Non-invasive electromyographic speech neuroprosthesis: a geometric perspective

Anonymous

Abstract

We present a neuromuscular speech interface that translates silently voiced articulations directly into text. We record surface electromyographic (EMG) signals from multiple articulatory sites on the face and neck as participants *silently* articulate speech, enabling direct EMG-to-text translation. Such an interface has the potential to restore communication for individuals who have lost the ability to produce intelligible speech due to laryngectomy, neuromuscular disease, stroke, or trauma-induced damage (e.g., radiotherapy toxicity) to the speech articulators. Prior work has largely focused on mapping EMG collected during *audible* articulation to time-aligned audio targets or transferring these targets to *silent* EMG recordings, which inherently requires audio and limits its applicability to patients who can no longer speak. In contrast, we propose an efficient representation of high-dimensional EMG signals and demonstrate direct sequence-to-sequence EMG-to-text conversion at the phonemic level without relying on time-aligned audio.

PROJECT PAGE. GITHUB[1].

1 Introduction

Electromyographic (EMG) signals collected from the orofacial neuromuscular system during the silent articulation of speech in an alaryngeal manner can be synthesized into personalized audible speech, potentially enabling individuals without vocal function to communicate naturally. Moreover, such systems could seamlessly interface with virtual environments where audible communication may be disruptive (e.g., multiplayer games) or to facilitate telephonic conversations in noisy settings. A key enabler of these advancements is the rich information encoded in EMG signals recorded from multiple spatially distributed locations, capturing muscle activation patterns across different muscles. This richness allows for the decoding of subtle and intricate articulatory details, potentially offering higher bandwidth and lower latency compared to exocentric or allocentric modalities, such as video-based lip-to-speech synthesis. By leveraging this information, EMG-based systems offer a promising foundation for natural and efficient communication across a range of applications.

The works in Willett et al. (2023) and Metzger et al. (2023) present invasive speech brain-computer interfaces (BCI). While invasive methods are suitable for individuals with complete anarthria, e.g., due to advanced amyotrophic lateral sclerosis, our EMG-based non-invasive speech prosthesis is appropriate for individuals with a broad range of speech impairments including dysarthria and dysphonia/aphonia, e.g., in those who have undergone laryngectomy. Work in Défossez et al. (2023) demonstrates a non-invasive BCI in which listened speech segments are reconstructed from magnetoencephalography (MEG) or electroencephalography (EEG) signals. However, such systems are not suitable for initiating communication (e.g., through speech).

Unlike invasive methods (Willett et al., 2023; Metzger et al., 2023), which can record neural activity at single-neuron resolution with high signal-to-noise ratios, EMG captures the aggregated activity of multiple muscle motor units, with signals further distorted as they propagate through the subcutaneous tissue and skin. These distortions lead to spatial signal correlations across electrodes, where activity at one sensor can influence measurements at others. To model this structure, we introduce symmetric positive definite (SPD) matrix representations that encode second-order inter-channel correlations, providing a compact and discriminative representation of EMG signals. In contrast to prior approaches (Défossez et al., 2023; Gaddy and Klein, 2020, 2021), which learn repre-

[1] We will open-source all data and model checkpoints upon publication of the manuscript. Due to IRB restrictions, we are unable to release the dataset anonymously.

sentations by mapping time-aligned MEG, EEG, or EMG signals to corresponding audio, we further improve the translation pipeline by directly predicting phoneme sequences from EMG without requiring time-aligned audio. This is achieved using connectionist temporal classification (CTC) loss (Graves et al., 2006), enabling alignment-free sequence prediction akin to standard speech-to-text (S2T) translation.

2 Prior work

The current benchmark for silent speech interfaces is established by Gaddy and Klein (2020, 2021). In these works, electromyographic (EMG) signals recorded during *silently* articulated speech (E_S) and *audibly* articulated speech (E_A), together with the corresponding audio (A), are used to train recurrent neural transduction models. These models learn a mapping from *time-aligned* EMG features (E_A or E_S) to audio (A). In the baseline formulation, joint representations between E_A and A are learned during training and subsequently evaluated on E_S. An improved variant further aligns E_S with E_A and uses the aligned features to strengthen the learned EMG-audio representation.

Despite strong performance, these approaches have several fundamental limitations that restrict their applicability in real-world clinical settings. Specifically, they require: ① access to high-quality E_A and audio A, which may be unavailable or unreliable in individuals with impaired articulation, such as laryngectomy (absence of laryngeal voicing) or ALS (degraded acoustic recordings due to bulbar impairment); ② the need for a *2x* sized training corpus for learning *x* representations (requiring both E_A and E_S); and ③ accurate temporal alignment between EMG and audio streams, which is computationally expensive and difficult to obtain robustly, thereby limiting scalability and near real-time deployment. In contrast, our approach eliminates these dependencies entirely by training directly on E_S paired only with phonemic transcriptions, without any EMG-audio alignment, using the CTC objective.

A geometric perspective on EMG representation is introduced in Gowda et al. (2024). That work shows that, unlike images or audio signals, which are functions sampled on Euclidean grids, multichannel EMG signals are more naturally modeled through covariance structure, whose intrinsic geometry lies on the manifold of symmetric positive-definite (SPD) matrices. While Gowda et al. (2024) focus primarily on classification of isolated articulatory gestures or phoneme segments, we extend this framework to sequence-to-sequence EMG-to-phoneme modeling, enabling continuous speech decoding. We present a detailed literature review and broad comparisons with other brain-computer interfaces (BCI) in appendix A.

2.1 Our contribution

We make two primary contributions.

First, we open-source one of the largest high-quality EMG-to-speech datasets collected during silent speech articulation (E_S). The dataset comprises approximately 8 hours of EMG speech data from a healthy participant, covering a large-vocabulary corpus with over 6500 unique words. To the best of our knowledge, it is among the most comprehensive publicly available resources for EMG-to-speech research to date.

Second, we demonstrate that symmetric positive definite (SPD) matrices provide a natural and sufficient *spatial representation* of EMG for EMG-to-text decoding. Motivated by the physiological view of EMG as arising from the additive superposition of motor unit action potentials (Farina et al., 2014), we use SPD matrices to model the spatial structure of multichannel muscle activity. Temporal dynamics are then captured using a simple vanilla GRU operating on eigenvalue-based representations of the SPD matrices, followed by CTC loss. This straightforward architecture aligns with modern speech-to-text modeling paradigms while remaining physiologically interpretable and enabling robust phoneme-by-phoneme decoding.

Our system is trained using only silently articulated EMG signals and their corresponding text transcriptions, and performs phoneme-level decoding directly from EMG followed by phoneme-to-word transcription. Unlike prior EMG-to-speech systems (Gaddy and Klein, 2020, 2021; Benster et al., 2024), our approach does not assume access to time-aligned EMG-audio pairs at any stage. Crucially, these results provide evidence that linguistically meaningful speech structure can be inferred from muscle activity alone and transcribed into words using only EMG.

Because our setting targets unaligned EMG-to-text generation without parallel audio supervision, there are no existing benchmarks that enable direct one-to-one comparisons. Nevertheless, we compare our methods against baselines from

EMG2QWERTY (Sivakumar et al., 2024). This comparison is well-motivated because both tasks involve decoding discrete linguistic sequences from EMG using closely related modeling and decoding pipelines, making EMG2QWERTY a robust and widely used benchmark for contextualizing our methods.

