Non-invasive electromyographic speech neuroprosthesis: a geometric perspective

Harshavardhana T. Gowda*
University of California, Davis
Davis, CA 95616
tgharshavardhana@gmail.com

Lee M. Miller
University of California, Davis
Davis, CA 95616
leemiller@ucdavis.edu

Abstract

We present a high-bandwidth, egocentric 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 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. All data, code, and model checkpoints are open-sourced at GitHub + Data repository.

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 [1] and [2] 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. Work in [3] 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).

*Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Learning from Time Series for Health.

Unlike invasive methods [1, 2], 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 contrast to prior approaches [3, 4, 5], 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 [6], 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 [4, 5]. 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 to 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.

Another notable approach is presented in [7], which demonstrates 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 in [7] 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 in [7], 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 [8, 9, 10, 11, 12, 13] 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* or implementing parameter- and data-efficient approaches. Moreover, we open-source eight hours of EMG data collected during *silent* speech articulation.

3 Methods

EMG signals are collected by a set of sensors $\mathcal{V}$ and represented as functions of time t. A sequence of EMG signals E_S corresponding to silently articulated speech, associated with an audio signal 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 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 example, 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 enables training with *unaligned* pairs of E_S and L, eliminating the need for precise alignment between input signals and their corresponding phoneme sequences. During testing, a sample E_S not seen during training produces probabilities over all possible phonemes (40 in our case) at each time step, and we construct L using beam search.

We represent EMG signals by constructing a complete graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}(\tau))$ that captures their functional connectivity, 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 $e_{12} = e_{21} = \mathbf{f}_{v_1}^T \mathbf{f}_{v_2}$, corresponding to the covariance of the signals at those nodes over 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 in [14]. (We provide background on the 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. Instead of recalculating U for each $\mathcal{E}$ at every time step τ, we fix an approximate common eigenbasis Q derived from the Fréchet mean $\mathcal{F}$ [14] of all adjacency matrices (across 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. 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 graph spectral sequences σ for all time windows τ and use them as inputs for EMG-to-language translation. We illustrate these concepts in figure 1 in appendix A. We implement a gated recurrent unit (GRU) architecture [15] for EMG-to-phoneme sequence-to-sequence modeling. The input to the GRU consists of a sequence of approximately diagonalized matrices σ, derived over different time windows τ.

4 Data and results

We adapt the language corpora from [1], who demonstrated a speech brain–computer interface by translating neural spikes from the motor cortex into speech. The dataset comprises an extensive English corpus containing approximately 6,500 unique words and 11,000 sentences. Unlike [4, 5], we collect only E_S (excluding E_A and A) and perform E_S-to-language translation without time alignment to E_A or A. The corpus includes sentences of varying lengths, with the subject articulating at a normal speaking rate, averaging 160 words per minute. Timestamps were used solely to mark sentence boundaries: the subject clicked a computer mouse at the start of articulation and again at its completion. Details of electrode placement and the experimental setup are provided in appendix B.

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 used as input to a GRU for EMG-to-phoneme sequence translation. We also compare our approach with a baseline method that uses EMG spectrograms similar to [16], where the spectrograms are provided as input to a GRU model (see appendix C.1 for further explanation). The dataset is divided into training, validation, and test sets containing 8000, 1000, and 1970 sentences, respectively. Sentences in the test set are not included in either the training or validation sets.

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**

The model shown in figure 1 (see appendix A) is trained with three GRU layers for 100 epochs, and the checkpoint corresponding to the lowest validation loss is 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. As shown in table 1, our method significantly outperforms the baseline. Words are reconstructed from phoneme sequences using weighted finite-state transducer (WFST)–based decoding, in which the CTC topology (H), lexicon (L), and a 4-gram language model (G) are composed into a single search graph (HLG) following [17]. Further details are provided in appendix C.

4.1 Comparison with prior work

To the best of our knowledge, no prior work has performed 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 [16]. In this dataset, subjects wear EMG wristbands on both hands and type on a QWERTY keyboard. The task is to decode the resulting EMG signals into a sequence of characters using CTC loss. Although the physical actions in EMG-to-speech decoding and EMG2QWERTY differ, the underlying machine learning principles are similar.

To ensure a fair comparison, we conduct controlled experiments in which we replace the original log-spectrogram features from [16] with 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 the 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 [16] in which personalized models are trained independently for each subject, starting from random weight initialization. Other paradigms—such as zero-shot transfer, where a model is trained on 100 subjects and evaluated on 8 unseen individuals, or personalized fine-tuning, where 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 summarized in table 2, with per-subject performance shown in figure 4 (appendix C). As shown, our proposed method outperforms the baseline reported by [16]. Without a language model, we achieve an 8.8% relative improvement in CER over the baseline. With a language model, the relative improvement increases to 16.8%[2].

