

Geometry of orofacial neuromuscular signals: speech articulation decoding using surface electromyography

Harshavardhana T. Gowda

*Department of Electrical and Computer Engineering
University of California, Davis*

TGHARSHAVARDHANA@GMAIL.COM

Zachary D. McNaughton

*Center for Mind and Brain
University of California, Davis*

ZDMCNAUGHTON@UCDAVIS.EDU

Lee M. Miller

*Center for Mind and Brain
Department of Neurobiology, Physiology, and Behavior
Department of Otolaryngology/Head and Neck Surgery
University of California, Davis*

LEEMILLER@UCDAVIS.EDU

Abstract

Each year, millions of individuals lose the ability to speak intelligibly due to causes such as neuromuscular disease, stroke, trauma, and head/neck cancer surgery (e.g. laryngectomy) or treatment (e.g. radiotherapy toxicity to the speech articulators). Effective communication is crucial for daily activities, and losing the ability to speak leads to isolation, depression, anxiety, and a host of detrimental sequelae. Non-invasive surface electromyography (sEMG) has shown promise to restore speech output in these individuals. The goal is to collect sEMG signals from multiple articulatory sites as people silently produce speech and then decode the signals to enable fluent and natural communication. Currently, many fundamental properties of orofacial neuromuscular signals relating to speech articulation remain unanswered. They include questions relating to ① the data structure of the orofacial sEMG signals, ② the signal distribution shift of sEMG across individuals, ③ ability of sEMG signals to span the entire English language phonetic space during silent speech articulations, and ④ the generalization capability of non-invasive sEMG based silent speech interfaces. We address these questions through a series of experiments involving healthy human subjects. We show that sEMG signals evince graph data structure and that the signal distribution shift is given by a

change of basis. Furthermore, we show that silently voiced articulations spanning the entire English language phonetic space can be decoded using small neural networks which can be trained with little data and that such architectures work well across individuals. To ensure transparency and reproducibility, we open-source all the data and codes used in this study.

Keywords: Silent speech, surface electromyogram (sEMG) signals, Riemannian manifolds, symmetric positive definite matrix learning.

Data and code availability

Data is available at [DOI 10.17605/OSF.IO/YM5JD](https://doi.org/10.17605/OSF.IO/YM5JD).
Codes are available at [GitHub](https://github.com).

1. Introduction

Individuals who lose intelligible speech output due to disease or damage of the articulators must learn new ways to communicate. Noninvasive brain-body-computer interfaces might provide the means for people with dysarthria to communicate fluently and with a naturalistic voice as demonstrated by [Gaddy and Klein \(2021\)](#). Among the modalities to capture “silent” speech information, neuromotor interfaces such as sEMG show great promise as such systems can work in many realistic environments, e.g. in noisy backgrounds or with visual occlusion,

where traditional methods based on video (such as generating audio from lip movements) may fail. In addition, neural interfaces encode rich information in multiple sensor nodes at different spatial locations and can detect subtle movements and gestures which may not be discernible with video or residual audio signals.

In this article, we record sEMG signals at multiple muscle locations on the neck, jaw, chin, and cheek for speech articulation decoding. Numerous muscles work in concert to produce a given articulation and these synergistic activation patterns are encoded across the spatially distributed sensor electrodes. We show that this rich spatial information embedded in multiple sensor nodes manifests non-Euclidean data structure that is amenable for analysis on a differentiable manifold of symmetric positive definite matrices equipped with a Riemannian metric. sEMG signals, due to their multivariate nature, display graph data structure that is defined by a set of orthogonal axes rather than functions sampled on 1-dimensional or 2-dimensional Euclidean grids, as is the case for audio signals and images, and the domain shift in sEMG across individuals is characterized by a change of basis.

2. Related work

The current benchmark in sEMG based silent-speech interfaces is given by [Gaddy and Klein \(2020\)](#) and [Gaddy and Klein \(2021\)](#). They synthesize audio from silently articulated speech using sEMG signals by time-aligning speech targets with audibly articulated sEMG signals. However, such an approach is not feasible for practical deployment of silent-speech interfaces due to technicalities including ① unavailability of audible articulations from people who have already lost speech articulators (e.g. people who have undergone laryngectomy), ② need for extensive training data including both audibly and silently articulated speech, and ③ impracticality of obtaining such alignments during actual device deployment.

Another publicly available dataset is given by [Wand et al. \(2014\)](#). However, working codes have not been made available to support a reproducible benchmark. Moreover, the use of hand-crafted sEMG features along with other rigidly defined methods such as hidden Markov models conditioned on ‘bundled phonetic features’ make this method an

unlikely candidate for practical deployment given the complex time-dependent non-linear dynamics of sEMG signals.

Other prior works in silent speech interfaces such as [Meltzner et al. \(2017\)](#), [Janke and Diener \(2017\)](#), [Jou et al. \(2006\)](#), [Diener et al. \(2018\)](#), and [Kapur et al. \(2020\)](#) perform sEMG to text or speech translation using private datasets and unpublished working codes.

