Non-invasive electromyographic speech neuroprosthesis: a geometric perspective

Harshavardhana T. Gowda **Ferdous Rahimi** **Lee M. Miller**
University of California, Davis
tgharshavardhana@gmail.com
frahimi@ucdavis.edu, leemiller@ucdavis.edu

Abstract

In this article, we present a high-bandwidth *egocentric* neuromuscular speech interface for translating silently voiced speech articulations into text. Specifically, we collect electromyographic (EMG) signals from multiple articulatory sites on the face and neck as individuals articulate speech in an alaryngeal manner to perform EMG-to-language translation. Such an interface is useful for restoring audible speech in individuals who have lost the ability to speak intelligibly due to laryngectomy, neuromuscular disease, stroke, or trauma-induced damage (e.g., from radiotherapy toxicity) to the speech articulators. Previous works have focused on training text or speech synthesis models by mapping EMG collected during *audible* speech articulation to corresponding time-aligned audio, or by transferring time-aligned audio targets from EMG collected during *audible* articulation to EMG collected during *silent* articulation. However, such paradigms are not suitable for individuals who have already lost the ability to *audibly* articulate speech. Here, we present an alignment-free EMG-to-language conversion approach using only EMG collected during *silently* articulated speech. Our method is trained on a large, general-domain English language corpus and is released in an open-sourced manner.

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 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.

Willett et al. (2023) and Metzger et al. (2023) present invasive speech brain computer interfaces (BCI). While invasive methods are viable for individuals with anarthria or amyotrophic lateral sclerosis, our EMG-based non-invasive speech prosthesis is appropriate for individuals who have undergone laryngectomy or experience dysarthria or dysphonia. Défossez et al. (2023) demonstrate a non-invasive BCI where 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 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 con-

trast to prior approaches (Défossez et al., 2023; Gaddy and Klein, 2020, 2021), which learn representations 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 in silent speech interfaces is established by Gaddy and Klein (2020, 2021). Using electromyographic (EMG) signals collected during *silently* articulated speech (E_S) and *audibly* articulated speech (E_A), along with corresponding audio signals (A), they develop a recurrent neural transduction model to map time-aligned features of E_A or E_S with A. In their baseline model, joint representations between E_A and A are learned during training, and the model is tested on E_S. To improve performance, a refined model aligns E_S with E_A, and subsequently uses the aligned features to learn joint representations with A. The methods described above have significant shortcomings that limit their practicality for real-world deployment. These include: ① the unavailability of good-quality E_A and A in individuals who have lost vocal and articulatory functions; ② the need for a *2x* sized training corpus for learning *x* representations (requiring both E_A and E_S); and ③ the requirement for aligned features, which are computationally expensive and time-consuming to obtain, making near real-time implementation challenging. We overcome these challenges by training a model using only E_S and corresponding phonemic transcriptions, without any alignments, employing CTC loss.

Schultz and Wand (2010) demonstrate EMG-to-language modeling using only E_S, relying on hidden Markov models (HMMs) trained on a small 101-word vocabulary. However, they do not demonstrate the scalability of this approach to large-vocabulary corpora.

Benster et al. (2024) present a cross-modal approach to training EMG-to-language model, using contrastive loss functions that leverage both E_A and E_S, as well as audio-only corpora (A) such as LibriSpeech (Panayotov et al., 2015), to learn shared representations between the audio and EMG modalities. In contrast to their work, our approach relies solely on the surface EMG modality (E_S), without leveraging additional data sources such as synchronized audio signals (E_A) or large-scale audio-only corpora (A). This single-modality design enables us to model EMG-to-language mappings without requiring access to multimodal datasets or external speech resources.

Another notable approach is presented by Gowda et al. (2024), who demonstrate that, unlike images and audio - which are functions sampled on Euclidean grids - EMG signals are defined by a set of orthogonal axes, with the manifold of SPD matrices as their natural embedding space. We build upon the methods described by Gowda et al. (2024) in our analysis and introduce the following key improvements: ① we train a recurrent model for EMG-to-phoneme sequence-to-sequence generation, as opposed to the classification models proposed by Gowda et al. (2024), ② we operate in the sparse graph spectral domain, effectively circumventing bottlenecks associated with repeated eigenvalue computation in neural networks, which, due to their iterative nature, often have limited parallelization capabilities on GPUs, and ③ demonstrate EMG-to-language conversion on continuously articulated speech as opposed to individual words or phonemes.

A substantial body of prior work (Jou et al., 2006; Kapur et al., 2020; Meltzner et al., 2018; Toth et al., 2009; Janke and Diener, 2017; Diener et al., 2018) has laid the groundwork for the development of silent speech interfaces. While these studies have been instrumental in shaping the field, they place less emphasis on understanding the *data structure* and the implementation of parameter and data-efficient approaches.

In the following sections, ① we explain the inherent non-Euclidean data structure of EMG signals, ② quantify the signal distribution shift across individuals, and ③ demonstrate that high fidelity phoneme-by-phoneme translation of EMG-to-language is possible using only E_S without E_A and A.

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-AY-D-IY>.

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 personalized audio A using few-shot learning (Choi et al., 2021), which requires as little as a single audio clip from the individual (an audio clip of about 2-5 minutes, not necessarily containing the same phonemic content as L, recorded before their clinical condition). By leveraging this sample, we generate audio A that captures both the predicted linguistic content and the speaker's unique vocal characteristics (we elaborate on this topic in section 8.2).

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 $e_{12} = e_{21} = \mathbf{f}_{v_1}^T \mathbf{f}_{v_2}$, 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) (we provide background on Riemannian geometry of SPD matrices in appendix A).