3 Methods

EMG signals are collected by a set of sensors $\mathcal{V}$ and are functions of time t. A sequence of EMG signals E_S corresponding to silently articulated speech, associated with audio A and phonemic content L, is represented as $E_S = \{\mathbf{f}_v(t)\}_{\forall\, v \in \mathcal{V}}$. Here, $\mathbf{f}_v(t)$ denotes the EMG signal captured at a sensor node v as a function of time t. The audio signal A encodes both phonemic (lexical) content and expressive aspects of speech, such as volume, pitch, prosody, and intonation, while L represents purely the phonemic content—a sequence of phonemes. For instance, the phonemic content L of the word <FRIDAY> is denoted by the phoneme sequence <F-R-IY-D-AY>.

To model the mapping from E_S to L, we employ a sequence-to-sequence model trained using CTC loss. This approach allows us to train the model with *unaligned* pairs of E_S and L, eliminating the need for precise alignment between the input signals and their corresponding phoneme sequences. During testing, a sample of E_S not in the training set outputs probabilities over all possible phonemes (40 of them in our case) at every time step, and we construct L using beam search. L is then converted to words using a language model.

3.1 EMG data representation

Gowda et al. (2024) demonstrate that the manifold of SPD matrices serves as an effective embedding space for EMG signals, enabling the natural distinction of different orofacial movements associated with speech articulation and all English phonemes using raw signals. We make significant improvements on their methods to perform phoneme-by-phoneme decoding as opposed to classification paradigms and demonstrate our methods on continuously articulated speech in the English language as opposed to discrete word or phoneme articulations.

We construct a complete graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}(\tau))$ to represent the functional connectivity of EMG signals, where $\mathcal{E}(\tau)$ denotes the set of edges over a time window $\tau = [t_{\text{START}}, t_{\text{END}}]$. The edge weight between two nodes v_1 and $v_2 \in \mathcal{V}$ within this time window is defined as $e_{12} = e_{21} = \frac{1}{\tau}\mathbf{f}_{v_1}^T \mathbf{f}_{v_2}$, which corresponds to the covariance of the signals at those nodes during the interval. Consequently, the edge (adjacency) matrix $\mathcal{E}(\tau)$ is symmetric and positive semi-definite. To ensure positive definiteness, we convert the semi-definite adjacency matrices to definite matrices by applying the transformation $\mathcal{E} \leftarrow (1-\eta)\mathcal{E} + \eta\, \texttt{trace}(\mathcal{E})\, \mathcal{I}$, where $\mathcal{I}$ is the identity matrix of the same dimension as $\mathcal{E}$. We empirically found that $\eta = 0.1$ suffices for all our data. We then model these symmetric positive definite (SPD) matrices using a Riemannian geometry approach via Cholesky decomposition, as described by Lin (2019).

For any adjacency matrix $\mathcal{E}$, we can express it as $\mathcal{E} = U\Sigma U^T$, where U is the matrix of eigenvectors, and Σ is a diagonal matrix containing the corresponding eigenvalues. However, instead of calculating U for each $\mathcal{E}$ at every time-step τ, we fix an approximate common eigenbasis Q derived from the Fréchet mean $\mathcal{F}$ (Lin, 2019) of all adjacency matrices (at different time points) in the training set. Specifically, we compute $\mathcal{F}$ as the geometric mean of all $\mathcal{E}$, and decompose it as $\mathcal{F} = Q\Lambda Q^T$, where Q contains the eigenvectors of $\mathcal{F}$, and Λ is a diagonal matrix of its eigenvalues.

Using this fixed eigenbasis Q, any adjacency matrix $\mathcal{E}$ can be approximately diagonalized as $Q^T\mathcal{E}Q$, yielding a sparse matrix σ (see figure 7). This formulation allows us to work in an approximate graph spectral domain with a consistent orthogonal basis across all time windows τ. For our task, we compute the graph spectral sequences σ for all time windows τ and use these as inputs for EMG-to-language translation. We illustrate these concepts in figure 1.

Fréchet mean: Given a set of SPD edge matrices $\mathcal{E}(\tau)$ over different time windows τ, we first calculate their corresponding Cholesky decompositions $\mathcal{L}(\tau) = \text{CHOLESKY}(\mathcal{E}(\tau))$, such that $\mathcal{E}(\tau) = \mathcal{L}(\tau)\mathcal{L}(\tau)^T$. Then, the Fréchet mean of the Cholesky decomposed matrices $\mathcal{L}(\tau)$ is given by

$$\mathcal{F}_{\text{CHOLESKY}} = \frac{1}{n}\sum_{i=1}^{n} \lfloor \mathcal{L}(\tau_i) \rfloor + \exp\left(\frac{1}{n}\sum_{i=1}^{n} \log\big(\mathbb{D}(\mathcal{L}(\tau_i))\big)\right).$$

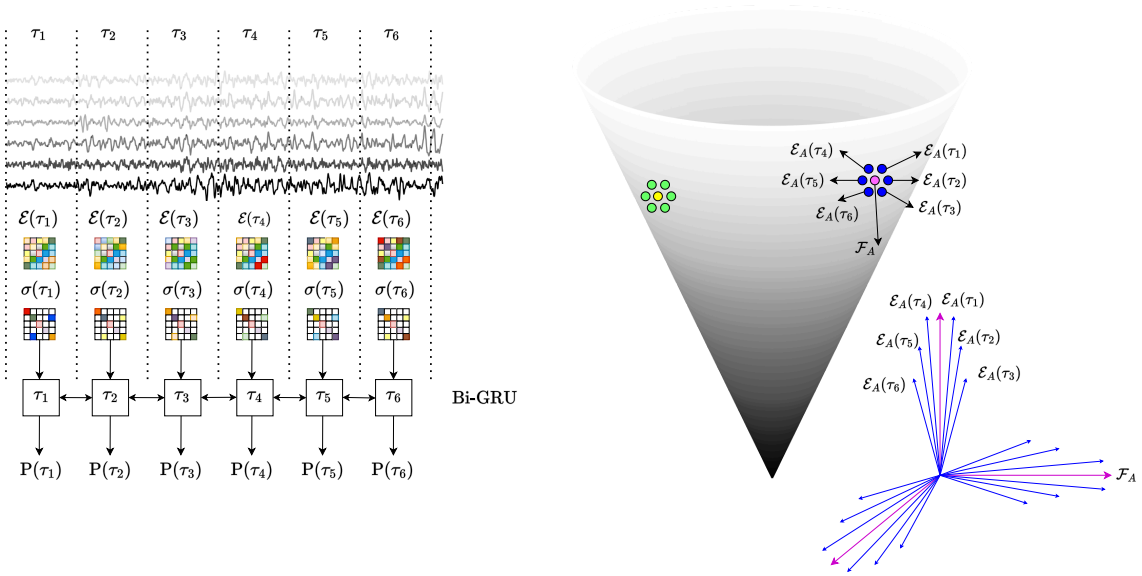


Figure 1: LEFT: EMG-to-phoneme translation pipeline. Bandpass-filtered and z-normalized EMG signals are converted into SPD edge matrices $\mathcal{E}(\tau)$, which are approximately diagonalized to $\sigma(\tau)$ and passed through a BiGRU. The model outputs phoneme probabilities $P(\tau)$ every 20 ms. The most probable phoneme sequence is decoded using beam search. RIGHT: Illustration of the geometry of SPD matrices in 3D. Edge matrices from individuals A (blue) and B (green) are shown on a convex cone manifold, with their corresponding Fréchet means in purple and yellow, respectively. The tangent spaces at A and B differ (because the surface is curved), and the induced transformations in $\mathbb{R}^{|\mathcal{V}|}$ reflect a change of basis. Inset: eigenvectors of individual A.

The Fréchet mean $\mathcal{F}$ on the manifold of SPD matrices is calculated as

$$\mathcal{F} = \mathcal{F}_{\text{CHOLESKY}} \mathcal{F}_{\text{CHOLESKY}}^{T}.$$

In the above equation, $\lfloor \mathcal{L}(\tau) \rfloor$ is the strictly lower triangular part of the matrix $\mathcal{L}(\tau)$, and $\mathbb{D}(\mathcal{L}(\tau))$ is the diagonal part of the matrix $\mathcal{L}(\tau)$.