Table 2: Comparison between our proposed methods and those presented by [16], with all results averaged over 8 subjects. Model size and FLOPs are identical across all three models. Lower CER is better. The CER improvement arising from our method is statistically significant ($p < 0.015$).

	NO LM		6-GRAM CHAR-LM	
	VAL CER (% ↓)	TEST CER (% ↓)	VAL CER (% ↓)	TEST CER (% ↓)
BASELINE (SPECTROGRAM) [16]	15.65 ± 5.95	15.38 ± 5.88	11.03 ± 4.45	9.55 ± 5.16
MATRICES $\sigma(\tau)$ (OURS)	$\mathbf{14.33 \pm 5.27}$	$\mathbf{14.03 \pm 5.27}$	$\mathbf{9.61 \pm 3.84}$	$\mathbf{7.95 \pm 4.54}$

5 Conclusion and discussion

We show that EMG signals collected from orofacial muscles can be efficiently converted into text through phonemic decoding. This demonstrates that a complex task such as speech can be decoded non-invasively from EMG at a fine temporal resolution of 20 ms, where the expected chance phoneme error rate is approximately 98%, compared to a much lower error rate of 49% achieved by our method—without requiring time alignment with audio or regression to the acoustic domain. This stands in contrast to other non-invasive neural decoding approaches, such as EEG- and MEG-based systems, which—even when restricted to closed-set classification tasks over small vocabularies and evaluated under favorable conditions (e.g., when alignment between neural signals and audio is provided)—still yield high error rates. For example, decoding listened speech from high-density EEG results in error rates around 95%, while MEG-based methods report rates near 59% [3]. In comparison, our approach performs phoneme-level decoding on large-vocabulary corpora with error rates around 50%, underscoring the potential of EMG as a more accurate and scalable non-invasive alternative for individuals with clinical conditions that impair voicing and articulator movement.

[2]For reference, in personalized-finetuning paradigm, [16] trained a generic model on 100 subjects (nearly 100× more data) and 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. This 6.95% result represents a strong performance ceiling achieved with large-scale pretraining. In comparison, our 7.95% CER—obtained using only per-subject training (100× less data)—is already close to this ceiling, highlighting the effectiveness of our approach.

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, concept development, data analysis, experiment design, data collection software design, data collection, manuscript preparation.
- Lee M. Miller: Conceptualization and manuscript preparation.

References

[1] 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. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023.

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

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

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

[5] David Gaddy and Dan Klein. 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, 2021.

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

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

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

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

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

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

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

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

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

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

[16] Viswanath Sivakumar, Jeffrey Seely, Alan Du, Sean R Bittner, Adam Berenzweig, Anuoluwapo Bolarinwa, Alexandre Gramfort, and Michael I Mandel. 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*, 2024.

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

[18] 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. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 30016–30030, 2022.

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

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

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

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

[23] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 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, 2015.

[24] Kenneth Heafield. KenLM: Faster and smaller language model queries. In Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar F. Zaidan, editors, *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, July 2011. Association for Computational Linguistics.

A Background on Riemannian geometry of SPD matrices

Speech articulation involves the coordinated activation of multiple muscles, with their activation patterns governed by the functional connectivity of the underlying neuromuscular system. Consequently, EMG signals collected from spatially distributed muscle sites exhibit a time-varying graph structure. [7] demonstrates 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, [7] shows 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. The overall decoding pipeline is illustrated in figure 1.