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. Gowda et al. (2024) show that such a matrix Q can be learned using neural networks constrained on the Stiefel manifold (Huang and Van Gool, 2017) and that such a Q is different for different individuals. However, neural networks constrained on the Stiefel manifold require performing repeated eigendecomposition operations, which have limited parallelization capability and lead to unstable gradients when using CTC loss. Therefore, we simply derive Q from the Fréchet mean $\mathcal{F}$ and use that Q to obtain sparse matrices σ. 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.

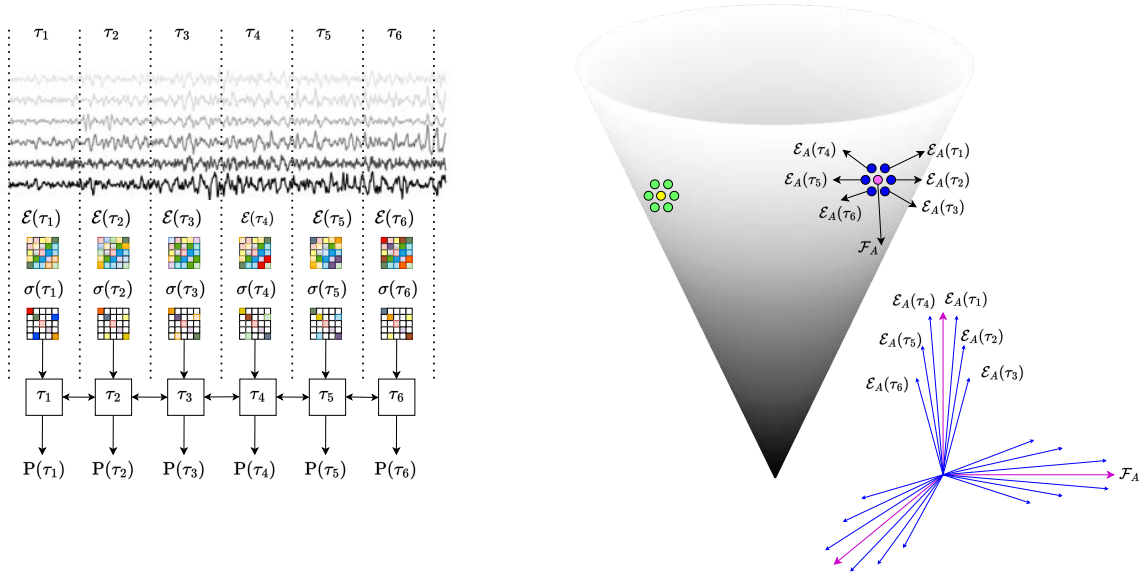


Figure 1: LEFT: EMG-to-phoneme translation pipeline. Bandpass-filtered and z-normalized EMG signals are converted into SPD edge matrices $\mathcal{E}(\tau)$, which are approximately diagonalized to $\sigma(\tau)$ and passed through a BiGRU. The model outputs phoneme probabilities $P(\tau)$ every 20 ms. 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.

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.

To investigate whether recurrent models defined on the manifold provide better representations of $\sigma(\tau)$ for sequence-to-sequence translation compared to those defined in Euclidean space, we draw motivation from prior works (Chakraborty et al., 2018; Jeong et al., 2024), which show that recurrent networks defined on manifolds outperform their Euclidean counterparts on classification and forecasting tasks involving manifold-valued data. Based on this, we construct three distinct GRU architectures:

① **GRU** $_{\text{EUCLIDEAN}}$: A GRU layer defined in the Euclidean domain, following the implementation described by Chung et al. (2014),

② **GRU** $_{\text{MANIFOLD}}$: A GRU layer formulated on the manifold of SPD matrices, as proposed by Jeong et al. (2024), and

③ **GRU** $_{\text{MANIFOLD + ODE}}$: A GRU layer defined on the manifold of SPD matrices, plus an implicit layer solved using neural ordinary differential equations, integrating methodologies from Jeong et al. (2024), Chen et al. (2018), and Lou et al. (2020).

$\text{GRU}_{\text{MANIFOLD}}$ and $\text{GRU}_{\text{MANIFOLD + ODE}}$ directly accept SPD matrices, σ, as input, whereas $\text{GRU}_{\text{EUCLIDEAN}}$ processes vectorized representations of σ. At each time step, the GRU models output probability distributions over 40 phonemes in the English language. The models are 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. We provide further details on the GRU architectures in appendix A.

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 sen-

sor 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 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, which are better modeled as time-varying filters applied to time-varying sources (Sivakumar et al., 2024). Importantly, the choice of eigenbasis Q is subject-specific. EMG signals from different individuals yield different underlying transformations $\mathcal{E}(\tau)$ and, consequently, different eigenspaces due to anatomical and physiological variability—including differences in subcutaneous fat, muscle fiber composition, conduction velocities, and neural drive properties. As a result, signal distribution shifts across individuals can be interpreted as *changes of basis* in the underlying space $\mathbb{R}^{|\mathcal{V}|}$.

4 Data

We evaluate our models using three datasets. They are DATA SMALL-VOCAB, DATA LARGE-VOCAB, and DATA NATO-WORDS. The duration of DATA SMALL-VOCAB is approximately 75 minutes, DATA LARGE-VOCAB is approximately 480 minutes (about 8 hours), and DATA NATO-WORDS is approximately 60 minutes (60 $\times$ 4 from four different individuals). The size of the data is comparable to that used in Willett et al. (2023); Metzger et al. (2023) in terms of the number of articulated sentences.

We begin with DATA SMALL-VOCAB, a timestamped dataset of isolated and connected words, to demonstrate that EMG-to-phoneme sequence mapping is feasible using only E_S, without relying on E_A or external audio corpora (A). This controlled setting serves as a proof of concept before extending to more complex and naturalistic speech.