3.2 EMG-to-phoneme sequence translation

We implement a gated recurrent unit (GRU) architecture for EMG-to-phoneme sequence-to-sequence modeling. The input to the GRU consists of a sequence of approximately diagonalized matrices, denoted as σ, derived over different time windows τ. At each time step, the GRU model outputs probability distributions over 40 phonemes in the English language. The model is trained using CTC loss, and during inference, the most probable phoneme sequence is reconstructed using beam search decoding. The end-to-end EMG-to-language translation model is depicted in figure 1.

3.3 Geometric perspective aligns well with biology

We model multivariate EMG signals recorded at $|\mathcal{V}|$ sensor nodes over different time windows τ using symmetric edge matrices $\mathcal{E}(\tau) \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$, which capture pairwise relationships between sensor channels. Each matrix $\mathcal{E}(\tau)$ can be interpreted as defining a linear transformation of the sensor space $\mathbb{R}^{|\mathcal{V}|}$, reflecting the spatial structure of EMG activity at time τ. This transformation admits a spectral interpretation: when $\mathcal{E}(\tau)$ is symmetric, it can be diagonalized as

$$\mathcal{E}(\tau) = U \Sigma(\tau) U^{\top},$$

where U is an orthonormal matrix whose columns are the eigenvectors of $\mathcal{E}(\tau)$, and $\Sigma(\tau)$ is a diagonal matrix of eigenvalues. In this eigenbasis coordinate system, the transformation of space is expressed as a weighted combination of the eigenvectors, with the eigenvalues in $\Sigma(\tau)$ serving as scaling coefficients. To reduce variability across time and to enable sequential modeling, we fix an approximate eigenbasis $Q \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$, and project each edge matrix into this basis:

$$\sigma(\tau) = Q^{\top} \mathcal{E}(\tau) Q,$$

yielding an approximately diagonal matrix $\sigma(\tau)$. The diagonals of $\sigma(\tau)$ approximate the eigenvalues of $\mathcal{E}(\tau)$ in the shared basis Q, providing a compact summary of the EMG activity at each time window. These sequences of approximate eigenvalues can

then be directly modeled using a recurrent neural network to capture temporal dynamics. This formulation aligns with the physiological origin of EMG signals: the surface EMG measurement arises from an additive superposition of motor unit action potentials, resulting in a structure that is naturally well-represented in an eigenbasis. This contrasts with modalities like speech audio, which are better modeled as time-varying filters applied to time-varying sources (Sivakumar et al., 2024). Importantly, the eigenbasis Q is subject-specific. EMG signals from different individuals induce different transformations $\mathcal{E}(\tau)$ and therefore have different eigenvectors, reflecting anatomical and physiological variability such as subcutaneous fat thickness, muscle fiber composition, conduction velocities, and neural drive. Consequently, distribution shifts across individuals can be interpreted as a *change of basis* in the sensor space $\mathbb{R}^{|\mathcal{V}|}$.

4 Data

We curate a large-vocabulary silent-speech EMG dataset with the number of articulated sentences comparable to those used in Willett et al. (2023); Metzger et al. (2023). We adapt the language corpora from Willett et al. (2023), which was originally developed for a speech brain-computer interface that translates motor-cortex neural activity into text. Our corpus contains approximately 6500 unique words and 11000 sentences. Unlike Gaddy and Klein (2020, 2021), we collect only silent-speech EMG (E_S), and do not collect audibly articulated EMG (E_A) or audio (A). As a result, our task is to translate E_S to language without relying on any time-aligned E_A or A supervision.

For phoneme-to-word decoding, we use a small weighted finite-state transducer (WFST) language model trained on transcripts from LibriSpeech-100 (Panayotov et al., 2015), which contain roughly 38000 sentences and 35000 unique words. See appendix C for additional experiments.

4.1 Experimental details

We record EMG signals from 31 sites distributed across the neck, chin, jaw, cheek, and lips using monopolar electrodes. Data are acquired using an ACTICHAMP PLUS amplifier with active electrodes from BRAIN VISION (Brain Vision), sampled at 5000 Hz. To ensure low-impedance contact between the electrodes and the skin surface, we apply SUPERVISC, a high-viscosity electrolyte gel from EASYCAP (Easycap). We develop a custom software suite in a PYTHON environment to present visual cues to participants and to collate and store timestamped EMG data. Time synchronization across data streams is handled using Lab Streaming Layer (LSL: LSL). Figure 2 illustrates the electrode placement. In addition to the 31 data electrodes, a GROUND electrode (marked GND) is placed on the left earlobe, and a REFERENCE electrode (marked as electrode 32) is placed on the right earlobe.

Before signal acquisition, participants are briefed on the experimental protocol and seated comfortably. For silent speech data (E_S), participants are instructed to articulate naturally but without producing audible speech. Sentence onset and offset are manually timestamped using mouse clicks by the participant. When ready to articulate a sentence, the participant clicks the mouse, causing the sentence to appear on the screen. After completing the articulation, the participant clicks again to mark the end of the sentence, at which point the sentence disappears from the display. This protocol allows participants to articulate each sentence at their own comfortable pace.

The data collection environment is carefully controlled to minimize AC electrical interference. EMG signals undergo minimal preprocessing. Specifically, the signal from the REFERENCE channel (electrode 32) is subtracted from all other channels. The resulting signals are bandpass filtered using a third-order Butterworth filter with cutoff frequencies of 80 and 1000 Hz, and segmented into individual sentences using synchronized start and end timestamps. Each segmented sentence is subsequently z-normalized along the time dimension on a per-channel basis. The preprocessed EMG signals are then used to construct a fully connected sensor graph, $\mathcal{E}(\tau)$, along with its approximately diagonalized representation, $\sigma(\tau)$.

The electrodes are positioned over anatomical regions that directly overlie muscle groups involved in speech articulation, providing coverage of key articulators such as the tongue, jaw, lips, and larynx. Electrode locations 19, 21, 3, and 1 approximately overlie the *hyoglossus*, *palatoglossus*, and *styloglossus* muscles. These muscles, located primarily in the lower cheek and tongue regions, play a central role in tongue shaping and movement and are consistently recruited across a wide range of articulatory gestures. Muscles in the upper and posterior cheek regions—such as the *masseter* and *temporalis*, which control jaw motion, and

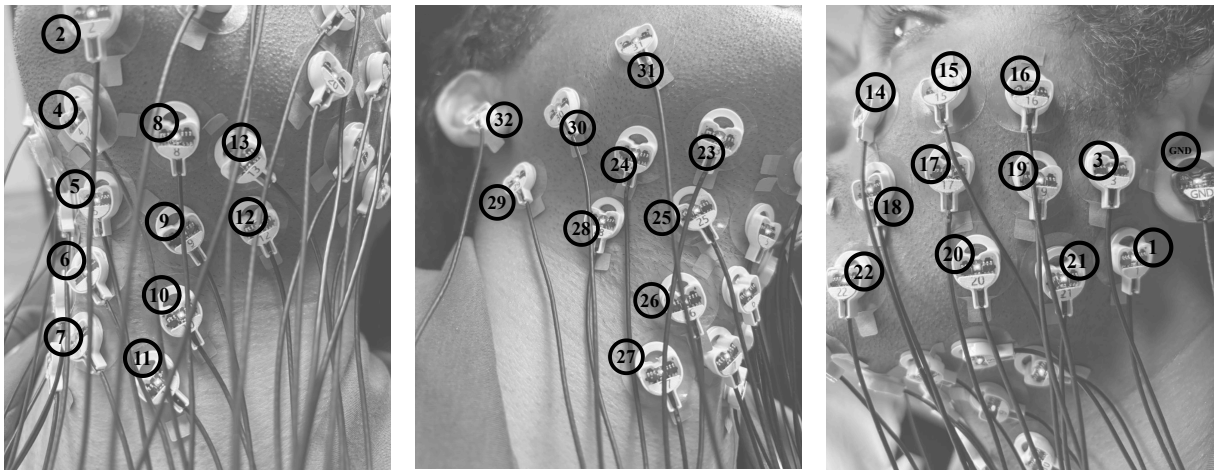


Figure 2: LEFT: Electrode placement on the left side of the neck. MIDDLE: Electrode placement on the right side of the neck. RIGHT: Electrode placement on the left cheek.