Furthermore, none of the above works address several fundamental questions regarding orofacial neuromuscular signals such as: ① What is the appropriate data-representation for multivariate sEMG signals? ② Given that muscle action potentials depend on various characteristics which may differ widely across individuals such as thickness of subcutaneous fat, spatial distribution of muscle fibers, distribution of muscle fiber conduction velocity ([Farina et al. \(2014\)](#)), circumstantial factors such as precise electrode locations, and neural properties such as discharge characteristics of the neural drive ([Farina et al. \(2014\)](#)), what quantifies the signal distribution shift across individuals? ③ And, is it possible to learn meaningful representations of sEMG signals using very little data (instead of tens of hours of data as in [Gaddy and Klein \(2020\)](#) and [Gaddy and Klein \(2021\)](#))?

We answer the above questions via a series of experiments on healthy human subjects. For the first time, we quantify and explain how the combined effect of individual idiosyncrasies, anatomy, and physiology affect sEMG signals. We summarize our original contributions in the next section.

3. Our contribution

To the best of our knowledge, we present the largest open-sourced data (and codes) from a series of articulation experiments from 16 subjects ([Gaddy and Klein \(2020\)](#) provide data from a single subject and [Wand et al. \(2014\)](#) provide data from 4 subjects). Our data is sampled at a higher frequency and from larger number of spatial locations compared to previously open-sourced data; we provide data from 22 sEMG channels sampled at 5000 Hertz ([Gaddy and Klein \(2020\)](#) provide data from 8 channels sampled at 1000 Hertz; [Wand et al. \(2014\)](#) provide data from 6 channels sampled at 600 Hertz).

We demonstrate the following. ① Multivariate sEMG signals evince graph data structure defined by a set of orthogonal axes and that the domain shift in sEMG signals due to combined effect of anatomical, physiological, and neural drive properties is characterized by a change of basis. ①a We show that the orofacial movements underlying speech are naturally distinguishable on the manifold of SPD matrices. That is, different orofacial movements can be classified in an unsupervised manner using raw data. ①b All phoneme articulations in the English language and individual word articulations spanning the entire phonetic space also evince structured representation on the manifold of SPD matrices and can be classified using simple algorithms such as minimum distance to mean with Riemannian geodesic distance using raw data. ② We demonstrate that sEMG signals can be leveraged to learn meaningful representations of silently voiced speech without relying on time-aligned signals from audibly articulated speech. This is accomplished by training compact neural networks that respect the geometry of SPD matrices and evaluating them on a test set nearly 4 times the size of the training set. Four subjects articulate English sentences in a spelled-out format using NATO phonetic alphabets (e.g., “alfa” for “a”) and we frame silently articulated speech recognition from sEMG as a 26-way classification task, with a top-5 chance accuracy of 0.178. ②a We first develop a tiny model with 9,734 parameters, achieving an average top-5 classification accuracy of 0.63. ②b A second, larger recurrent model with about 150,000 parameters yields an average top-5 accuracy of 0.78, significantly above the chance level. ③ Notably, these models were trained on a dataset nearly four times smaller than the test set. This finding shows that we can build small, easy-to-train architectures that require minimal training data and work well across subjects. Additionally, these models effectively filter out breathing and swallowing patterns intertwined with speech-related sEMG signals, enabling them to learn meaningful speech representations. ④ We show that when represented on the manifold of SPD matrices, besides allowing for design of efficient models as described above, sEMG signal embeddings also allow us to distinguish among different individuals using raw data without needing to train any neural networks. This is unlike audio, where distinguishing among different individuals needs obtaining embeddings by training a neural

network such as the one in Desplanques et al. (2020). This further demonstrates that the manifold of SPD matrices is a powerful representational space for multivariate sEMG timeseries. ④ We also explain the weights learned by the neural networks and present their correlations with the activities of the underlying neuromuscular system.

Geometrical structure of the orofacial neuromuscular signals described here along with the open-sourced data and code sets a strong foundation for development of non-invasive silent speech interfaces and their deployment across various population groups.

4. Data description

We collect sEMG signals from twenty-two muscle sites on the neck, chin, jaw, cheek, and lips. An actiChamp Plus amplifier and associated active electrodes by Brain Vision¹ are used for sEMG signal recording at 5000 Hertz. For establishing proper contact between skin surface and electrodes, we use SuperVisc high viscosity electrolyte gel by Easycap². We develop a software suite in Python environment for providing visual cues to subjects and collating and storing timestamped data³ (see figures 12 and 13 for electrode placement).

We conduct the study in two parts. In the first part, 12 healthy subjects⁴ perform various orofacial gestures related to speech articulation, articulate phonemes in the English language in audible and silent manners, articulate a set of words spanning the entire English language phonetic space in audible and silent manners, and enunciate the Rainbow Passage in audible and silent manners (Fairbanks (1960)). In the second part, 4 healthy subjects articulate NATO phonetic alphabets in a silent manner, enunciate the Rainbow Passage and the Grandfather passage (see supplementary materials for the text of these passages) character-

1. (<https://brainvision.com/products/actichamp-plus/>)

2. (<https://shop.easycap.de/products/supervisc>)

3. We use Lab Streaming Layer for time synchronization. Refer to LSL at <https://labstreaminglayer.org>

4. Although we aim silent speech interfaces for use by people with laryngectomy or dysarthria, since the main focus of this article is about exploring the data structure and verifying it across individuals, we use healthy subjects; each experiment session lasted for about four to five hours, which would have been too strenuous for afflicted individuals.

by-character (a word is articulated continuously character-by-character followed by a break after each word) using NATO phonetic alphabets in a silent manner, and enunciate the Grandfather Passage normally in a silent manner. The first part of the experiment is summarized in table 7. The second part of the experiment is detailed in table 8.

