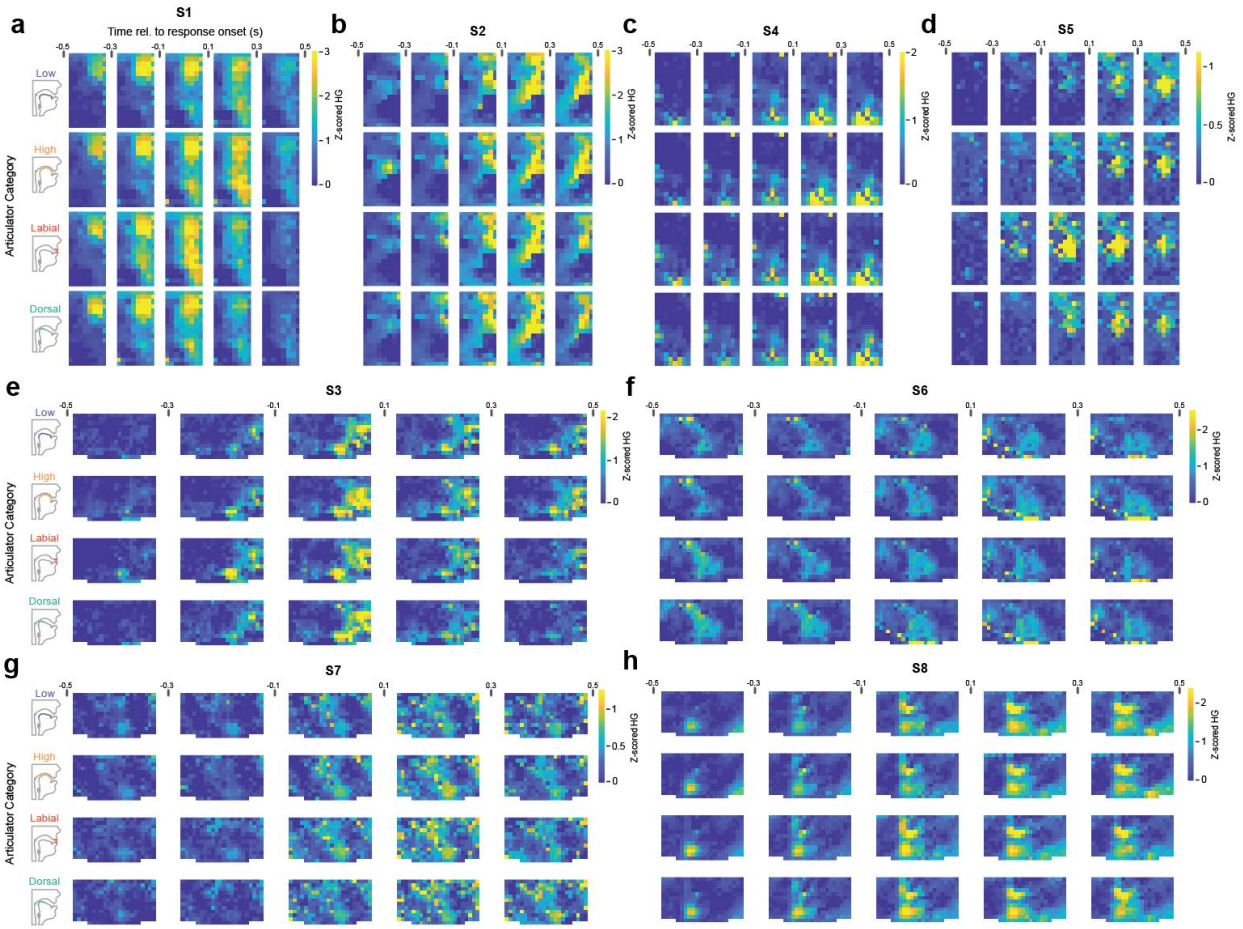

Figure S3. Spatiotemporal high gamma activations. Spatiotemporal activation patterns of HG across the μECoG arrays for all patients (S1-S8, **a**-**h**) grouped by different articulatory gestures. Vocal tract diagrams show visual representation of articulatory category (low, high, labial, dorsal). Note that colorbars reflect different scales per patient to show active regions despite differences in HG z-scores across patients.