Next, we use DATA LARGE-VOCAB to evaluate our models on silently articulated speech in unconstrained, conversational settings using a large, general-domain English corpus. This dataset reflects realistic usage scenarios and challenges in large-vocabulary decoding.

Finally, DATA NATO-WORDS is used to demonstrate that a generalizable language-to-spelling model can be trained with minimal data by using a compact set of codewords, such as the NATO phonetic alphabet. We train the same model architecture separately for different individuals and find that performance is consistent across subjects. This indicates that our proposed architecture is effective across users.

We describe each dataset in more detail below. Additional information on data collection and experimental setup is provided in appendix B.

4.1 DATA SMALL-VOCAB

Following Gaddy and Klein (2020), we create a limited-vocabulary dataset consisting of 67 unique words. These words include weekdays, ordinal dates, months, and years. Sentences are constructed from these words in the format <WEEKDAY-MONTH-DATE-YEAR>. A single individual silently articulated 500 such sentences, and the resulting EMG data, denoted as E_S, is translated into output phoneme sequences. We also have timestamps that mark the beginning and end of each word within a sentence.

We collect EMG data from 31 sites at a sampling rate of 5000 Hz. For details about electrode placement and the experimental setup, please refer

to appendix B.

4.2 DATA LARGE-VOCAB

We adapt the language corpora from Willett et al. (2023), who demonstrated a speech brain-computer interface by translating neural spikes from the motor cortex into speech. The dataset comprises an extensive English language corpus containing approximately 6,500 unique words and 11,000 sentences. Unlike Gaddy and Klein (2020, 2021), we collect only E_S (excluding E_A and A) and perform E_S-to-language translation without time-aligning with E_A and A. The data collection setup follows the methodology described for DATA SMALL-VOCAB. This corpus includes sentences of varying lengths, with the subject articulating sentences at a normal speed, averaging 160 words per minute. Timestamps were used solely to mark the beginning and end of each sentence, with the subject clicking the computer mouse at the start of articulation and again upon completion (unlike DATA SMALL-VOCAB, there are no timestamps to demarcate between words within a sentence). For details about electrode placement and the experimental setup, please refer to appendix B.

4.3 DATA NATO-WORDS

We use the dataset provided by Gowda et al. (2024)[1]. Specifically, we use data from their second experiment, in which 4 individuals articulated English sentences in a spelled-out manner using NATO phonemic codes in a silent manner. For instance, 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 phonemically balanced RAINBOW and GRANDFATHER passages in this spelled-out format. In total, 1968 NATO code articulations were recorded across both passages. The EMG data was collected from 22 sites in the neck and cheek regions at a sampling rate of 5000 Hz.

5 Results

Here, we describe the experimental setup and results for DATA LARGE-VOCAB, DATA SMALL-VOCAB, and DATA NATO-WORDS. During preprocessing, raw EMG signals are bandpass filtered between 80 and 1000 Hz and z-normalized per channel along the time dimension. A complete time-dependent graph, $\mathcal{E}(\tau)$, and its diagonalized form, $\sigma(\tau)$, are then constructed from the EMG signals.

5.1 Results for DATA LARGE-VOCAB

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 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 using dictionary lookup, scored with a 3-gram language model following Heafield (2011) (we elaborate on this in section 8.1).

Table 1: Mean PER and WER on DATA LARGE-VOCAB. Lower values indicate better performance. Gray inset values correspond to models trained on raw SPD matrices without approximate diagonalization.

MODEL	PER	WER
6.4 million parameters	**0.478** (0.51)	**0.80** (0.82)

In figure 2, we show the phoneme error rate (PER) and the corresponding word error rate (WER) for the decoded sentences in test set. A lower PER can still result in a high WER, depending on the nature of the transcription errors. For example, the phrase <BELIEVE EVERYTHING> with phonemic transcription <B-IH-L-IY-V-SPACE-EH-V-R-IY-TH-IH-NG>, when decoded by the model as <B-IH-L-IY-SPACE-V-EH-M-R-IY-SPACE-TH-IH-NG-K>, results in the sentence <REALLY VERY THINK>. Although the PER is only 0.38, the WER is 1.5.

In figure 3, we examine how model size affects the PER. To do this, we vary the number of GRU

[1]The dataset is available at Gowda et al. (2024) dataset.

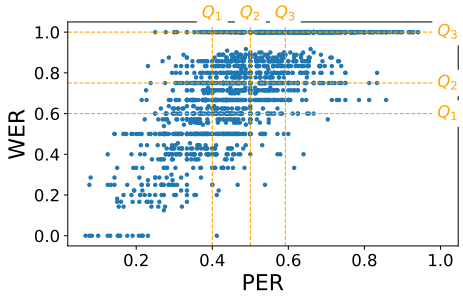


Figure 2: Scatter plot of PER versus WER with quartile annotations for DATA LARGE-VOCAB. A lower PER can still result in a high WER.

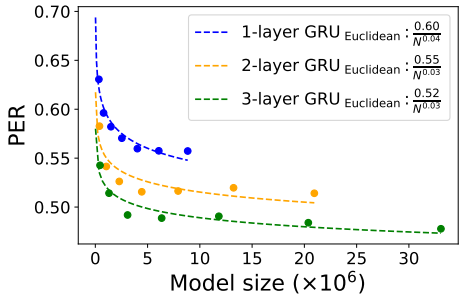


Figure 3: Model size versus PER for EMG-to-phoneme translation for DATA LARGE-VOCAB.

layers and, within each configuration, the dimensionality of the GRU's hidden units. We observe that the relationship between PER and model size approximately follows the power law, as described by Hoffmann et al. (2022) in the context of LLMs. Specifically, when trained with sufficient data, the error E can be expressed as $E = \frac{\alpha}{N^{\beta}}$, where N is the model size and α and β are positive constants. This relationship allows us to predict model performance based on its size. Notably, even a single-layer model achieves a reasonably low PER of 0.56, although deeper models perform better.