5. Methods and data properties

We construct a complete graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ representing functional connectivity of the orofacial neuromuscular system, where \mathcal{V} denotes the set of vertices (physically — the sensors at different muscle locations) and \mathcal{E} the set of edges ($\mathcal{E} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$). Let $\mathbf{f}_v(t)$ denote the signal at node $v \in \mathcal{V}$ as a function of time t . We update the edge weights over a time window $[t_{\text{Start}}, t_{\text{End}}]$. In a given time window $[t_{\text{Start}}, t_{\text{End}}]$, edge weight between two nodes $v_1, v_2 \in \mathcal{V}$ is calculated as $e_{12} = e_{21} = \mathbf{f}_{v_1}^T \mathbf{f}_{v_2}$ (that is, covariance of the signals at those two nodes in that time window). Therefore, in a time duration $[t_{\text{Start}}, t_{\text{End}}]$, edge (adjacency) matrix is a symmetric positive semi-definite matrix. A positive semi-definite matrix \mathcal{E} can be converted to a positive definite matrix by computing $\mathcal{E} \leftarrow (1 - \eta)\mathcal{E} + \eta \text{trace}(\mathcal{E})\mathcal{I}$ (\mathcal{I} is an identity matrix whose dimensions are the same as \mathcal{E} ; η is a small positive constant usually between 0.1 and 0.2). Therefore, we consider \mathcal{E} as a symmetric positive definite (SPD) matrix and analyze the functional connectivity induced by $\mathbf{f}_v, v \in \mathcal{V}$ on a manifold of SPD matrices. We adapt the Riemannian geometry of symmetric positive definite matrices via Cholesky decomposition as described in Lin (2019).

Edge matrices belonging to different orofacial movements that are used in eliciting articulations (described in table 7) naturally cluster separately on the manifold of symmetric positive definite matrices. t -SNE visualization of the edge matrices using Riemannian geodesic distance is shown in figure 1 and the unsupervised classification accuracy of the edge matrices using k -medoids algorithm with Riemannian geodesic distance is shown in table 1. These show that the edge matrices constructed as described above effectively capture the underlying functional connectivity of the orofacial neuromuscular system.

Idiosyncratic factors unique to individuals such

as physiological and anatomical differences influencing sEMG action potential patterns, neural drive characteristics of an individual, and articulatory features distinctive to individuals induce signal distribution shift across people. This domain shift is captured by the second order signal statistics given by pairwise signal covariances among different sensor nodes represented by the edge matrix. t -SNE visualization of SPD edge matrices as subjects articulated 36 words (each word is repeated 10 times) spanning the entire phonetic space (described in table 7) in an audible manner is shown in figure 2. t -SNE visualization of SPD edge matrices as subjects articulated sentences from the Rainbow Passage is shown in figure 3 (the entire passage is articulated over a duration of 215 seconds and SPD edge matrices are constructed by considering a time window of 100ms; therefore, we have 2150 edge matrices for the entire passage duration). We observe that, for the same linguistic content, edge matrices belonging to different individuals occupy different neighborhoods on the manifold of SPD matrices. This suggests that the underlying functional connectivity of the neuromuscular system across individuals is different. Within an individual, SPD edge matrices evince strong spatial geometric structure that allows us to naturally distinguish different word and phoneme articulations. For example, simple minimum distance to mean (MDM) algorithm using Riemannian distance and Fréchet mean as given in Lin (2019) gives high classification accuracy (see table 2 and table 3).

We now elucidate the domain shift from the perspective of linear transformations in the Euclidean space $\mathbb{R}^{|\mathcal{V}|}$. We make the following observations. SPD edge matrices of a given individual can be approximately diagonalized for all phonetic articulations. That is, edge matrix \mathcal{E} for any articulation in a given individual can be expressed as $\mathcal{E} = Q\Sigma Q^T$, where Σ is approximately diagonal (since all \mathcal{E} are real symmetric matrices, Q can be constrained to be orthogonal). Q that can approximately diagonalize a set of edge matrices is different for different individuals (see figure 5 and figure 6).

In essence, SPD edge matrices constructed using pairwise signal covariances of sEMG sensor signals at different spatial locations encode information that can readily distinguish different articulations (figure 1, tables 1, 2, and 3). All edge matrices

(spanning the entire phonetic articulation space) of a given individual have approximately the same eigenbasis vectors. That is, orofacial articulations can be described as linear combination of $|\mathcal{V}|$ vectors, each with dimension $|\mathcal{V}|$. The eigenbasis vectors and their linear combination are different for different individuals. This basis change is precisely what domain shift in sEMG signals means (figure 4).