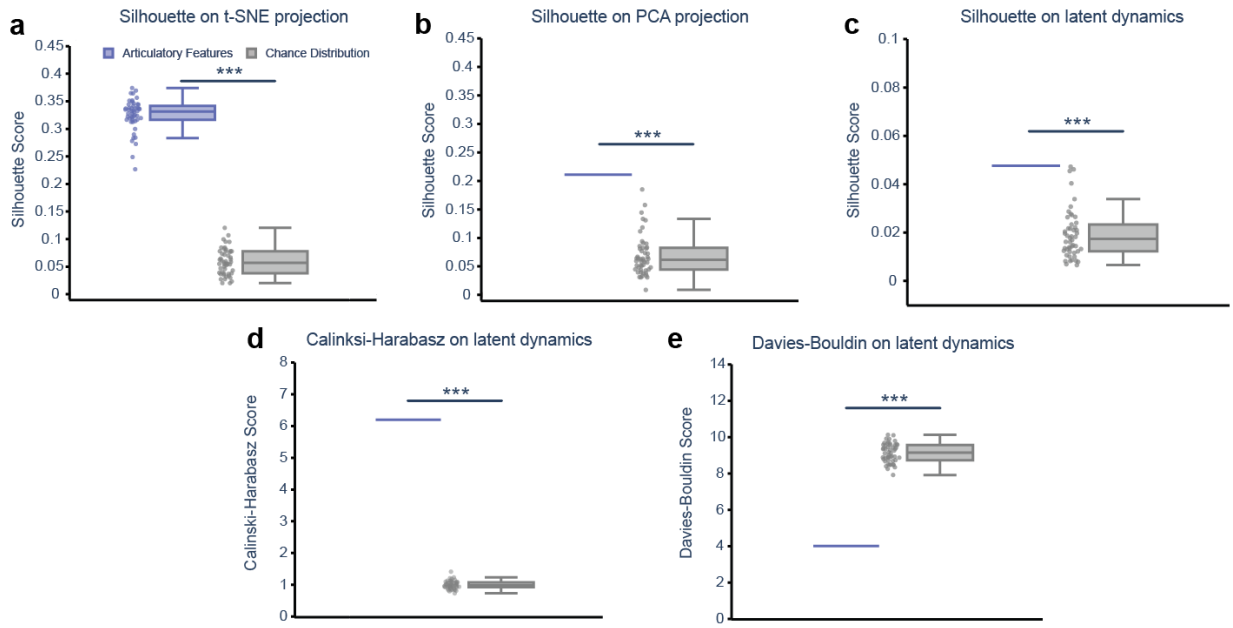

Figure S4. Clustering of S1 latent dynamics with various metrics. a. Clustering strength of articulatory features as quantified by the silhouette score on 2D t-SNE-projected latent dynamics (as reported in Fig. 2). Articulatory clustering is stronger than in a shuffled chance distribution (***$p < 0.001$, Mann-Whitney U test). As in the main text, the chance distribution is obtained by applying the same clustering quantification after randomly permuting the labels of articulatory features across trials). **b.** Clustering strength as quantified by the silhouette score on 2D PCA-projected latent dynamics. Since PCA results in a deterministic projection, only a single value is generated for the true articulatory features. Articulatory clustering is stronger than in a shuffled chance distribution ($p < 0.001$, one-sample T-test). **c.** Clustering strength directly on the latent dynamics. Articulatory clustering is stronger than in a shuffled chance distribution ($p < 0.001$, one-sample T-test). Due to the relatively high dimensionality of the space (latent components by timepoints), the silhouette score is lower than in **a** and **b**. **d.** Clustering strength as quantified by the Calinski-Harabasz score directly on latent dynamics. Articulatory clustering is stronger than in a shuffled chance distribution ($p < 0.001$, one-sample T-test). **e.** Clustering strength as quantified by the Davies-Bouldin score directly on latent dynamics. A lower score indicates stronger clustering for this metric. Articulatory clustering is stronger than in a shuffled chance distribution ($p < 0.001$, one-sample T-test).