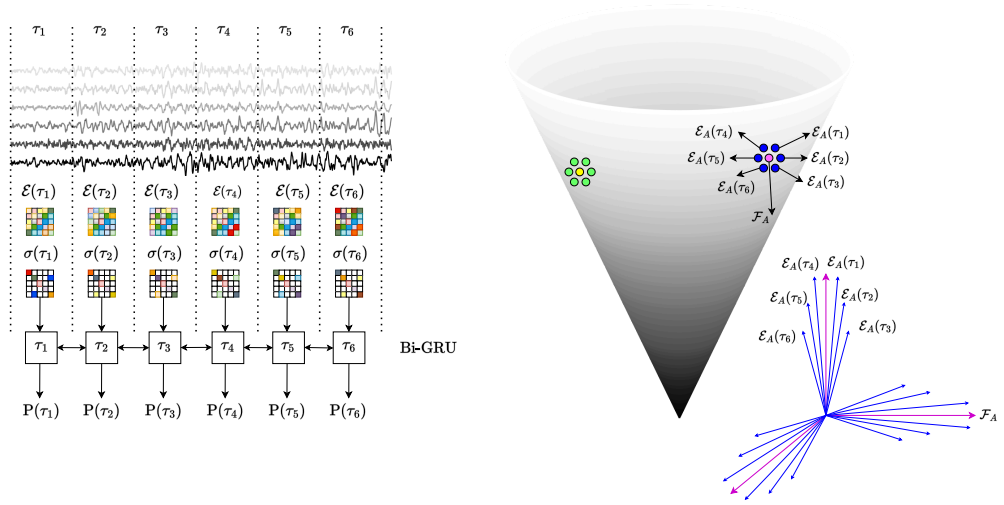

Figure 1: LEFT: EMG-to-phoneme translation pipeline. Bandpass-filtered and z-normalized EMG signals are converted into SPD edge matrices $\mathcal{E}(\tau)$, which are approximately diagonalized to $\sigma(\tau)$ and passed through a BiGRU. The model outputs phoneme probabilities $P(\tau)$ every 20 ms.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

Since SPD matrices ($\mathcal{E}(\tau)$, $\sigma(\tau)$) inherently encode articulatory information, directly feeding them into a neural network that models temporal dependencies (such as a GRU) enables effective decoding of continuously articulated speech at the phonemic level. This aligns with the fact that phonemes are determined by articulatory placement. In contrast, spectrogram features do not exhibit this property and perform significantly worse than SPD-based representations, as shown in table 1. Furthermore, figure 2 illustrates that the relationship between phoneme error rate (PER) and model size approximately follows a power law, consistent with the scaling behavior described by [18] for large language models. 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 $\alpha, \beta > 0$ are constants. This scaling relationship allows us to predict model performance based on size. Notably, even a single-layer model achieves a reasonably low PER of 56%, while deeper models further improve performance. These findings demonstrate that our approach is both effective and theoretically well-grounded.

A.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 $\mathcal{E}(\tau) = \mathcal{L}(\tau)\mathcal{L}(\tau)^T$.

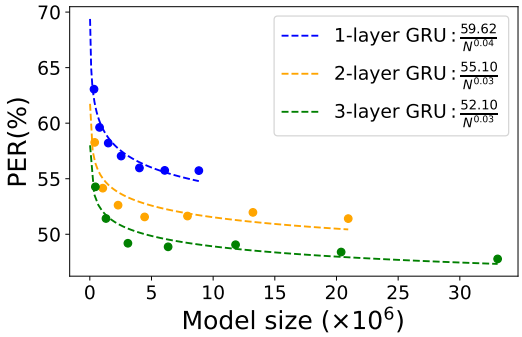

Figure 2: Model size versus PER for EMG-to-phoneme translation.

Then, the Fréchet mean of the Cholesky decomposed matrices $\mathcal{L}(\tau)$ is given by

$$\mathcal{F}_{\text{CHOLESKY}} = \frac{1}{n}\sum_{i=1}^{n}\lfloor\mathcal{L}(\tau_i)\rfloor \quad + \quad \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log(\mathbb{D}(\mathcal{L}(\tau_i)))\right) [14] \quad .$$

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

Previous work in [19] 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, [20, 21] employed Riemannian geometry frameworks for classification tasks in EEG-based brain-computer interfaces, while [22] 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.

A.2 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 modeled using a recurrent neural network to capture temporal dynamics. This formulation aligns with the physiological origin of EMG