With sEMG articulations represented on the Riemannian manifold, we can apply methods described in Huang and Van Gool (2017) for SPD matrix learning to decode different articulations. We describe the architecture of the neural network in figure 14 and briefly recapitulate the core formulations in Huang and Van Gool (2017) here. The neural network architecture for learning discriminative SPD matrix representations is made of three layer types. First, a layer is defined by $\mathcal{E}_k = W_k^T \mathcal{E}_{k-1} W_k$, where an SPD edge matrix \mathcal{E}_{k-1} of dimensions $c_{k-1} \times c_{k-1}$ is input to k -th layer. W_k is of dimensions $c_{k-1} \times c_k$ giving rise to \mathcal{E}_k of dimensions $c_k \times c_k$. W_k is constrained to be a full-rank semi-orthogonal matrix on a compact Stiefel manifold such that $W_k^T W_k = I$. Second, a non-linear layer is defined by $\mathcal{E}_k = U_{k-1} \max(\epsilon I, \Sigma_{k-1}) U_{k-1}^T$. Third, a layer to map SPD matrices from the manifold space to its tangent space (so that the Euclidean operations can be applied) is defined by $\mathcal{E}_k = U_{k-1} \log(\Sigma_{k-1}) U_{k-1}^T$. U_{k-1} , Σ_{k-1} are obtained by eigendecomposition of matrix \mathcal{E}_{k-1} and ϵ is a small constant > 0 .

For backpropagation, the gradient of the loss function L with respect to W_k , when restricted to the tangent space of the Stiefel manifold (denoted by $\mathbb{R}^{c_{k-1} \times c_k}$, the space of all full-rank matrices of dimension $c_{k-1} \times c_k$) is given by

$$\nabla L_{W_k}^{(k)} = \nabla L_{W_k(\text{Euclidean})}^{(k)} - W_k \nabla L_{W_k(\text{Euclidean})}^{T(k)} W_k.$$

The gradient is updated as

$$W_k \leftarrow W_k - \lambda \nabla L_{W_k}^{(k)},$$

where λ is the learning rate. W_k is then mapped back to the Stiefel manifold (from the tangent space) via orthogonalization. We use Gram-Schmidt method for matrix orthogonalization. Refer to Huang and Van Gool (2017) for backpropagation through non-linear and tangent space mapping layers. Matrices Q in figures 5 and 6 are obtained by training the network in figure 14. The network

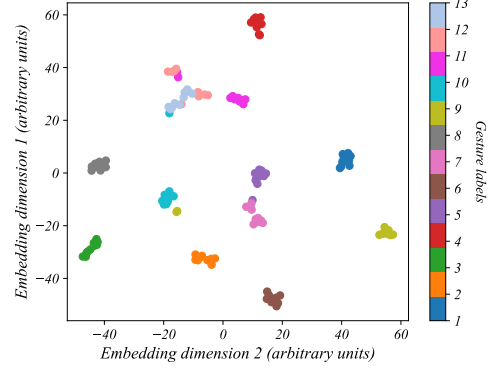


Figure 1: Different orofacial gestures are naturally distinguishable on the manifold of SPD matrices. t -SNE of edge matrices of various orofacial movements described in table 7 for subject 1 using Riemannian geodesic distance. Edge matrices are constructed with $\eta = 0.1$, $t_{\text{start}} = 0$, and $t_{\text{end}} = 1.5$ s. Embedding is colored according to gestures.

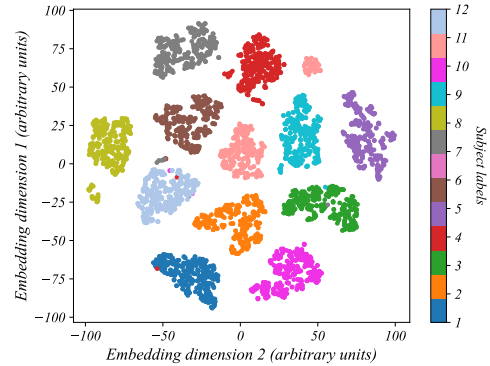


Figure 2: Different subjects are naturally distinguishable on the manifold of SPD matrices. t -SNE of edge matrices of voiced word articulations described in table 7 for all subjects using Riemannian geodesic distance. Edge matrices are constructed with $\eta = 0.1$, $t_{\text{start}} = 0$, and $t_{\text{end}} = 1.5$ s. Embedding is colored according to subjects.

is trained separately for each individual and matrix Q for an individual is $W_{22 \times 22}^{(1)}$ when the model gave the best decoding accuracy on the test set.

Subject number	Unsupervised k -medoids clustering using Riemannian geodesic distance
1	0.877
2	0.862
3	0.677
4	0.638
5	0.654
6	0.915
7	0.554
8	0.515
9	0.846
10	0.854
11	0.731
12	0.715
Mean	0.736

Table 1: Different orofacial gestures are naturally distinguishable on the manifold of SPD matrices. Classification accuracy of 13 orofacial movements described in table 7. Chance accuracy is 0.077. Edge matrices are constructed with $\eta = 0$, $t_{\text{start}} = 0$, and $t_{\text{End}} = 1.5\text{s}$.

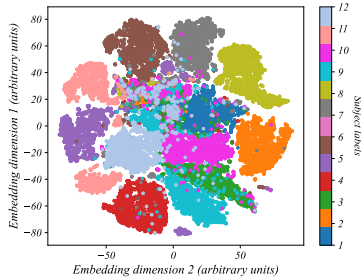


Figure 3: Different subjects are naturally distinguishable on the manifold of SPD matrices. t -SNE of edge matrices of voiced Rainbow Passage articulations described in table 7 for all subjects using Riemannian geodesic distance. Edge matrices are constructed with $\eta = 0.1$, $t_{\text{End}} - t_{\text{Start}} = 100\text{ms}$. Embedding is colored according to subjects.

Classification accuracy of the model as it decoded 36 different word and 38 different phoneme articulations (table 7) are given in tables 6 and 4 respectively.