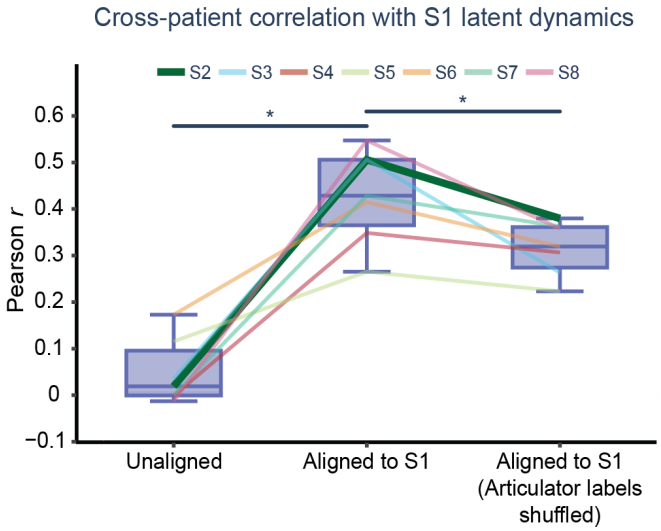

Figure S5. Speech information improves cross-patient correlation. Cross-patient correlation between the latent dynamics of S1 and S2-S8. The first two columns are as reported in Fig. 2. For correspondence between trial types across patients during alignment, trials are condition-averaged by articulator type (see Methods). In the third column, these articulator labels are shuffled prior to condition-averaging and subsequent alignment. Here, we show that label-agnostic condition-averaging prior to alignment results in significantly worse cross-patient correlation of latent dynamics, supporting the idea that alignment is occurring between speech-relevant neural latent dynamics *(*p* < 0.05, Wilcoxon signed-rank test)