5.2 Results for DATA SMALL-VOCAB

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.

We also provide a comparative analysis of GRU EUCLIDEAN, GRU MANIFOLD, and GRU MANIFOLD + ODE. Training GRU EUCLIDEAN takes approximately 2 minutes, GRU MANIFOLD approximately 40 minutes, and GRU MANIFOLD + ODE approximately 80 minutes on an RTX 4090 GPU.

In table 2, 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 2: Mean PER and WER for DATA SMALL-VOCAB. Lower PER and WER are better.

MODEL SIZE	PER	WER
4 MILLION	0.13	0.14

In figure 4, 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. For GRU EUCLIDEAN and GRU MANIFOLD, we observe that the relationship between PER and model size approximately follows a power-law trend. In contrast, GRU MANIFOLD-ODE deviates from this pattern: while it performs competitively at smaller model sizes, its performance deteriorates as the model size increases. The cause of this degradation remains unclear.

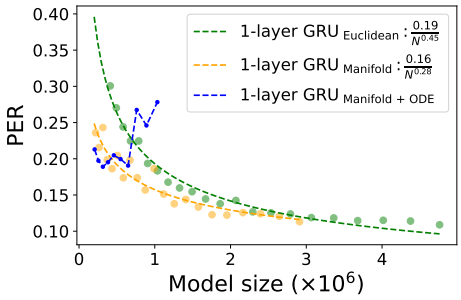


Figure 4: Model size versus PER for EMG-to-phoneme translation for DATA SMALL-VOCAB.

Importantly, both GRU MANIFOLD and GRU MANIFOLD-ODE incur higher computational and training costs, yet fail to yield significant performance improvements at larger scales. As a result, the practical utility of neural ordinary

differential equation-based architectures for sequence-to-sequence model is limited.

The primary motivation for this analysis was to investigate whether performance improvements observed in prior work on manifold-based models for classification tasks (Chakraborty et al., 2018; Jeong et al., 2024) extend to more complex sequence-to-sequence modeling. Our empirical findings suggest that this is not the case.

5.3 Results for DATA NATO-WORDS

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 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 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 3, 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 3: Mean CER for DATA NATO-WORDS. Lower CER indicates better performance. The chance error rate is 0.96, and our error rates are significantly lower than this baseline.

SUBJECT	CER
1	0.557
2	0.550
3	0.704
4	0.564

In figure 5, 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 4 across all subjects.

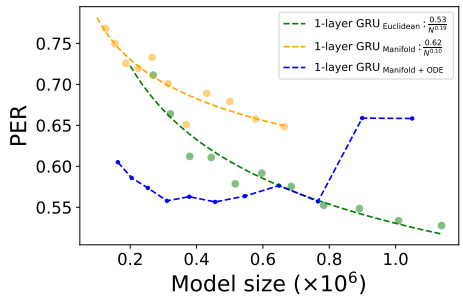
5.4 Comparison with prior work

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-speech decoding and EMG2QWERTY differ, the underlying machine learning principles remain similar.

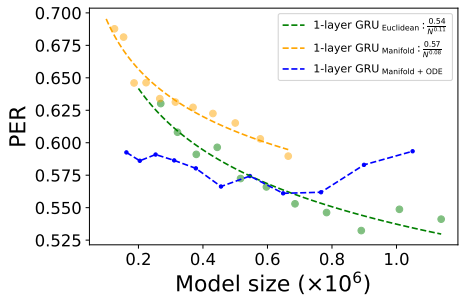
To enable a fair comparison, we conduct controlled experiments in which we replace the original log-spectrogram features from Sivakumar et al. (2024) with SPD covariance matrices. We perform two variants: one using SPD matrices directly, and another using their approximately diagonalized versions. Apart 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 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.

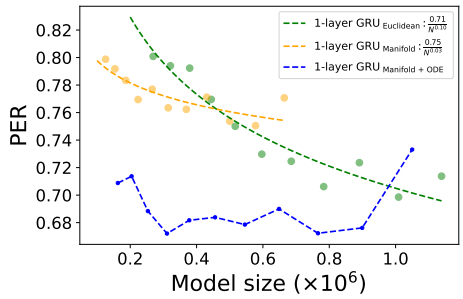
The results are presented in table 4. As shown, our proposed methods outperform the baseline approaches reported by Sivakumar et al. (2024). Furthermore, the use of approximately diagonal-



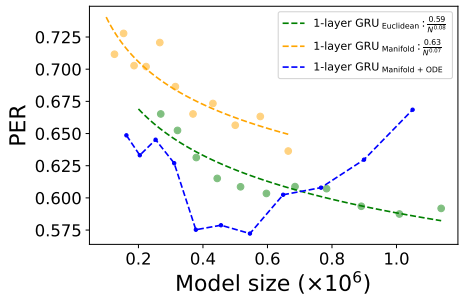
(a) Subject 1



(b) Subject 2



(c) Subject 3

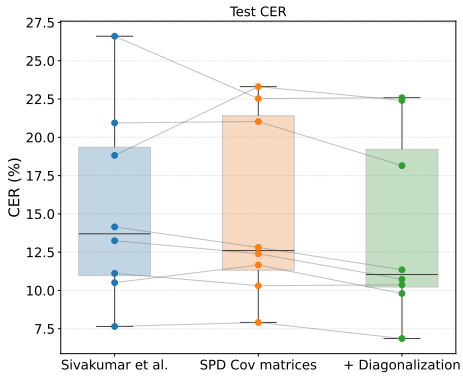


(d) Subject 4

Figure 5: Model size versus PER for EMG-to-phoneme translation for DATA NATO-WORDS.