We adapt methods from Jeong et al. (2023) for continuous learning of manifold valued multivariate

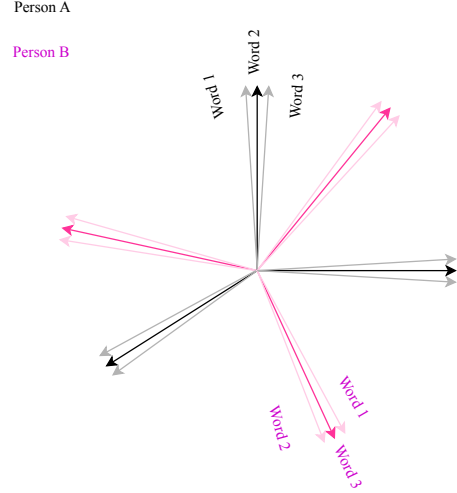


Figure 4: Conceptual depiction of 3 orthogonal axes - basis vectors in shades of pink belong to person A and basis vectors in shades of black belong to person B. SPD edge matrices of sEMG signals from all articulations of a given individual have approximately a same set of eigenbasis vectors that span the space $\mathcal{R}^{|\mathcal{V}|}$. Signals from different individuals have different eigenbasis vectors.

sEMG timeseries data. Jeong et al. (2023) use a recurrent neural network with gated recurrent units by redefining the arithmetic operations in the Euclidean space in Cho et al. (2014) with the corresponding operations that respect the geometry of the manifold of SPD matrices. Specifically, weighted summations are replaced by Fréchet mean and bias addition is replaced by a smooth commutative group operator (\odot) in the SPD manifold space. We briefly summarize the concepts explained in Jeong et al. (2023) here. An SPD edge matrix \mathcal{E} is mapped to Cholesky space by a differentiable function \mathcal{L} such that $\mathcal{E} = LL^T$, where L is a lower triangular matrix. When the diagonal of L is restricted to be positive, such a mapping is unique. Therefore, \mathcal{L} and its inverse function \mathcal{S} are diffeomorphisms. `split`(L) splits the matrix L into $\mathbb{D}(L)$, the diagonal part of matrix L and $\lfloor L \rfloor$, the strictly lower triangular part of matrix L . `combine` is the inverse operation of `split` that adds the matrices together such that

$L = \mathbb{D}(L) + \lfloor L \rfloor$. The relationships are summarized below.

$$\mathcal{E} \xrightarrow[\mathcal{S}]{\mathcal{L}} L \xrightarrow[\text{combine}]{\text{split}} \{\mathbb{D}(L), \lfloor L \rfloor\}. \quad (1)$$

For L and K in the Cholesky space, operation \odot is defined as (Lin (2019))

$$L \odot K = \lfloor L \rfloor + \lfloor K \rfloor + \mathbb{D}(L)\mathbb{D}(K). \quad (2)$$

For L_1, L_2, \dots, L_n in the Cholesky space, and positive weights w_1, w_2, \dots, w_n , Fréchet mean \mathbb{E}_F is defined as

$$\mathbb{E}_F = \frac{1}{n} \sum_{i=1}^n w_i \lfloor L_i \rfloor + \exp \left(\frac{1}{n} \sum_{i=1}^n w_i \log(\mathbb{D}(L_i)) \right). \quad (3)$$

We update the gates of the recurrent neural network as below. Update-gate z_t at time-step t is

$$z_t = \sigma(w_z \lfloor l_t \rfloor + u_z \lfloor h_{t-1} \rfloor + b_z) + \sigma(b_{z'} [\exp(w_{z'} \log(\mathbb{D}(l_t))) + u_{z'} \log(\mathbb{D}(h_{t-1}))]). \quad (4)$$

Reset-gate r_t at time-step t is

$$r_t = \sigma(w_r \lfloor l_t \rfloor + u_r \lfloor h_{t-1} \rfloor + b_r) + \sigma(b_{r'} [\exp(w_{r'} \log(\mathbb{D}(l_t))) + u_{r'} \log(\mathbb{D}(h_{t-1}))]). \quad (5)$$

Candidate-activation vector \hat{h}_t is

$$\begin{aligned} \hat{h}_t = & \text{tanh}(w_h \lfloor l_t \rfloor + u_h (\lfloor r_t \rfloor * \lfloor h_{t-1} \rfloor) + b_h) + \\ & \text{softplus}(b_{h'} \exp(w_{h'} \log(\mathbb{D}(l_t))) \\ & + u_{h'} \log(\mathbb{D}(r_t) * \mathbb{D}(h_{t-1}))). \end{aligned} \quad (6)$$

Output vector h_t is

$$h_t = (1 - \lfloor z_t \rfloor) * \lfloor h_{t-1} \rfloor + \lfloor z_t \rfloor * \lfloor \hat{h}_t \rfloor + \exp((1 - \mathbb{D}(z_t)) * \log(\mathbb{D}(h_{t-1})) + \mathbb{D}(z_t) * \log(\mathbb{D}(\hat{h}_t))). \quad (7)$$