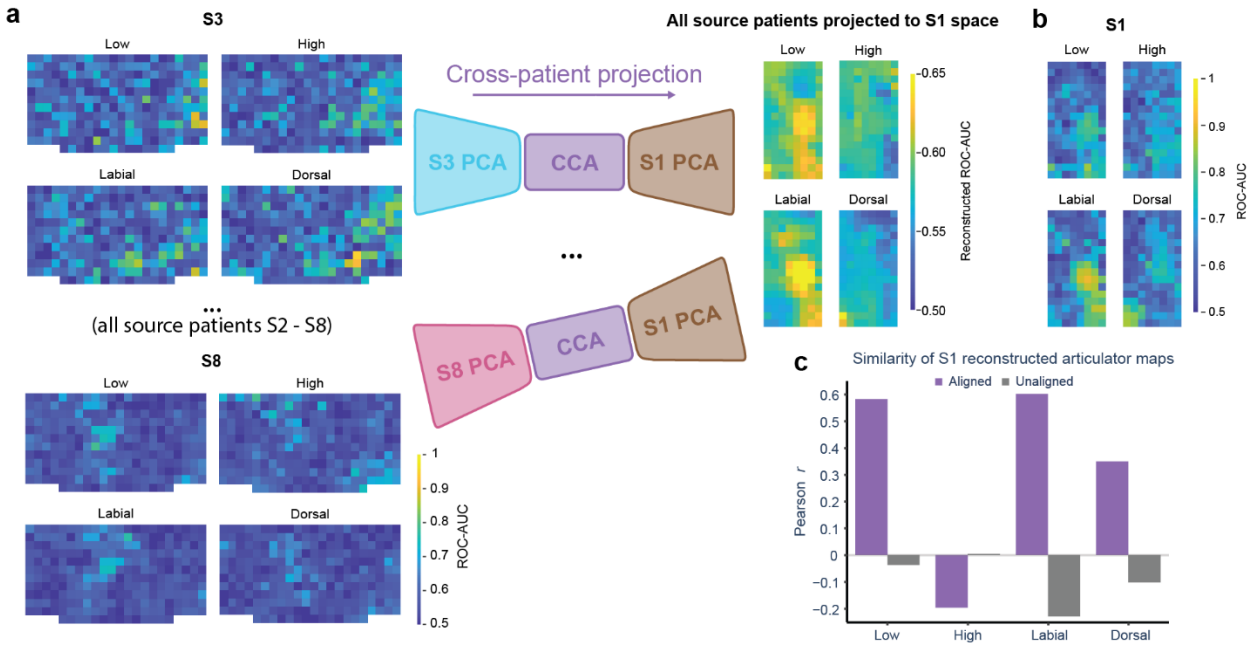

Figure S6. Cross-patient projection averaged across source patients. a. Cross-patient project, as in Fig. 3, but across multiple source patients. The articulator maps of all patients S2-S8 (left, S3 and S8 shown here) are projected to S1's electrode-space (see Methods) and the resultant reconstructed maps are averaged across source patients within articulator type (right). This method considers reconstructed speech information across all source patients instead of a singular source patient as shown in the Fig. 3b. **b.** Ground-truth S1 articulator maps for reference (as reported in Fig. 3a). **c.** Similarity between source-patient-averaged reconstructed articulator maps and ground truth S1 articulator maps, as measured by Pearson correlation. All articulator types except "high" show improved similarity when compared to reconstructions of S1 articulator maps by only S3 (as in Fig. 3c, left).

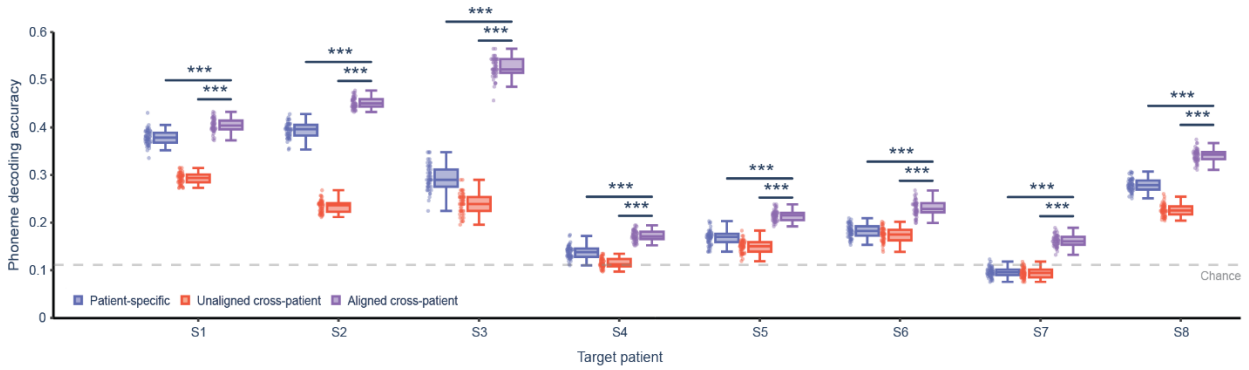

Figure S7. Cross-patient phoneme decoding by target patient. Cross-patient phoneme decoding for each target patient separately. Aligned, cross-patient data results in significantly higher phoneme decoding than both unaligned and patient-specific data for all patients (***$p < 0.001$, post-hoc Tukey HSD test following one-way ANOVA within target patient). The gray dotted line indicates chance levels of decoding.