the *zygomaticus*, which contributes to upper lip elevation—correspond approximately to electrode regions around nodes 22, 18, 17, and 15 in figure 2. Electrodes located beneath the jaw capture activity from muscles involved in tongue protrusion and jaw-tongue coordination, including the *genioglossus* (near electrodes 8, 9, 23, and 25) and the *digastric*. Finally, electrodes near the laryngeal region (nodes 6, 7, 10, 11, 26, and 27) reflect activity from muscles that modulate laryngeal and hyoid position—such as the *sternohyoid*, *stylohyoid*, and *digastric*—which contribute to pitch modulation, vowel shaping, and coordinated jaw movement.

5 Results

We use a timestep τ of 20 ms, implemented as a sliding window with 50 ms of overlapping context and a 20 ms step size, to compute $\mathcal{E}(\tau)$ and $\sigma(\tau)$, both of which are SPD matrices of size 31×31. The matrices $\sigma(\tau)$ are then input to a GRU for EMG-to-phoneme sequence translation. The dataset is split into training, validation, and test sets consisting of 8000, 1000, and 1970 sentences, respectively. Sentences in the test set are not present in the training and validation sets. The model depicted in figure 1 is trained using 3 GRU layers for 100 epochs, and the weights corresponding to the lowest validation loss are selected.

In table 1, we report phoneme error rate (PER) and word error rate (WER), computed as the normalized Levenshtein distance between the reference and predicted sequences at the phoneme and word levels, respectively. To compute WER, we convert predicted phoneme sequences into word sequences using weighted finite-state transducer (WFST) decoding. We use transcripts from LibriSpeech-100 (Panayotov et al., 2015) (approximately 38000 sentences and 35000 unique words) to build the lexicon and language model. Following Mohri et al. (2008)[2], we compose the CTC topology FST H, lexicon FST L, and an n-gram language model FST G into a single decoding graph, $HLG = H \circ L \circ G$. Specifically, H encodes the allowable label sequences under the CTC criterion, L maps phoneme sequences to word sequences, and G is constructed from a 4-gram language model trained with KenLM (Heafield, 2011). At inference time, we perform beam search over HLG with a beam width of 50 to obtain the best-scoring word sequence.

For comparison with prior work, we derive EMG spectrograms, in which we match the temporal resolution to that of the SPD features (50 ms window and 20 ms hop). We compute STFT (short-time Fourier transform) with $n_{\text{FFT}} = 256$ (129 linear-frequency bins) and then average-pool the frequency axis down to 31 bins per channel. This produces per-frame tensors of shape 31 channels $\times$ 31 frequency bins, paralleling the 31×31 shape of $\sigma(\tau)$.

Unlike SPD matrices $\sigma(\tau)$, which encode cross-channel articulatory structure and allow a vanilla GRU to learn meaningful temporal dependencies, raw spectrograms do not provide an equivalent inductive bias. When we use spectrograms as GRU inputs, the model collapses to predicting a small set of phoneme sequences largely independent of the

[2]We use the WFST decoding implementation provided by ICEFALL (github.com/k2-fsa/icefall).

Table 1: Mean PER and WER. Lower values indicate better performance.

MODEL	PER(% ↓)	WER(% ↓)
BASELINE (SPECTROGRAM)	89.25	100
MATRICES $\sigma(\tau)$ (OURS)	**48.47**	**73.53**

input, which made phoneme-to-word decoding unreliable and resulted in a WER of 100% (table 1).[3]

In figure 3, we study how model capacity affects phoneme error rate (PER). We vary the number of GRU layers and the hidden-state dimensionality, which changes the total number of trainable parameters N. Across the range of models explored, PER decreases as N increases and is approximately consistent with a power-law-like trend, $PER \propto N^{-\beta}$ for $\beta > 0$, similar in spirit to empirical scaling behaviors reported for neural language models (Kaplan et al., 2020). The observed trend suggests a predictable relationship between model capacity and decoding accuracy in this setting. Notably, even a single-layer GRU attains a PER of 56%, with deeper and wider models yielding further improvements.

In figure 4, we examine how the amount of training data affects PER. We vary the number of training sentences M while keeping the validation and test sets fixed. PER again decreases with more data and is approximately consistent with a power-law-like trend over the range explored, $PER \propto M^{-\beta}$ for $\beta > 0$ (Kaplan et al., 2020). The result indicates a regular relationship between data availability and decoding accuracy in our EMG-to-phoneme setting. Together, these trends suggest that scaling both model capacity and data can improve performance, with diminishing returns at larger sizes.

[3]One might argue that spectrogram inputs could be made more amenable to recurrent modeling through careful normalization. To enable a fair comparison with SPD matrices, we keep the decoding pipeline identical across representations: we feed either SPD matrices $\sigma(\tau)$ or raw spectrogram features into the same GRU decoder. However, on the EMG2QWERTY benchmark (Sivakumar et al., 2024), we still find that $\sigma(\tau)$ outperforms normalized spectrogram features. Our primary motivation for this comparison is to highlight that, with an SPD-matrix representation, EMG-to-text translation is feasible even with an architecture that mainly models temporal dependencies and performs no explicit spatial modeling (i.e., a vanilla GRU), because cross-channel structure is already captured by second-order (covariance) features. In contrast, spectrogram features do not appear to confer the same inductive bias under an otherwise identical recurrent decoder. This result should be interpreted in that context.

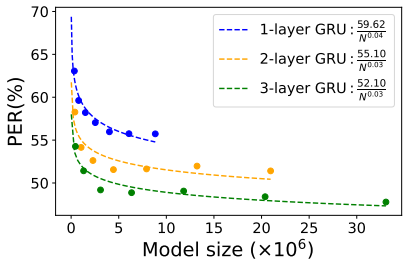


Figure 3: Model size versus PER for EMG-to-phoneme translation.

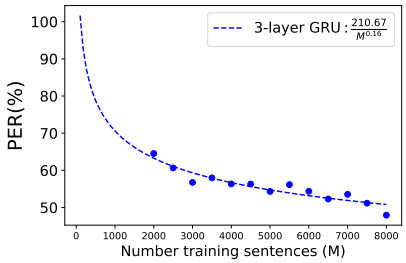


Figure 4: Training data size versus PER for EMG-to-phoneme translation.

5.1 Comparison with prior work

To the best of our knowledge, there is no prior work that performs E_S-to-language conversion without using E_A or A on large English language corpora with CTC loss. Therefore, we compare our methods on the EMG2QWERTY dataset introduced by Sivakumar et al. (2024). In this dataset, subjects wear EMG wristbands on both hands and touch-type on a QWERTY keyboard. The goal is to decode the resulting EMG signals into a sequence of characters using CTC loss. Although the physical actions involved in EMG-to-text decoding and EMG2QWERTY differ, the underlying machine learning principles remain similar.