Table 4: Comparison between our proposed methods and those presented by Sivakumar et al. (2024), with all results averaged over 8 subjects. Model size and FLOPs are identical across all three models. Lower CER is better.

	NO LM		6-GRAM CHAR-LM	
	VAL CER (% ↓)	TEST CER (% ↓)	VAL CER (% ↓)	TEST CER (% ↓)
SIVAKUMAR ET AL. (2024)	15.65 ± 5.95	15.38 ± 5.88	11.03 ± 4.45	9.55 ± 5.16
SPD COV MATRICES	15.66 ± 5.70	15.25 ± 5.66	10.48 ± 4.38	8.71 ± 4.51
+ DIAGONALIZATION	$\mathbf{14.33 \pm 5.27}$	$\mathbf{14.03 \pm 5.27}$	$\mathbf{9.61 \pm 3.84}$	$\mathbf{7.95 \pm 4.54}$



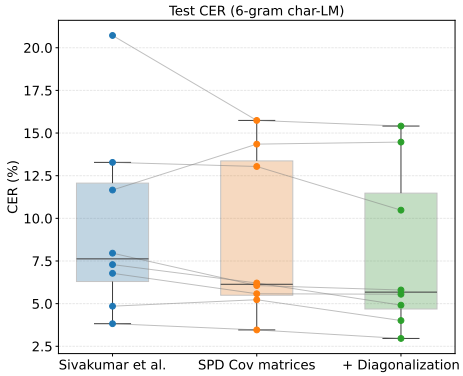


Figure 6: Each dot represents an individual test subject, with connecting lines indicating within-subject performance across different models. The boxplots summarize the median and interquartile range of the results. Our SPD COV MATRICES + DIAGONALIZATION method improves performance for all subjects except USER6.

ized SPD matrices leads to improved performance compared to raw SPD matrices. These findings support the effectiveness of our approach, which is specifically designed to reflect the underlying biological structure of EMG signals.

In figure 6, 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% improvement in CER. With a language model, the improvement increases to 16.8%.

6 Discussion and conclusion

We present a non-invasive, EMG-based speech neuroprosthesis designed for individuals with clinical conditions that impair voicing and articulator movement. Unlike previous approaches that rely on mapping EMG features to time-aligned acoustic signals, our method bypasses audio supervision entirely. Instead, we treat EMG-to-phoneme conversion as a standalone sequence-to-sequence problem, modeling the EMG modality directly without requiring alignment or regression to the audio domain.

Our approach is lightweight, computationally efficient, and grounded in principled design. By releasing our dataset and codebase as open-source resources, we aim to establish a reproducible and extensible foundation for future research in neuromuscular speech interfaces. The effectiveness of our method is demonstrated on large-vocabulary speech corpora, where we show that direct EMG-to-language modeling is feasible.

This contrasts with other non-invasive neural decoding approaches, such as EEG- and MEG-based systems, which—even when constrained to closed-set classification tasks over small vocabularies and evaluated under favorable conditions, such as when alignment between neural signals and audio is known—yield high error rates. For example, decoding listened speech from EEG results in error rates around 95%, while MEG-based methods report error rates near 59% (Défossez et al., 2023). In comparison, our approach performs phoneme-level decoding on large-vocabulary corpora with error rates below 50%, underscoring the potential of EMG as a more accurate and scalable non-invasive alternative.

It should also be noted that invasive BCI and our non-invasive EMG-based method serve different use cases. Invasive systems, such as those reported by Willett et al. (2023) (PER = 21%) and Metzger et al. (2023) (PER = 46%), offer higher accuracy but rely on more than 250 implanted electrodes and operate at slower speaking rates (62–78 words per minute). These systems are especially valuable for individuals with severe motor impairments, such as anarthria or ALS, where residual muscle movement may be absent. In contrast, our approach is well-suited for individuals who retain articulatory muscle control but are unable to produce voiced speech. Additionally, EMG-based systems may be advantageous in contexts such as virtual or augmented reality, where silent communication is desirable.

7 Limitation

While our results are promising, this work has a few limitations. Currently, sentence processing is performed only after the entire utterance has been completed, rather than in an online, real-time fashion. This introduces latency and limits the system's suitability for naturalistic, real-time conversational interaction.

8 Supplementary Technical Details

Here, we provide further information regarding the decoding paradigms employed in our experiments, and our approach to personalized audio synthesis.

8.1 Decoding paradigms

We decode CTC output probabilities using beam search without incorporating external language models or prior linguistic constraints. At each timestep, the algorithm retains the top hypotheses based solely on CTC symbol probabilities, maintaining a fixed beam width of 5. The final output is the sequence with the highest cumulative probability under the CTC model.

To convert phoneme sequences into words, we use a decoding method that combines hard segmentation, lexicon-based dictionary lookup, and language model scoring. The phoneme sequence is first segmented into candidate word units using a designated blank token, with each segment retained if it exceeds a minimum length. Each segment is then matched against a lexicon using normalized Levenshtein distance, and the top-k closest candidates are considered. Each candidate word is scored using:

$$\text{score} = \lambda_{\text{LM}} \cdot \text{LM}(\text{sentence}) - \lambda_{\text{Dist}} \cdot \text{normalized Levenshtein distance},$$

where LM(sentence) is the log-probability of the decoded word sequence under a trained n-gram language model, and the normalized distance is the Levenshtein distance between the phoneme segment and the candidate word's phoneme representation, divided by the maximum of the two lengths. The weights $\lambda_{\text{LM}}, \lambda_{\text{Dist}} \in [0, 1]$ control the tradeoff between language model guidance and acoustic similarity, and are constrained such that $\lambda_{\text{LM}} + \lambda_{\text{Dist}} = 1$. The candidate with the highest score is selected for each segment. This approach enables decoding of noisy phoneme sequences into linguistically plausible word sequences.