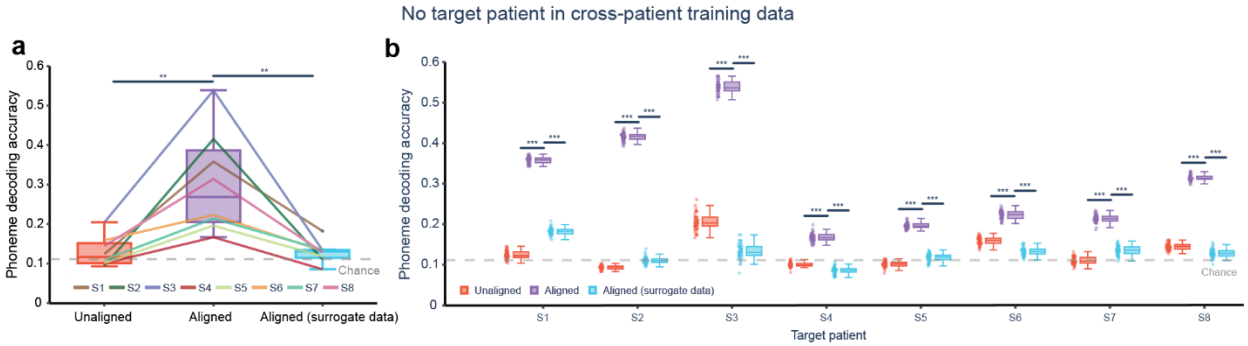

Figure S8. Cross-patient phoneme decoding on non-target patient data. As in Fig. 4 and Fig. S7, but data from the target patient is excluded from the training set of the decoding model (i.e. target patient data is only used to align other patients to the target space).The gray dotted line indicates chance levels of decoding. **a,** Without alignment, data from other patients results in chance levels of phoneme decoding. With alignment, we see significantly higher phoneme decoding accuracies across all patients (***p < 0.01*, repeated measures ANOVA and follow-up FDR-corrected paired t-tests). To confirm that the increased decoding accuracies from alignment were not artificially induced by the alignment procedure, surrogate data was generated that preserved second-order statistics of neural data from each patient but was otherwise random (see Methods) and was aligned to a non-surrogate target patient prior to phoneme decoding. This aligned surrogate data achieved approximately chance levels of decoding and had significantly lower phoneme decoding accuracy than the aligned non-surrogate context, validating the aligned decoding approach. **b,** As in **a** but for each target patient individually. Alignment of true neural data results in higher accuracies than unaligned and aligned surrogate data for all target patients, even without the inclusion of the target patient in the training data (****p<0.001,* post-hoc Tukey test following one-way ANOVA within target patient).

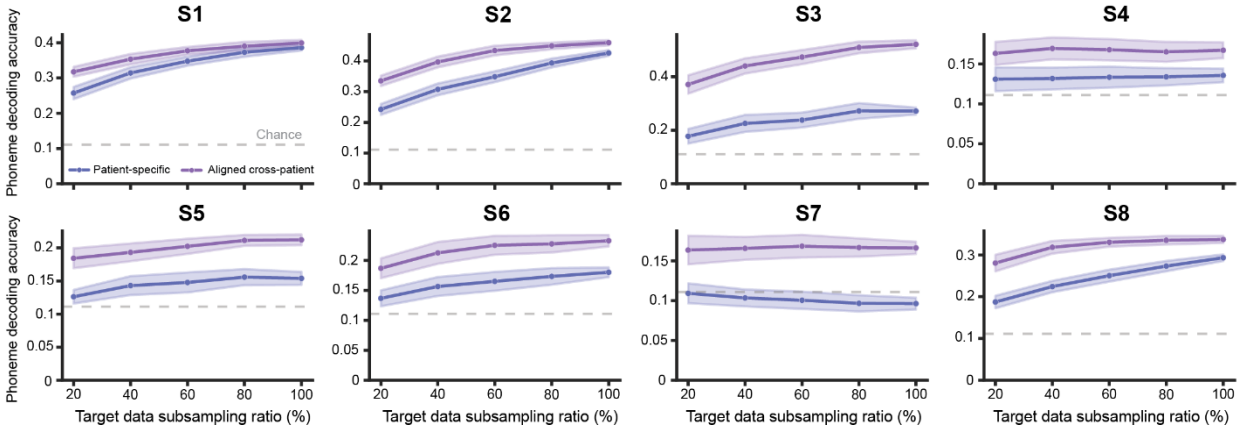

Figure S9. Target patient trial subsampling by patient. As reported in Fig. 4c, but plotted separately for each target patient. Patient-specific and aligned cross-patient decoding is performed as the amount of data from the target patient (used for both alignment and decoding) is subsampled. The gray dotted line indicates chance levels of decoding.