In the above equations, $w_z, u_z, b_z, w_{z'}, u_{z'}, w_r, u_r, b_r, w_{r'}, u_{r'}, w_h, u_h, b_h, w_{h'},$ and $u_{h'}$ are real weights. $b_{z'}, b_{r'},$ and $b_{h'}$ are real positive weights. $*$ is Hadamard matrix multiplication. l_t is the input SPD matrix at time-step t and h_{t-1} is the hidden-state at time-step $t - 1$. As given in Jeong et al. (2023), we use manifold neural ordinary differential equations (Lou et al. (2020), Chen et al. (2018)) for modeling the dynamics of multivariate sEMG data. The dynamics f of sEMG data is modeled by a neural network with parameters Θ . The output state h_t is updated as,

$$\begin{aligned} h_{t-1} \leftarrow \text{ODESolve}(f_\Theta, \widetilde{\text{Log}}(h_{t-1}), (t-1, t)) \\ h_t = \text{GRU}(l_t, \widetilde{\text{Exp}}(h_{t-1})), \end{aligned} \quad (8)$$

where $\widetilde{\text{Log}}$ is the logarithm mapping from the manifold space of SPD matrices to its tangent space and $\widetilde{\text{Exp}}$ is its inverse operation as defined in Lin (2019). GRU is a gated recurrent unit whose gates are given by equations 4 - 7. We create a model by combining SPD-matrix learning and manifold GRU with neural ordinary differential equations as shown in figure 15.

Since multiple muscles must synchronize their activities for articulations whose functional connectivity is captured by SPD edge matrices, domain shift is approximately characterized as linear transformation in the Euclidean space $\mathbb{R}^{|\mathcal{V}|}$. This is unlike audio, where statistical pooling with attention-mechanisms on a single dimension can lead to good representation learning of linguistic content (such as in Baevski et al. (2020) and Hsu et al. (2021)) or subject specific content (speaker identification such as in Desplanques et al. (2020)). Therefore, multivariate sEMG signals require appropriate geometric priors unavailable in convolutional neural networks (which owe their effectiveness to shared weights in convolutional layers that can efficiently encode statistics of functions sampled on 1D or 2D grids) for efficient learning. We show that the manifold of SPD matrices is a space that naturally encodes information that distinguishes linguistic content (tables 2 and 3) as well as idiosyncrasies particular to individuals (figures 2 and 3). We leverage this geometric prior using models in figure 14 and figure 15 and report the classification accuracy in tables 6, 4, and 5 and in figures 9, 10, and 11.

6. Results

6.1. Experiment 1

Word and phonetic articulations in table 7 are produced in 4 different sessions. Subjects articulated phonemes and words in an audible manner in the first 2 sessions and in a silent manner in the next 2 sessions. Within each session, subjects articulated a given word or phoneme 5 times. Therefore, we have 10 instances each of audible and silent articulations for all words and phonemes. We use first 3 instances from each session for training and last two instances for testing the models. Therefore, we have 6 instances of each word or phoneme for training and 4 instances for testing. Phonetic articulation data of subject 11 was corrupted during acquisition and is unavailable.

For MDM and architecture in figure 14, edge matrices are created using the time-context window of 1.5s (the entire duration of articulation). For GRU architecture in figure 15, edge matrices are created with a sliding window of context size of 150ms and step size of 30ms. Results for the experiments in table 7 are given in tables 6, 4, and 5. The presented accuracy values are averaged over 10 random seeds (the model in figure 14 is trained for 1000 epochs and the model in figure 15 is trained for 150 epochs. Best accuracy on the test set in reported). Model in figure 14 contains about 5000 to 11000 parameters (depending on the size of the linear classification head) and the model in figure 15 has approximately 150000 parameters.

Subject number	Accuracy of audible words	Accuracy of silent words
1	0.632	0.389
2	0.660	0.590
3	0.389	0.264
4	0.264	0.243
5	0.632	0.708
6	0.688	0.569
7	0.382	0.125
8	0.549	0.472
9	0.611	0.382
10	0.625	0.444
11	0.587	0.674
12	0.5	0.403
Mean	0.544	0.439

Table 2: Word articulations are naturally distinguishable on the manifold of SPD matrices. Classification accuracy of word articulations using minimum distance to mean algorithm (see appendix C for algorithm description). It is a 36-way classification problem with chance accuracy of 0.028.

6.2. Experiment 2

We present the results for articulations described in table 8 here. All articulations in table 8 are performed in a silent manner. Subjects articulated individual NATO phonetic alphabet codes in 4 different sessions. In each session, individual words are repeated 5 times. Therefore, we have 20 repetitions for each code. We use 16 repetitions of each code for training the models in figure 14 and 15 (first 4 repetitions from each session; model in figure 14 has 9734 parameters and the GRU model in figure 15 has about 150000 parameters), train the model in figure 14 for 1000 epochs and the

Subject number	Accuracy of audible phonemes			Accuracy of silent phonemes		
	A	C	V	A	C	V
1	0.38	0.34	0.52	0.5	0.47	0.57
2	0.51	0.47	0.67	0.49	0.42	0.63
3	0.26	0.18	0.45	0.28	0.29	0.3
4	0.20	0.23	0.25	0.33	0.38	0.43
5	0.47	0.52	0.53	0.45	0.41	0.58
6	0.5	0.46	0.58	0.45	0.41	0.58
7	0.19	0.25	0.2	0.27	0.28	0.33
8	0.33	0.46	0.33	0.30	0.33	0.42
9	0.36	0.26	0.6	0.23	0.17	0.5
10	0.41	0.41	0.48	0.31	0.23	0.63
12	0.35	0.41	0.38	0.32	0.39	0.4
Mean	0.36	0.36	0.45	0.36	0.34	0.49

Table 3: Phoneme articulations are naturally distinguishable on the manifold of SPD matrices. Classification accuracy of phoneme articulations using minimum distance to mean algorithm (see appendix C for algorithm description). It is a 38-way classification problem with chance accuracy of 0.026. We also show classification accuracy for 23 consonant phonemes and 15 vowel phonemes separately. A-all phonemes, C-consonants only, V-vowels only.