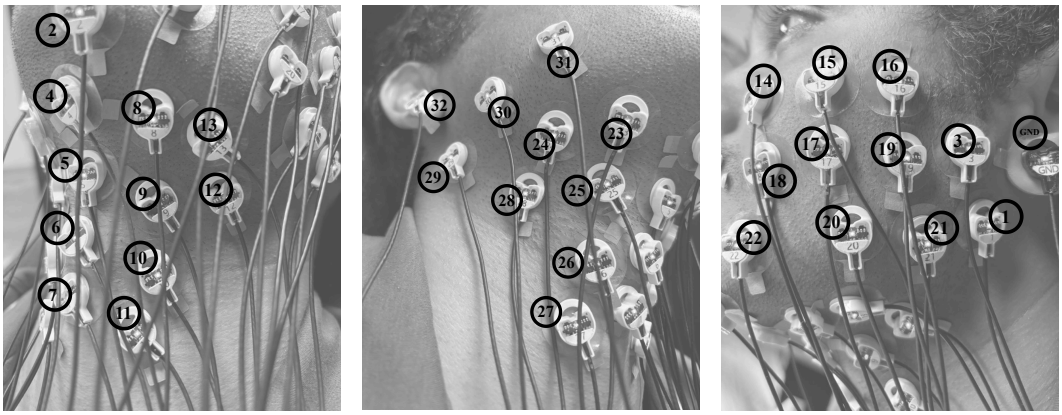

Figure 3: LEFT: Electrode placement on the left side of the neck. MIDDLE: Electrode placement on the right side of the neck. RIGHT: Electrode placement on the left cheek.

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 [16]. 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 $\mathbb{R}^{|\mathcal{V}|}$.

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 3 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 earlobe and the REFERENCE electrode is placed on the right earlobe.

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

C Additional technical details

For calculation of the phoneme error rate (PER) in section 4, we use beamsearch over the CTC output probabilities without incorporating a language model, with a beam width of 50.

For phoneme-to-word decoding in section 4, we use the LibriSpeech-100 transcripts [23], which contain roughly 38000 sentences and 35000 unique words. From these, we construct a lexicon finite-state transducer (FST), L, mapping words to their phonemic transcriptions. A grammar FST, G, is built from a 4-gram language model trained with KenLM [24]. The CTC topology FST, H, encodes allowable symbol sequences under the CTC criterion. These components are then composed into the decoding graph $HLG = H \circ L \circ G$, which integrates the acoustic model constraints (H), lexicon mapping (L), and language model probabilities (G). During inference, we perform beam search over HLG with a beam width of 50, and compute the word error rate (WER) as the normalized Levenshtein distance between the reference and the transcribed sequences.

C.1 Spectrogram representations lack sufficient phonemic structure

Unlike matrices $\sigma(\tau)$, which inherently encode articulatory structure and enable a vanilla GRU to learn meaningful temporal dependencies, raw spectrograms lack such structure. Using spectrograms as GRU inputs caused the model to collapse to a few phoneme sequences regardless of the input, making phoneme-to-word decoding infeasible and resulting in a WER of 1 (table 1). To ensure a fair comparison, we matched the temporal resolution of the spectrograms to that of the SPD features (50 ms window and 20 ms hop). We computed an STFT with $n_{\text{FFT}} = 256$ (129 linear-frequency bins) and then average-pooled the frequency axis down to 31 bins per channel. This produced per-frame tensors of shape $(31 \text{ channels}) \times (31 \text{ frequency bins})$, paralleling the 31×31 shape of $\sigma(\tau)$. Spectrogram inputs did not support robust temporal modeling, whereas $\sigma(\tau)$ enabled the vanilla GRU to learn stable and discriminative dynamics (table 1).

C.2 Individual results on EMG2QWERTY dataset.

In figure 4, we present subject-wise results for the EMG2QWERTY dataset described in subsection 4.1. Our SPD covariance matrices with approximate diagonalization (matrices $\sigma(\tau)$) outperform the baseline spectrogram features for all subjects except USER6.

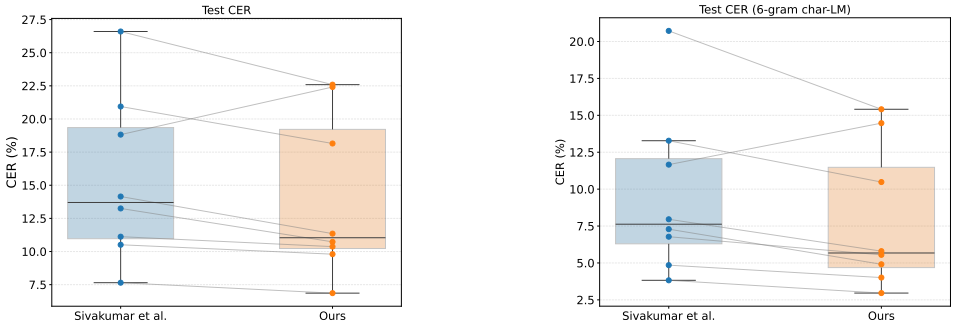

Figure 4: Results for individual subjects in EMG2QWERTY dataset. Each dot represents an individual test subject, with connecting lines indicating within-subject performance across different models. The boxplots summarize the median and interquartile range of the results. Our method improves performance for all subjects except USER6.