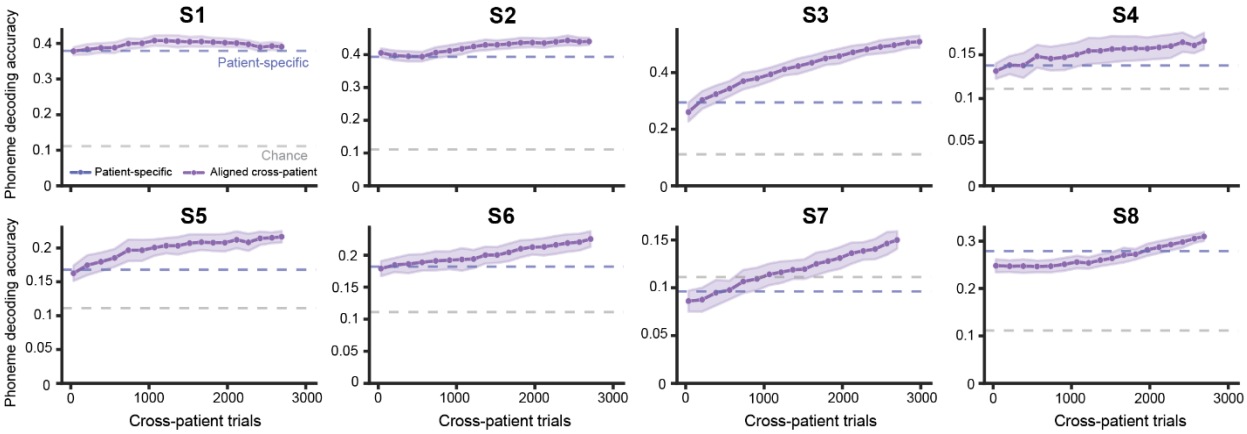

Figure S10. Cross-patient trial subsampling by patient. As reported in Fig. 4d (only S5), but plotted separately for each target patient. Aligned cross-patient decoding is performed as the amount of data from the target patient (used for both alignment and decoding) is subsampled. The gray dotted line indicates chance levels of decoding. The blue dotted line indicates the mean patient-specific decoding accuracy for each target patient.

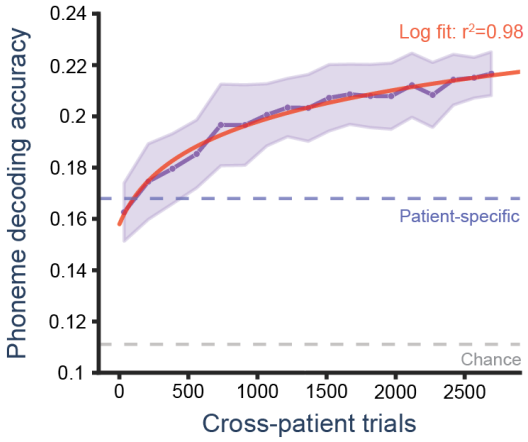

Figure S11. Log fit to S5 cross-patient subsampling data. As reported in Fig. 4d, but with a log fit to S5's phoneme decoding accuracy vs. cross-patient trials included data. The log fit was used to confirm diminishing gains in decoding accuracy for the SVM model as more cross-patient data is added. The log-fit was well-matched to the data with an r^2 of 0.98. As the fit is a non-linear regression, no p-value was obtained. The gray dotted line indicates chance levels of decoding. The blue dotted line indicates the mean S5's patient-specific decoding accuracy.