To enable a fair comparison, we conduct controlled experiments in which we replace the original log-spectrogram features from Sivakumar et al. (2024) with SPD matrices $\sigma(\tau)$. Apart from substituting the features, we omit their SPECAUGMENT data augmentation strategy—this should not compromise the fairness of the comparison, as SPECAUGMENT was shown to improve their performance. Additionally, we train our models for 250 epochs (compared to 150 in their setup, where their model converged early), and apply a weight decay of 10^{-3} to the Adam optimizer to ensure stable training.

We focus on a specific case from Sivakumar et al. (2024), in which personalized models are

Table 2: Comparison between our proposed methods and those presented by (Sivakumar et al., 2024), with all results averaged over 8 subjects. Parameter size and FLOPs are identical across all the models. Lower CER is better. The CER improvement arising from our method is statistically significant ($p < 0.015$). LM: language model.

	No LM		6-gram char-LM	
	Val CER (% ↓)	Test CER (% ↓)	Val CER (% ↓)	Test CER (% ↓)
Baseline (spectrogram) (Sivakumar et al., 2024)	15.65 ± 5.95	15.38 ± 5.88	11.03 ± 4.45	9.55 ± 5.16
Matrices $\sigma(\tau)$ (ours)	$\mathbf{14.33 \pm 5.27}$	$\mathbf{14.03 \pm 5.27}$	$\mathbf{9.61 \pm 3.84}$	$\mathbf{7.95 \pm 4.54}$

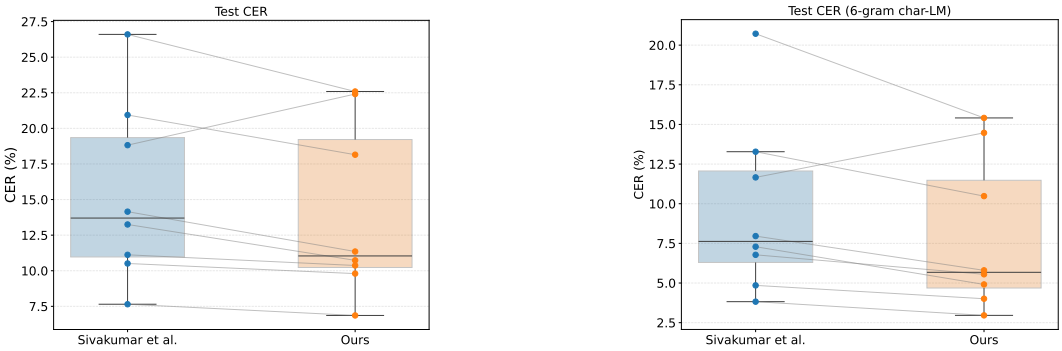


Figure 5: Results for individual subjects in EMG2QWERTY dataset. Each dot represents an individual test subject, with connecting lines indicating within-subject performance across different models. The boxplots summarize the median and interquartile range of the results. Our method improves performance for all subjects except USER6.

trained independently for each individual, starting from random weight initialization. The zero-shot paradigm, in which a model is trained on data from 100 subjects and evaluated on 8 unseen individuals, as well as the personalized fine-tuning paradigm, in which individual models are initialized with generic weights pretrained on 100 subjects, are beyond the scope of this work. In this paper, we restrict our investigation to personalized models trained from scratch.

The results are presented in table 2. As shown, our proposed methods outperform the baseline approaches reported by Sivakumar et al. (2024), with just *one simple* modification. These findings support the effectiveness of our approach, which is specifically designed to reflect the underlying biological structure of EMG signals.

In figure 5, we present subject-wise character error rates (CER). Our method improves performance for all users except USER6. When decoding is performed without a language model, we observe an 8.8% relative improvement in CER. With a 6-gram character level language model (6-gram char-LM), the relative improvement is 16.8%[4].

[4]For reference, in a personalized finetuning paradigm, Sivakumar et al. (2024) first trained a generic model on data from 100 subjects (nearly 100× more data) and then finetuned it on 8 individual subjects, achieving a CER of 11.29% without a language model and 6.95% with a 6-gram character LM. The 6.95% result reflects a strong performance ceiling enabled by large-scale pretraining. In comparison, our 7.95% CER is obtained using only per-subject training (approximately 100× less data) and is already close to this ceiling, highlighting the effectiveness of our approach.

6 Conclusion

We show that continuous speech can be inferred from orofacial EMG by decoding articulations *phoneme-by-phoneme*. This formulation is well matched to the physiology of speech production: phonemes are defined by place and manner of articulation, which should be expressed in coordinated patterns of orofacial muscle activity. Our approach is simple and interpretable: SPD matrices capture spatial structure in multichannel EMG, and a GRU models temporal dynamics. In our evaluation, we obtain a phoneme error rate (PER) of 49%, substantially below the chance-level PER of approximately 98%, providing strong evidence that direct EMG-to-text transcription is feasible. Although word error rate (WER) remains modest and experiments are currently limited to a single subject, these results establish a concrete baseline and motivate future work on improved modeling, stronger decoding, and broader validation across participants.

7 Limitations

This work primarily focuses on demonstrating that linguistic content can be decoded from EMG signals alone, phoneme-by-phoneme, and then reconstructed into words. To our knowledge, this has not previously been shown in the context of general English corpora using methodologies widely adopted in modern speech-to-text paradigms. We view this as a necessary milestone in a longer-term research effort on which future advances can build; however, several limitations remain.

First, our current model uses a bidirectional GRU, which requires access to the full sentence before decoding and therefore does not support streaming, low-latency use. In follow-up work (https://myoverse.github.io/), we address this limitation by using causal models that rely only on local past context and by directly converting EMG sequences to speech.

Second, this study is demonstrated on a single healthy participant, so the results do not yet establish robustness across individuals or clinical populations. In a follow-up study (https://myoverse.github.io/), we extend the same overall approach to an individual with amyotrophic lateral sclerosis (ALS).

Third, we do not evaluate sustained, long-term performance of this non-invasive neuroprosthesis across days or months, including the effects of electrode shifts, fatigue, and other sources of day-to-day variability. In contrast, prior work on invasive neuroprostheses has reported stability over extended periods in related decoding settings, including brain-to-text (Fan et al., 2023) and cursor-based brain-computer interfaces (Wilson et al., 2025).

Finally, we do not explore whether large-scale pretrained EMG models can improve decoding performance or reduce the amount of subject-specific data required. Related work on EMG-based keyboard typing (EMG2QWERTY) suggests that pretraining on data from many individuals can improve accuracy after fine-tuning, although zero-shot performance remains limited (Sivakumar et al., 2024). Speech is likely more challenging than discrete key typing, and future work should investigate how to build and effectively leverage large-scale pretrained models for EMG-to-speech translation.

We are actively addressing these limitations through ongoing longitudinal studies and by expanding data collection to build larger EMG-to-speech corpora from individuals with diverse clinical etiologies, including ALS and laryngectomy.

8 Ethical considerations

Research was conducted in accordance with the principles embodied in the Declaration of Helsinki and with approval from the home institution's Institutional Review Board (IRB). All participants provided written informed consent. All participants also provided consent for publication of deidentified data. Volunteers of any gender and from all racial and ethnic groups were eligible to participate. Participants were required to be at least 18 years old, able to understand spoken and written English, and able to follow task instructions. Participants had no skin conditions or wounds at electrode placement sites and were excluded if they had uncorrected vision problems. Children, individuals unable to provide informed consent, and prisoners were not included in the experiments.

References

Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. 2011. Multiclass brain–computer interface classification by riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928.

Alexandre Barachant, StéPhane Bonnet, Marco Congedo, and Christian Jutten. 2013. Classification of covariance matrices using a riemannian-based kernel for bci applications. *Neurocomput.*, 112:172–178.