We create a 3-gram language using LibriSpeech TRAIN-CLEAN-100 (Panayotov et al., 2015) transcriptions. The lexicon consists of all unique words from the LibriSpeech TRAIN-CLEAN-100 (which is about 35000 words).

We also experimented with jointly training a model using both CTC and attention losses, following the approach described in Hori et al. (2017), and incorporated probabilities from both the CTC and attention-based decoders during beam search. However, this joint model underperformed compared to the CTC-only decoding approach.

8.2 TEXT TO PERSONALIZED AUDIO SYNTHESIS

We synthesize constructed phoneme sequences into personalized audio using methods described by Choi et al. For this, we train the model proposed by Choi et al. on speech corpora provided by Panayotov et al. (LibriSpeech TRAIN-CLEAN-360 and TRAIN-CLEAN-100) and Veaux et al. (VCTK corpus). For few-shot learning, we use a 2-minute reference audio clip from the subject to capture the speaker's vocal characteristics.

The process involves converting the predicted text into audio using Google Text-to-Speech (gTTS). The gTTS-generated audio is then personalized using the model by Choi et al., leveraging the 2-minute reference audio data. This approach ensures that the synthesized audio closely mimics the speaker's unique vocal attributes.

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 in DATA $_{\text{LARGE-VOCAB}}$
T-START <IT WAS EIGHT 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 in DATA $_{\text{LARGE-VOCAB}}$
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

Ethical statement

Research was conducted in accordance with the principles embodied in the Declaration of Helsinki and in accordance with the University of California, Davis Institutional Review Board Administration protocol 2078695-1. All participants provided written informed consent. Consent was also given for publication of the deidentified data by all participants. Participants were healthy volunteers and were selected from any gender and all ethnic and racial groups. Subjects were aged 18 or above, were able to fully understand spoken and written English, and were capable of following task instructions. Subjects had no skin conditions or wounds where electrodes were placed. Subjects were excluded if they had uncorrected vision problems or neuromotor disorders that prevented them from articulating speech. Children, adults who were unable to consent, and prisoners were not included in the experiments.

Acknowledgments

This work was supported by awards to Lee M. Miller from: Accenture, through the Accenture Labs Digital Experiences group; CITRIS and the Banatao Institute at the University of California; the University of California Davis School of Medicine (Cultivating Team Science Award); the University of California Davis Academic Senate; a UC Davis Science Translation and Innovative Research (STAIR) Grant; and the Child Family Fund for the Center for Mind and Brain.

Harshavardhana T. Gowda is supported by Neuralstorm Fellowship, NSF NRT Award No. 2152260 and Ellis Fund administered by the University of California, Davis.

We appreciate Sergey D. Stavisky for reviewing the manuscript and providing insightful feedback.

Conflict of interest

H. T. Gowda and L. M. Miller are inventors on intellectual property related to silent speech owned by the Regents of University of California, not presently licensed.

Author contributions

- Harshavardhana T. Gowda: Conceptualization, Mathematical formulation, concepts development, data analysis, experiment design, data collection software design, data collection, manuscript preparation.
- Ferdous Rahimi: Data collection.
- Lee M. Miller: Conceptualization and manuscript preparation.

References

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. 2007. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29(1):328–347.

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*.

R Chakraborty, CH Yang, X Zhen, M Banerjee, D Archer, D Vaillancourt, V Singh, and BC Vemuri. 2018. A statistical recurrent model on the manifold of symmetric positive definite matrices. *Advances in neural information processing systems*.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. 2018. Neural ordinary differential equations. *Advances in neural information processing systems*, 31.

Hyeong-Seok Choi, Juheon Lee, Wansoo Kim, Jie Lee, Hoon Heo, and Kyogu Lee. 2021. Neural analysis and synthesis: Reconstructing speech from self-supervised representations. *Advances in Neural Information Processing Systems*, 34:16251–16265.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

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.

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. *arXiv preprint arXiv:2411.02591*.

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.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 30016–30030.

Takaaki Hori, Shinji Watanabe, and John Hershey. 2017. Joint CTC/attention decoding for end-to-end speech recognition. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 518–529, Vancouver, Canada. Association for Computational Linguistics.

Zhiwu Huang and Luc Van Gool. 2017. A riemannian network for spd matrix learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

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.

Seungwoo Jeong, Wonjun Ko, Ahmad Wisnu Mulyadi, and Heung-Il Suk. 2024. Deep Efficient Continuous Manifold Learning for Time Series Modeling . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 46(01):171–184.

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*.

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.

Aaron Lou, Derek Lim, Isay Katsman, Leo Huang, Qingxuan Jiang, Ser Nam Lim, and Christopher M De Sa. 2020. Neural manifold ordinary differential equations. *Advances in Neural Information Processing Systems*, 33:17548–17558.

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, et al. 2023. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620(7976):1037–1046.

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.

Tanja Schultz and Michael Wand. 2010. Modeling coarticulation in emg-based continuous speech recognition. *Speech Communication*, 52(4):341–353.

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-eight 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.

Christophe Veaux, Junichi Yamagishi, and Simon King. 2013. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 1–4.

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, et al. 2023. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036.