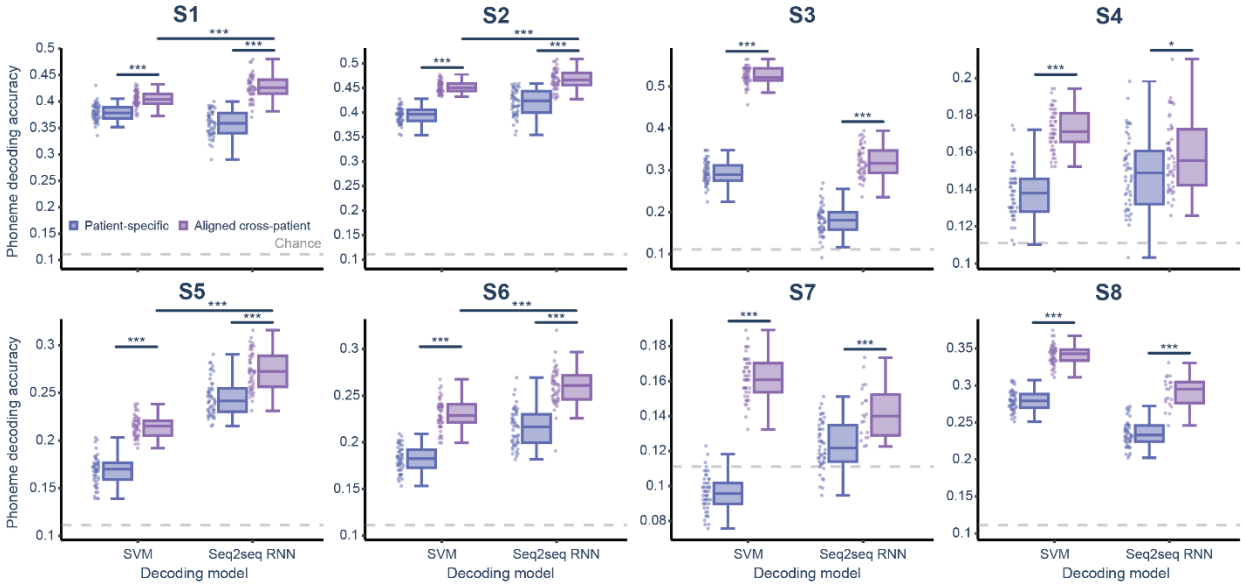

Figure S12. Cross-patient recurrent neural network decoding by patient. As reported for S5 in Fig. 4e, but plotted for all target patients. SVM patient-specific and cross-patient decoding accuracies are as reported in Fig. 4 and Fig. S7. For all patients, the Seq2seq RNN achieved significantly higher phoneme decoding accuracy when trained on aligned cross-patient data relative to patient-specific data (*$*p < 0.05$, $***p < 0.001$,* post-hoc Tukey test following two-way ANOVA within each target patient). For patients S1, S2, S5, and S6, the aligned RNN outperformed the aligned SVM *$***p < 0.001$,* post-hoc Tukey test following two-way ANOVA within each target patient). The gray dotted line indicates chance levels of decoding.

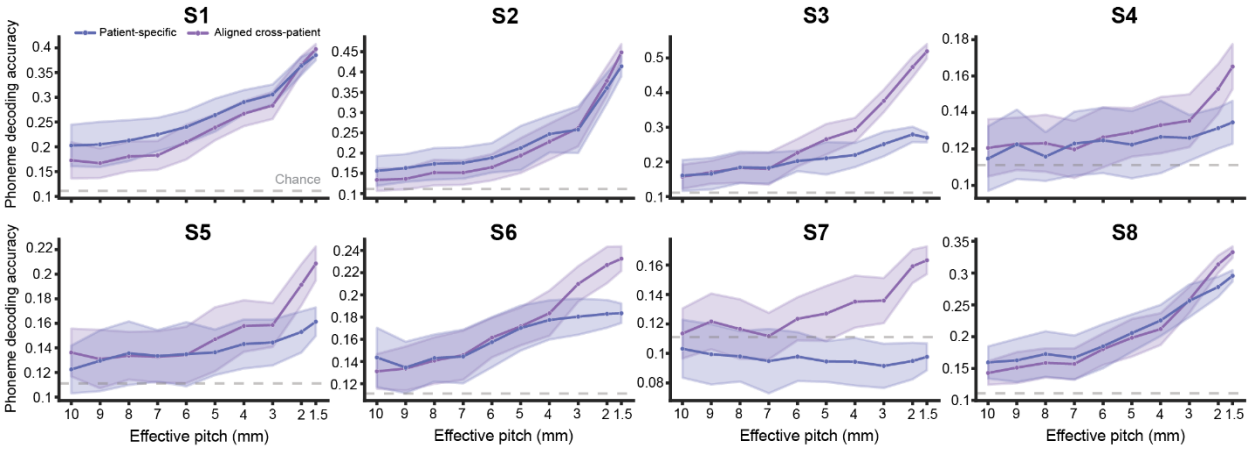

Figure S13. Pitch-subsampled phoneme decoding by patient. As reported in Fig. 5a, but plotted separately for each target patient. Patient-specific and aligned cross-patient decoding is performed with Poisson-disk-subsampled electrode data to simulate different electrode array densities. The gray dotted line indicates chance levels of decoding.