Tyler Benster, Guy Wilson, Reshef Elisha, Francis R Willett, and Shaul Druckmann. 2024. A cross-modal approach to silent speech with llm-enhanced recognition. *arXiv preprint arXiv:2403.05583*.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. 2023. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10):1097–1107.

Lorenz Diener, Gerrit Felsch, Miguel Angrick, and Tanja Schultz. 2018. Session-independent array-based emg-to-speech conversion using convolutional neural networks. In *Speech Communication; 13th ITG-Symposium*, pages 1–5.

Chaofei Fan, Nick Hahn, Foram Kamdar, Donald Avansino, Guy Wilson, Leigh Hochberg, Krishna V Shenoy, Jaimie Henderson, and Francis Willett. 2023. Plug-and-play stability for intracortical brain-computer interfaces: a one-year demonstration of seamless brain-to-text communication. *Advances in neural information processing systems*, 36:42258–42270.

Dario Farina, Roberto Merletti, and Roger M. Enoka. 2014. The extraction of neural strategies from the surface emg: an update. *Journal of Applied Physiology*, 117(11):1215–1230. Epub 2014 Oct 2.

David Gaddy and Dan Klein. 2020. Digital voicing of silent speech. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5521–5530.

David Gaddy and Dan Klein. 2021. An improved model for voicing silent speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 175–181.

Harshavardhana T Gowda, Zachary D McNaughton, and Lee M Miller. 2024. Geometry of orofacial neuromuscular signals: speech articulation decoding using surface electromyography. *Journal of Neural Engineering*.

Harshavardhana T Gowda and Lee M Miller. 2024. Topology of surface electromyogram signals: hand gesture decoding on riemannian manifolds. *Journal of Neural Engineering*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Matthias Janke and Lorenz Diener. 2017. Emg-to-speech: Direct generation of speech from facial electromyographic signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2375–2385.

Szu-Chen Jou, Tanja Schultz, Matthias Walliczek, Florian Kraft, and Alex Waibel. 2006. Towards continuous speech recognition using surface electromyography. In *Ninth International Conference on Spoken Language Processing*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Arnav Kapur, Utkarsh Sarawgi, Eric Wadkins, Matthew Wu, Nora Hollenstein, and Pattie Maes. 2020. Non-invasive silent speech recognition in multiple sclerosis with dysphonia. In *Machine Learning for Health Workshop*, pages 25–38. PMLR.

Zhenhua Lin. 2019. Riemannian geometry of symmetric positive definite matrices via cholesky decomposition. *SIAM Journal on Matrix Analysis and Applications*, 40(4):1353–1370.

Kaylo T Littlejohn, Cheol Jun Cho, Jessie R Liu, Alexander B Silva, Bohan Yu, Vanessa R Anderson, Cady M Kurtz-Miott, Samantha Brosler, Anshul P Kashyap, Irina P Hallinan, and 1 others. 2025. A streaming brain-to-voice neuroprosthesis to restore naturalistic communication. *Nature neuroscience*, pages 1–11.

Geoffrey S Meltzner, James T Heaton, Yunbin Deng, Gianluca De Luca, Serge H Roy, and Joshua C Kline. 2018. Development of semg sensors and algorithms for silent speech recognition. *Journal of neural engineering*, 15(4):046031.

Sean L Metzger, Kaylo T Littlejohn, Alexander B Silva, David A Moses, Margaret P Seaton, Ran Wang, Maximilian E Dougherty, Jessie R Liu, Peter Wu, Michael A Berger, and 1 others. 2023. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620(7976):1037–1046.

Mehryar Mohri, Fernando C. N. Pereira, and Michael Riley. 2008. Speech recognition with weighted finite-state transducers. In *Handbook on Speech Processing and Speech Communication, Part E: Speech recognition*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

David Sabbagh, Pierre Ablin, Gaël Varoquaux, Alexandre Gramfort, and Denis A. Engemann. 2019. *Manifold-regression to predict from MEG/EEG brain signals without source modeling*. Curran Associates Inc., Red Hook, NY, USA.

Viswanath Sivakumar, Jeffrey Seely, Alan Du, Sean R Bittner, Adam Berenzweig, Anuoluwapo Bolarinwa, Alexandre Gramfort, and Michael I Mandel. 2024. emg2qwerty: A large dataset with baselines for touch typing using surface electromyography. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Arthur R. Toth, Michael Wand, and Tanja Schultz. 2009. Synthesizing speech from electromyography using voice transformation techniques. In *Interspeech 2009*, pages 652–655.

Francis R Willett, Erin M Kunz, Chaofei Fan, Donald T Avansino, Guy H Wilson, Eun Young Choi, Foram Kamdar, Matthew F Glasser, Leigh R Hochberg, Shaul Druckmann, and 1 others. 2023. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036.

Guy H Wilson, Elias A Stein, Foram Kamdar, Donald T Avansino, Tsam Kiu Pun, Ronnie Gross, Tommy

Hosman, Tyler Singer-Clark, Anastasia Kapitonava, Leigh R Hochberg, and 1 others. 2025. Long-term unsupervised recalibration of cursor-based intracortical brain–computer interfaces using a hidden markov model. *Nature Biomedical Engineering*, pages 1–19.

A Detailed literature review

Here, we review prior work on speech neural and neuromuscular interfaces and contextualize our results relative to state-of-the-art methods. A substantial body of research (Jou et al., 2006; Kapur et al., 2020; Meltzner et al., 2018; Toth et al., 2009; Janke and Diener, 2017; Diener et al., 2018; Littlejohn et al., 2025) has laid the groundwork for EMG-based speech interfaces. Among the earliest studies, Jou et al. (2006) demonstrate EMG-to-speech conversion on a small corpus of 50 sentences. Kapur et al. (2020) use a corpus of 15 sentences and, rather than performing phoneme-level decoding, formulate the task as a 15-way classification problem. Meltzner et al. (2018) study EMG-to-text recognition for isolated words, phrases drawn from a ∼200-word vocabulary, and continuous sentences using a custom grammar-based recognition model over a set of 1200 scripted phrases. Toth et al. (2009) present EMG-to-speech conversion on a corpus of 500 sentences. Janke and Diener (2017) demonstrate EMG-to-speech conversion using up to two hours of data and 2000 utterances.

Overall, these studies rely on private datasets and task-specific pipelines, and they typically evaluate on small, constrained corpora. In addition, the works do not release full implementations (e.g., code repositories) or sufficient methodological details to enable direct reproducibility. As a result, it is difficult to directly compare performance across systems, and all the above results do not establish generalization to open-vocabulary English settings.

A reproducible benchmark for open-vocabulary EMG-to-speech conversion was introduced by Gaddy and Klein (2020, 2021). However, these works rely on time-aligned EMG-audio pairs for training. Building on Gaddy and Klein (2021), Benster et al. (2024) propose an approach that leverages an audio-only corpus in addition to paired EMG-audio data. *While effective in the benchmark setting*, such methods cannot be deployed in clinical scenarios where parallel EMG-audio recordings may be unavailable or unreliable. On the large-vocabulary corpus, Gaddy and Klein (2020) report a word error rate (WER) of 68%, and Gaddy and Klein (2021) reduce this to 42%. Littlejohn et al. (2025) report a WER of 74% on the Gaddy and Klein (2020) dataset using a CNN+RNN transducer model; however, their train-test splits and implementation details are not publicly available, which prevents direct comparison. In our setting, we address a harder learning problem by not assuming time-aligned EMG-audio pairs during training, and we report a WER of 73% on an open-vocabulary corpus. We emphasize that these WER values should not be compared one-to-one across studies, since the data collection setup, training targets and alignment assumptions, problem formulation, and evaluation methodology differ substantially. We report these results to provide context relative to prior EMG-based speech interfaces.