A Background on Riemannian geometry of SPD matrices

Speech articulation involves the coordinated activation of various muscles, with their activation patterns defined by the functional connectivity of the underlying neuromuscular system. Consequently, EMG signals collected from multiple, spatially separated muscle locations exhibit a time-varying graph structure. Gowda et al. (2024) demonstrate that the graph edge matrices corresponding to orofacial movements underlying speech articulation are inherently distinguishable on the manifold of SPD matrices. Through experiments with 16 subjects, they highlight the effectiveness of using SPD manifolds as an embedding space for these edge matrices. Building on this foundation, we investigate the temporal evolution of graph connectivity using edge matrices to enable EMG-to-language translation.

Directly working with SPD matrices using affine-invariant or log-Euclidean metrics (Arsigny et al., 2007) involves computationally expensive operations, such as matrix exponential and matrix logarithm calculations. These operations make mappings between the manifold space and the tangent space, and vice versa, computationally intensive. To address this, Lin (2019) proposed methods to operate on SPD matrices using Cholesky decomposition. They established a diffeomorphism between the Riemannian manifold of SPD matrices and Cholesky space, which was later uti-

lized by Jeong et al. (2024) to develop computationally efficient recurrent neural networks. In Cholesky space, the computational burden is significantly reduced: logarithmic and exponential computations are restricted to the diagonal elements of the matrix, making them element-wise operations. Additionally, the Fréchet mean is derived in a closed form.

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

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

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

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

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 $\mathcal{L}(\tau)$.

GRU $_{\text{EUCLIDEAN}}$ is a standard GRU (Chung et al., 2014). GRU $_{\text{MANIFOLD}}$ is constructed from GRU $_{\text{EUCLIDEAN}}$ by replacing the arithmetic operations of GRU$_{\text{EUCLIDEAN}}$ defined in the Euclidean domain with the corresponding operations on the SPD manifold. Gates of GRU $_{\text{MANIFOLD}}$ as defined by Jeong et al. (2024) are given below. Given the sparse SPD edge matrices $\sigma(\tau)$ over different time windows τ, we first calculate their corresponding Cholesky decompositions $l(\tau) = \text{CHOLESKY}(\sigma(\tau))$, such that $\sigma(\tau) = l(\tau)l(\tau)^T$.

Update-gate z_τ at time-step τ is

$$z_\tau = \text{SIGMOID}(w_z\lfloor l_\tau\rfloor + u_z\lfloor h_{\tau-1}\rfloor + b_z) + \text{SIGMOID}(b_{z'}[\exp(w_{z'}\log(\mathbb{D}(l_\tau)) + u_{z'}\log(\mathbb{D}(h_{\tau-1}))]), \quad (1)$$

where w_z, u_z, b_z, $w_{z'}$, and $u_{z'}$ are real weights and $b_{z'}$ is a real positive weight.

Reset-gate r_τ at time-step τ is

$$r_\tau = \text{SIGMOID}(w_r\lfloor l_\tau\rfloor + u_r\lfloor h_{\tau-1}\rfloor + b_r) + \text{SIGMOID}(b_{r'}[\exp(w_{r'}\log(\mathbb{D}(l_\tau)) + u_{r'}\log(\mathbb{D}(h_{\tau-1}))]), \quad (2)$$

where w_r, u_r, b_r, $w_{r'}$, and $u_{r'}$ are real weights and $b_{r'}$ is a real positive weight.

Candidate-activation vector $\hat{h}_\tau$ is

$$\hat{h}_\tau = \text{TANH}(w_h\lfloor l_\tau\rfloor + u_h(\lfloor r_\tau\rfloor * \lfloor h_{\tau-1}\rfloor) + b_h) + \text{SOFTPLUS}(b_{h'}\exp(w_{h'}\log(\mathbb{D}(l_\tau)) + u_{h'}\log(\mathbb{D}(r_\tau) * \mathbb{D}(h_{\tau-1})))), \quad (3)$$

where w_h, u_h, b_h, $w_{h'}$, and $u_{h'}$ are real weights and $b_{h'}$ is a real positive weight.

Output vector h_τ is

$$h_\tau = (1 - \lfloor z_\tau\rfloor) * \lfloor h_{\tau-1}\rfloor + \lfloor z_\tau\rfloor * \lfloor\hat{h}_\tau\rfloor + \exp((1 - \mathbb{D}(z_\tau)) * \log(\mathbb{D}(h_{\tau-1})) + \mathbb{D}(z_\tau) * \log(\mathbb{D}(\hat{h}_\tau))). \quad (4)$$

In the above equations, $h_{\tau-1}$ is the hidden-state at time-step $\tau - 1$.

In GRU $_{\text{MANIFOLD + ODE}}$, we define an additional implicit layer solved using neural ODEs. The dynamics f of the EMG data are modeled by a neural network with parameters Θ. The output state $h\tau$ is updated as

$$h_{\tau-1} \leftarrow \text{ODESOLVE}(f_\Theta, \widetilde{\text{LOG}}(h_{\tau-1}), (\tau-1, \tau))$$
$$h_\tau = \text{GRU}(l_\tau, \widetilde{\text{EXP}}(h_{\tau-1})), \quad (5)$$

where $\widetilde{\text{LOG}}$ is the logarithmic mapping from the manifold space of SPD matrices to its tangent space, and $\widetilde{\text{EXP}}$ is its inverse operation, as defined by Lin and given below. GRU denotes a gated recurrent unit whose gates are defined by equations 1–4.