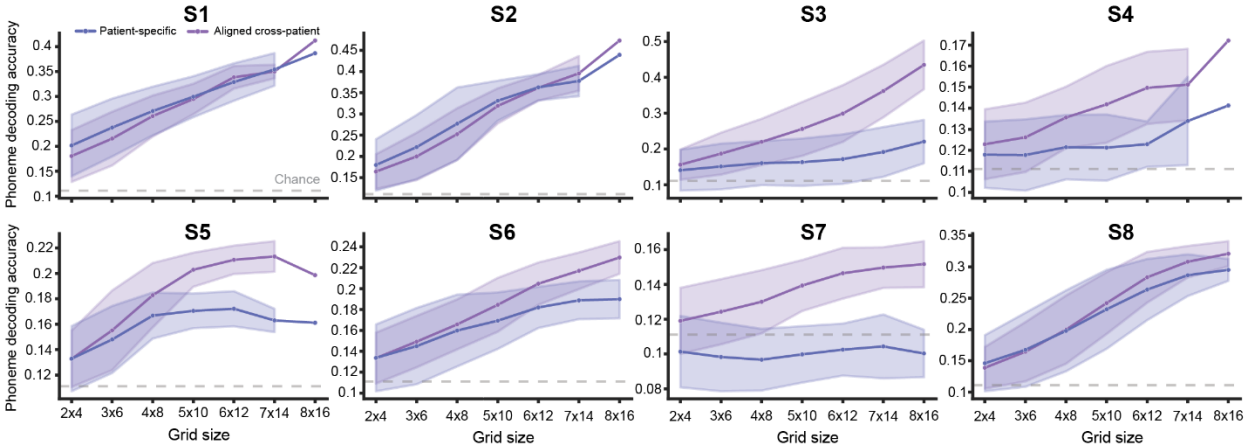

Figure S14. Grid-subsampled phoneme decoding by patient. As reported in Fig. 5b, but plotted separately for each target patient. Patient-specific and aligned cross-patient decoding is performed with data from subsampled grids of electrodes to simulate different electrode array coverages. The gray dotted line indicates chance levels of decoding. For patients with 128-channel arrays (S1, S2, S4, S5), there is a singular point for the 8x16 grid size since that encompasses the entire electrode array.

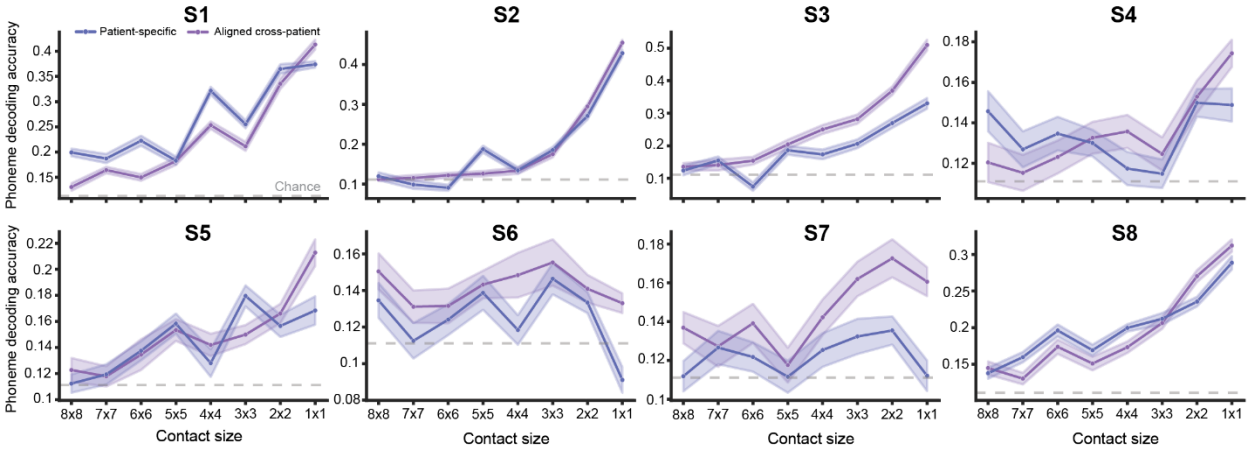

Figure S15. Spatially-averaged phoneme decoding by patient. As reported in Fig. 5c, but plotted separately for each target patient. Patient-specific and aligned cross-patient decoding is performed with HG data extracted from averaged electrode contacts of various sizes to simulate recording with larger electrode contacts. The gray dotted line indicates chance levels of decoding.

1426