To address these limitations, we build on widely used methods in speech-to-text (S2T) domain by adapting them to the EMG setting through principled, articulatorily motivated design choices.

Previous work by Gowda and Miller (2024) demonstrated the effectiveness of SPD matrices in decoding *discrete* hand gestures from EMG signals collected from the upper limb. Furthermore, SPD matrix representations have been extensively utilized to model electroencephalogram (EEG) signals, although they have never been applied to complex tasks such as sequence-to-sequence speech decoding. For example, Barachant et al. (2011, 2013) employed Riemannian geometry frameworks for classification tasks in EEG-based brain-computer interfaces, while Sabbagh et al. (2019) developed regression models based on Riemannian geometry for biomarker exploration using EEG data.

The novelty of our work lies in the algebraic interpretation of manifold-valued data through linear transformations, and the development of models for complex sequence-to-sequence tasks. This approach moves beyond the conventional applications of classification and regression.

B Additional results

In figure 6, we summarize the distribution of WER and PER across all 1970 sentences in the test set. The mean PER is 48.47%, far below the chance-level PER of approximately $1 - \frac{1}{40} = 97.5\%$ under uniform random guessing over 40 phoneme labels. The mean WER is 73.53%, and the gap between WER and PER suggests that, even when word-level transcriptions are incorrect, the predicted sequences often remain phonetically plausible. We show qualitative EMG-to-text transcription exam-

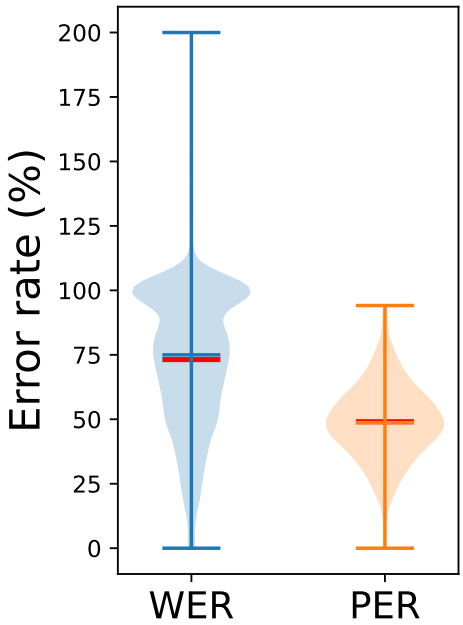


Figure 6: Distributions of WER and PER across all 1970 sentences in the test set. Means are shown in red.

ples in table 5.

$\sigma(\tau)$ **are sparse matrices:** in figure 7, we show that $\sigma(\tau)$ is sparser than $\mathcal{E}(\tau)$; its off-diagonal entries are small relative to its diagonal, i.e., $\sigma(\tau)$ is closer to a diagonal matrix than $\mathcal{E}(\tau)$.

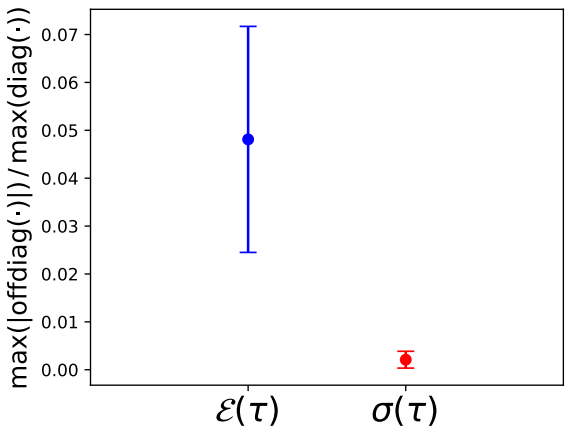


Figure 7: Blue: dataset average of $\frac{\max\left(|\text{offdiag}(\mathcal{E}(\tau))|\right)}{\max\left(\text{diag}(\mathcal{E}(\tau))\right)}$ over all τ in the train, validation, and test sets. Red: dataset average of $\frac{\max\left(|\text{offdiag}(\sigma(\tau))|\right)}{\max\left(\text{diag}(\sigma(\tau))\right)}$ over all τ in the train, validation, and test sets. The consistently lower ratio for $\sigma(\tau)$ indicates that it is closer to diagonal (and thus sparser) than $\mathcal{E}(\tau)$. We use the sparse SPD matrices $\sigma(\tau)$ for EMG-to-text translation.

C Additional experiments

Here, we present additional experiments on small-vocabulary data and on recordings from multiple subjects. The motivation for using small-vocabulary data is to test whether we can achieve high decoding accuracy in a closed-vocabulary setting, supporting a minimum-viable neuroprosthesis. The motivation for evaluating multiple subjects is to assess whether our approach generalizes across individuals.

We curate DATA $_{\text{SMALL-VOCAB}}$, a timestamped dataset of isolated and connected words, and DATA $_{\text{NATO-WORDS}}$, a compact codeword dataset based on the NATO phonetic alphabet that enables training a generalizable language-to-spelling model with minimal data.

C.1 DATA $_{\text{SMALL-VOCAB}}$

We curate a limited-vocabulary dataset consisting of 67 unique words. These words include weekdays, ordinal dates, months, and years. Sentences are constructed in the format <WEEKDAY-MONTH-DATE-YEAR>. A single participant silently articulated 500 such sentences, and the resulting EMG data, denoted as E_S, are translated into output phoneme sequences. We have timestamps that mark the beginning and end of each word (or grouped words) within a sentence. We record EMG from 31 electrode sites at a sampling rate of 5000 Hz. For details about electrode placement and the experimental setup, please refer to section 4.1. The sentences were presented as individual words (or grouped words), demarcated by timestamps, and displayed as follows:

$$\underset{t=0}{<} \text{WEEKDAY} \underset{t=2\,\text{s}}{>} - \underset{t=2\,\text{s}}{<} \text{MONTH} \underset{t=4\,\text{s}}{>}$$

$$\underset{t=4\,\text{s}}{<} \text{DATE} \underset{t=6\,\text{s}}{>} - \underset{t=6\,\text{s}}{<} \text{YEAR} \underset{t=9\,\text{s}}{>},$$

where each segment occurs sequentially in time.

Results: we use a timestep τ of 50 ms, implemented as a sliding window with 100 ms of overlapping context and a 50 ms step size, to compute $\mathcal{E}(\tau)$ and $\sigma(\tau)$, both of which are SPD matrices of size 31×31. The matrices $\sigma(\tau)$ are then input to a GRU for EMG-to-phoneme sequence translation. The dataset is split into training, validation, and test sets consisting of 370, 30, and 100 sentences, respectively. The model depicted in figure 1 is trained using a single GRU layer for 100 epochs, and the weights

corresponding to the lowest validation loss are selected.

In table 3, we report the phoneme error rate (PER) and word error rate (WER), computed using the Levenshtein distance between the original and reconstructed sequences. Words are reconstructed from phoneme sequences by matching them to the word sequence with the lowest Levenshtein distance in a 67-word corpus.

Table 3: Mean PER and WER for DATA SMALL-VOCAB. Lower values indicate better performance.

PER(% ↓)	WER(% ↓)
13	14

In figure 8, we analyze the impact of model size on phoneme error rate (PER) across different GRU configurations by varying the dimensionality of the GRU's hidden units. We observe that the relationship between PER and model size approximately follows a power-law-like trend similar to figure 3.

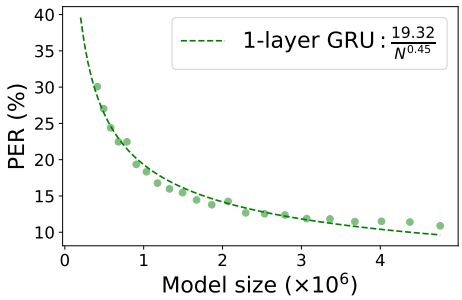


Figure 8: Model size versus PER for EMG-to-phoneme translation for DATA SMALL-VOCAB.