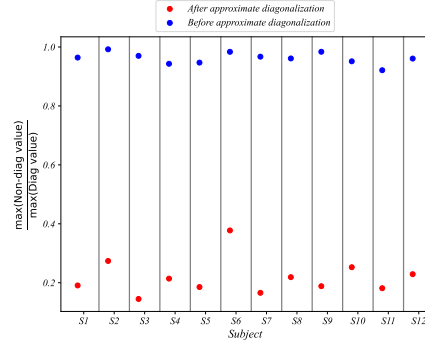


Figure 5: All edge matrices within an individual can be approximately diagonalized. A matrix Q is learned for a given subject such that for any edge matrix \mathcal{E} , $\mathcal{E} = Q\Sigma Q^T$ (Q is $W_{22 \times 22}^{(1)}$ from figure 14). Ratio of maximum non-diagonal value to maximum diagonal value averaged over all 360 trials before and after approximate diagonalization are shown. Matrices \mathcal{E} are constructed with $\eta = 0$.

GRU model in figure 15 for 100 epochs and validate it on the validation set (last repetition of a word from each session. In total, we have 4 repetitions for every word in the validation set). We use the model weights that give the best accuracy on the validation set and test it on Rainbow Passage (1429

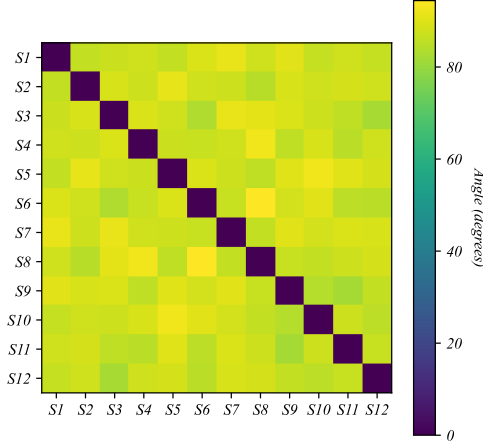


Figure 6: Articulations differ substantially across individuals. Approximate eigenbasis vectors ($Q = W_{22 \times 22}^{(1)}$) are different for different individuals.

$$\theta = \cos^{-1} \left(\frac{\text{trace}(Q_i Q_j^T)}{\sqrt{\text{trace}(Q_i Q_i^T)} \sqrt{\text{trace}(Q_j Q_j^T)}} \right)$$

between approximate eigenbasis matrices Q_i and Q_j of different individuals i and j .

character articulations) and Grandfather Passage (541 character articulations) articulations using NATO phonetic codes and present the top- k ($k = 1$ to 5) decoding accuracy for each passage in figure 9 and figure 10. Chance top-5 decoding accuracy is 0.178 (it is 26-way classification). The average top-5 decoding accuracy using model in figure 14 is 0.64 for Rainbow Passage and 0.612 for Grandfather Passage across all 4 subjects. The average top-5 decoding accuracy using model in figure 15 is 0.82 for Rainbow Passage and 0.73 for Grandfather Passage across all 4 subjects. SPD matrices for model in figure 14 are created with a time context window of 1.5s and the SPD matrices for model in figure 15 are created with a sliding window of context size 150ms and step size of 30ms.

In passages, a word is articulated continuously character-by-character with a break between different words. For example, the word **RAINBOW** is articulated as **Romeo-Alfa-India-November-Bravo-Oscar-Whiskey**. Cues to articulate characters are given every 1.5s and articulation of a given character is influenced by preceding and succeeding characters (unlike training set, where every articulation was preceded and succeeded by a rest period).

As seen in figures 7 and 8, frequency of occurrence for many letters is higher than 16, the number of instances of each character in the training set. This demonstrates that by utilizing the inherent geometric structure of sEMG signals, we can potentially build small models that can be trained with small datasets which can generalize well.

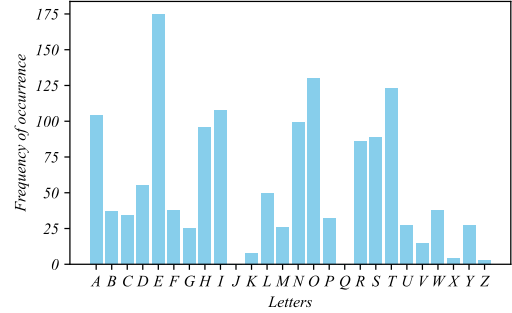


Figure 7: Frequency of letters in the Rainbow Passage

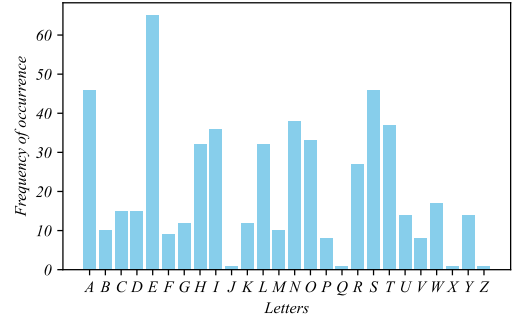


Figure 8: Frequency of letters in the Grandfather Passage

Next, Grandfather Passage was split into 21 sentences and subjects articulated the sentences as they would naturally speak in the English language in a silent manner (each sentence is repeated 6 times). We use the first 4 articulations of each sentence for training and last 2 articulations for testing the models in figures 14 and 15. For training the model in figure 14, edge matrices are created using the entire 4s window duration. For training the model in figure 15, edge