For a matrix $\mathcal{X}$ in the tangent space, the exponential map to the manifold space at $\mathcal{L}$ is given by

$$\widetilde{\text{EXP}}_{\mathcal{L}}(\mathcal{X}) = \lfloor\mathcal{L}\rfloor + \lfloor\mathcal{X}\rfloor + \mathbb{D}(\mathcal{L})\exp\left(\mathbb{D}(\mathcal{X})\mathbb{D}(\mathcal{L})^{-1}\right). \quad (6)$$

For a matrix $\mathcal{L}$ in the manifold space, the logarithmic map to the tangent space at $\mathcal{L}$ is given by

$$\widetilde{\text{LOG}}_{\mathcal{L}}(\mathcal{K}) = \lfloor\mathcal{K}\rfloor - \lfloor\mathcal{L}\rfloor + \mathbb{D}(\mathcal{L})\log\left(\mathbb{D}(\mathcal{L})^{-1}\mathbb{D}(\mathcal{K})\right). \quad (7)$$

In the above, $\mathcal{L}$ denotes a point on the manifold.

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 Experimental details

We collect EMG signals from 31 sites on the neck, chin, jaw, cheek, and lips using monopolar electrodes. An ACTICHAMP PLUS amplifier and associated active electrodes from BRAIN VISION (Brain Vision) are used to record EMG signals at 5000 Hertz. To ensure proper contact between the skin surface and electrodes, we use SUPERVISC, a high-viscosity electrolyte gel from EASYCAP (Easycap). We develop a software suite in a PYTHON environment to provide visual cues to subjects and to collate and store timestamped data. For time synchronization, we use lab streaming layer (LSL). See figure 7 for electrode placement. Besides 31 data electrodes, we also have a GROUND electrode (marked as GND) and a REFERENCE electrode (marked as 32). GROUND electrode is placed on the left ear lobe and the REFERENCE electrode is placed on the right ear lobe.

Before signal acquisition, participants were briefed on the experimental protocol and seated comfortably in a chair. For silent speech data (E_S), participants were instructed to articulate naturally but inaudibly. For DATA SMALL-VOCAB, the sentences were presented as individual or grouped words, demarcated by timestamps, and displayed in the following manner:

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

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

with each segment temporally following the previous one. In DATA LARGE-VOCAB, there are no such intra-sentence timestamps; only the start and end of the sentence are timestamped using mouse clicks from the subject. When a subject is ready to articulate a sentence, they click the mouse, prompting the sentence to appear on the screen. Once articulation is complete, they click the mouse again to indicate the end, causing the sentence to disappear from the screen—thus allowing them to articulate at their own pace.

The data collection environment was carefully controlled to eliminate AC electrical interference. EMG signals underwent minimal preprocessing. The signal from the REFERENCE channel (electrode 32) was subtracted from all other EMG data channels. The resulting signals were then bandpass filtered using a third-order Butterworth filter between 80 and 1000 Hz and segmented according to sentence start and end times based on synchronized timestamps. The segmented sentences were subsequently z-normalized along the time dimension for each channel. The preprocessed EMG signals were then used to construct a fully connected sensor graph, $\mathcal{E}(\tau)$, and its approximately diagonalized form, $\sigma(\tau)$.

The electrodes are positioned over regions that directly overlay 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 in the lower cheek region, play a vital role in tongue 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 the **zygomaticus**, involved in upper lip elevation—are associated with electrode regions approximately around nodes 22, 18, 17, and 15 in figure 7. Electrodes located beneath the jaw capture activity from muscles involved in tongue protrusion and jaw–tongue coordination, such as the **genioglossus** (near electrodes 8, 9, 23, and 25) and the **digastric**. Additionally, electrodes near the laryngeal region (nodes 6, 7, 10, 11, 26, and 27) reflect the activity of muscles that modulate laryngeal and hyoid position—such as the **sternohyoid**,

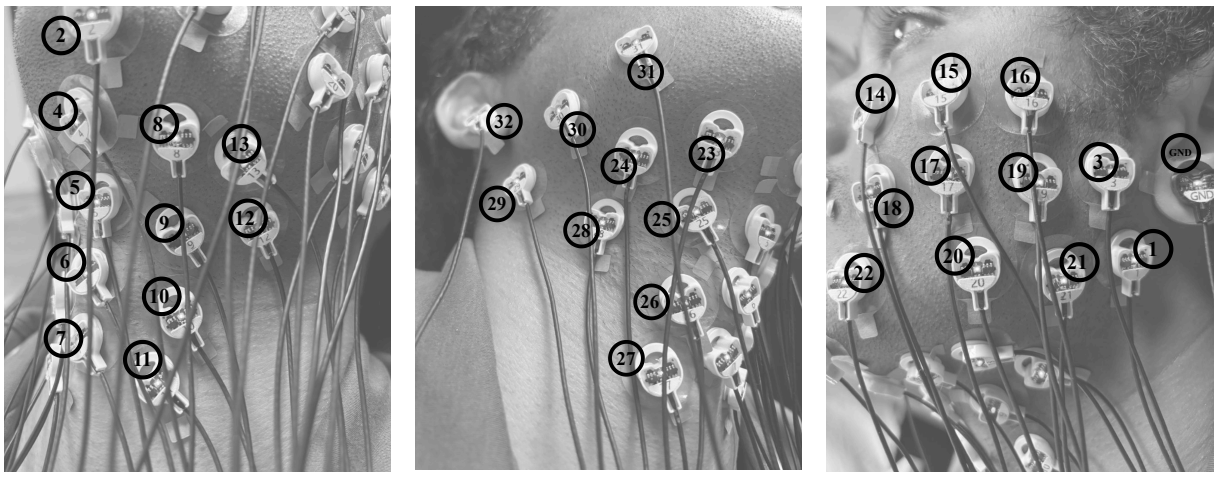


Figure 7: LEFT: Electrode placement on the left side of the neck. MIDDLE: Electrode placement on the right side of the neck. RIGHT: Electrode placement on the left cheek.

stylohyoid, and **digastric**—which are instrumental in pitch control, vowel shaping, and jaw movement.