C.2 DATA NATO-WORDS

In this dataset, 4 individuals silently articulated English sentences in a spelled-out format using NATO phonetic codewords. For example, the word <RAINBOW> was articulated as <ROMEO-ALFA-INDIA-NOVEMBER-BRAVO-OSCAR-WHISKEY>, with phonemic transcription <R-OW-M-IY-OW SPACE AE-L-F-AH SPACE IH-N-D-IY-AH SPACE N-OW-V-EH-M-B-ER SPACE B-R-AA-V-OW SPACE AO-S-K-ER SPACE W-IH-S-K-IY>. Subjects articulated the phonetically balanced RAINBOW and GRANDFATHER passages in this spelled-out format. In total, 1968 NATO codeword articulations were recorded across both passages, along with an additional 520 isolated codeword recordings used for training. EMG was recorded from 22 sites on the neck and cheek at a sampling rate of 5000 Hz (electrodes were not placed on the right side of the neck; middle image in figure 2).

Results: we use a timestep τ of 30 ms, implemented as a sliding window with 150 ms of overlapping context and a 30 ms step size, to compute $\mathcal{E}(\tau)$ and $\sigma(\tau)$, both of which are SPD matrices of size 22×22. The matrices $\sigma(\tau)$ are then input to a GRU for EMG-to-phoneme sequence translation. The dataset is split into training, validation, and test sets consisting of 416, 104, and 1968 articulations, respectively. The model depicted in figure 1 is trained using a single GRU layer for 100 epochs, and the weights corresponding to the lowest validation loss are selected.

In table 4, we report the character error rate (CER). For a given character articulation—for example, <R>, which corresponds to the spoken form <ROMEO: R-OW-M-IY-OW>—we consider the decoded character to be <R> if the predicted phoneme sequence most closely matches that of <R> among the 26 alphabet characters. It is worth noting that the test set is nearly five times larger than the training set. This experimental paradigm is designed to evaluate whether a model can be trained effectively using very limited data—an important consideration for clinical applications, where collecting large amounts of data can be too strenuous for patients. In our case, the model is trained on just 10 minutes of data and evaluated on 50 minutes of data.

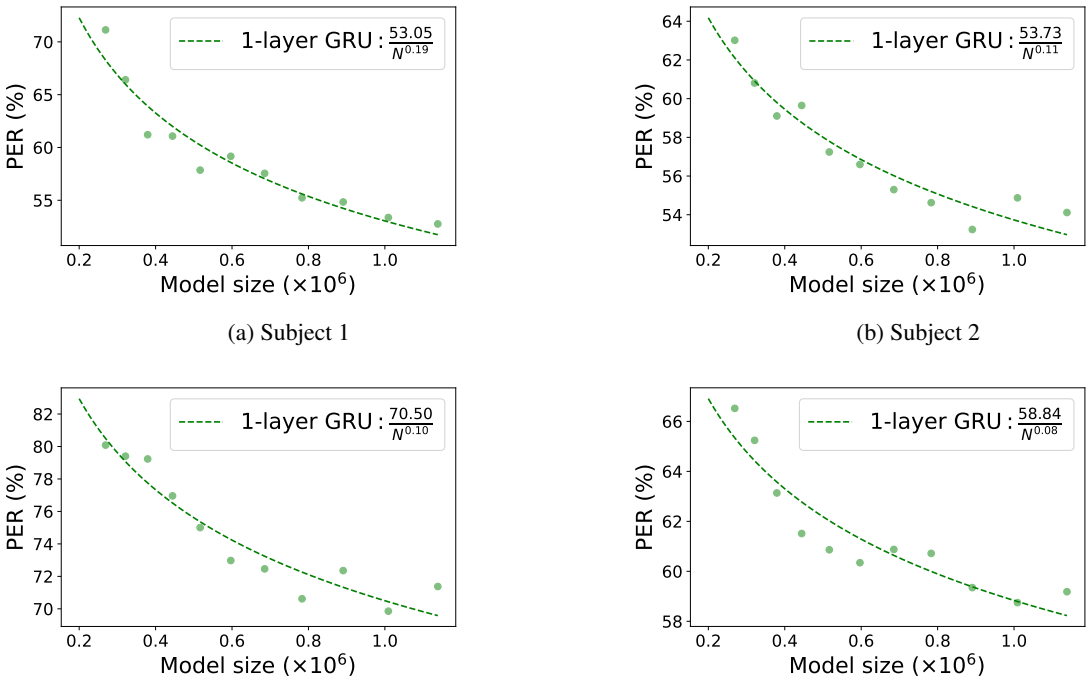
Table 4: Mean character error rate (CER) on DATA NATO-WORDS. Lower CER indicates better performance. The chance-level CER is 96%, and all subjects achieve substantially lower error rates.

SUBJECT	CER (% ↓)
1	55.7
2	55.0
3	70.4
4	56.4

In figure 9, we examine how model size across various GRU configurations affects the PER. To do this, we vary the dimensionality of the GRU's hidden units. We observe similar trends as noted in figure 3 across all subjects.

C.3 Discussion

These results indicate that accurate decoding is achievable in a small-vocabulary setting, suggesting that a minimum-viable neuromuscular speech



(a) Subject 1

(b) Subject 2

(c) Subject 3

(d) Subject 4

Figure 9: Model size versus PER for EMG-to-phoneme translation for DATA $_{\text{NATO-WORDS}}$.

prosthesis may be feasible even with limited training data.

Table 5: Examples of EMG-to-phoneme sequence translations. We do translations using EMG collected during *silent* articulations (E_S) with CTC loss without making use of corresponding time aligned *audio* (A) and EMG collected during *audible* articulation (E_A). Ground truth sentences with corresponding timestamps. Ground truth phonemic transcriptions. Decoded phonemic transcriptions. Decoded sentences.

Top-3 (best) transcribed sentences
T-START <IT WAS PAID FOR>T-END IH-T SPACE W-AA-Z SPACE P-EY-D SPACE F-AO-R IH-T SPACE W-AA-Z SPACE P-EY-T SPACE F-AO-R IT WAS PAY FOR
T-START <IT'S A COMMUNITY CENTER>T-END IH-T-S SPACE AH SPACE K-AH-M-Y-UW-N-AH-T-IY SPACE S-EH-N-T-ER IH-T-S SPACE AH SPACE K-AH-M-Y-UW-N-IH-T-IY SPACE S-EH-N-T-ER-N IT'S A COMMUNITY CENTER
T-START <JUST ALL DIFFERENT COLORS>T-END J-AH-S-T SPACE AO-L SPACE D-IH-F-ER-AH-N-T SPACE K-AH-L-ER-Z J-AH-S-T SPACE AO-L SPACE D-IH-F-ER-AH-N SPACE SPACE K-IH-L-ER-Z JUST ALL DIFFERENT COLORS
Bottom-3 (worst) transcribed sentences.
T-START <THE DEATH PENALTY>T-END DH-AH SPACE D-EH-TH SPACE P-EH-N-AH-L-T-IY IH SPACE DH-IH-T SPACE IH-K SPACE P-AY SPACE AE-K THAT THICK MY BACK
T-START <HE DOES THE YARD>T-END HH-IY SPACE D-AH-Z SPACE DH-AH SPACE Y-AA-R-D IH-IH-T SPACE IH-S SPACE N-IH-N-T SPACE AY-T IT ITS KNIT MIGHT
T-START <THAT'S AWFUL>T-END TH-AE-T-S SPACE AA-F-AH-L DH-EH-R SPACE AH SPACE T-OY-T THERE A POINT