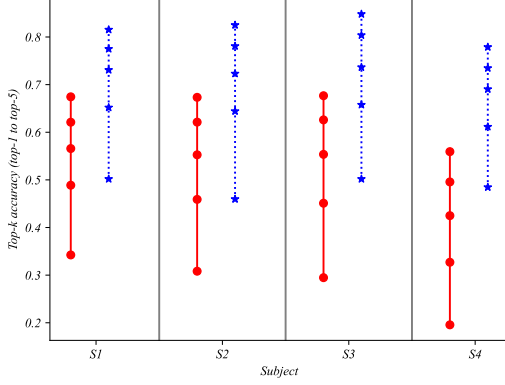


Figure 9: Top-1 to top-5 accuracy of Rainbow Passage articulation with 1429 characters. It is a 26-way classification and chance top-5 accuracy is 0.178. Chance top-1 accuracy is 0.038. Red - accuracy for model in figure 14. Blue - accuracy for model in figure 15.

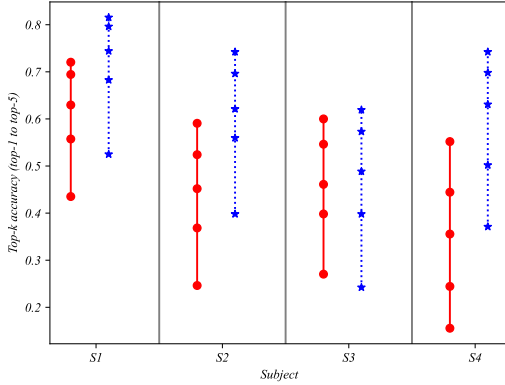


Figure 10: Top-1 to top-5 accuracy of Grandfather Passage articulation with 541 characters. It is a 26-way classification and chance top-5 accuracy is 0.178. Chance top-1 accuracy is 0.038. Red - accuracy for model in figure 14. Blue - accuracy for model in figure 15.

matrices are created with a sliding window of context size of 400ms and step size of 100ms. Accuracy averaged over 10 random seeds are reported in figure 11. It is a 21-way classification problem with chance accuracy of 0.047 (the model in figure 14 has about

8000 parameters and the model in figure 15 has about 150000 parameters).

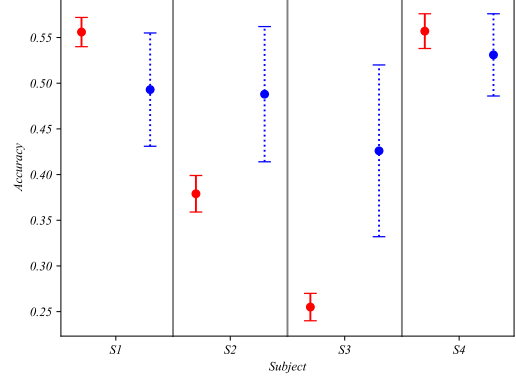


Figure 11: Red - accuracy of Grandfather passage sentence classification as subjects articulate the sentences as they would naturally speak in the English language in a silent manner using model in figure 14. Blue - accuracy of Grandfather passage sentence classification using model in figure 15.

7. Importance of electrodes in the signal graph

We analyze the weights learned by the model in figure 14 and explore corresponding correlations with the muscle activities. We show that the model learns meaningful representations that reflect the underlying functions of the orofacial neuromuscular system.

In figure 14, each column of $\mathcal{E}_{22 \times 22}^{(0)}$ captures the relationship of a given node with every other node. Each column of $W_{22 \times 22}^{(1)}$ can be approximately expressed as a linear combination of columns of $\mathcal{E}_{22 \times 22}^{(0)}$ (since $W_{22 \times 22}^{(1)}$ is approximately the eigenbasis of any $\mathcal{E}_{22 \times 22}^{(0)}$). That is,

$$w \approx \kappa \mathcal{E}_{22 \times 22}^{(0)}. \quad (9)$$

In the above equation w is a column of $W_{22 \times 22}^{(1)}$ and κ is a vector of $|\mathcal{V}|$ coefficients. We solve for κ using least squares solution (`numpy.linalg.lstsq`). We obtain κ_{λ_G} corresponding to w_{λ_G} - the column vector with the largest eigenvalue.

We compute κ_{λ_G} for all 360 audible word articulations (36 words in table 7, each repeated 10 times). For each articulation, we rank the coefficients κ_{λ_G} according to their absolute values in decreasing order. We present the locations of three electrodes that are ranked at the top, the most number of times across all 360 trials in figure 17. The frequency of top-rank for each of the 22 electrodes across 360 trails is summarized in figure 18. That is, a node that contributes the most to the eigenvector with the largest eigenvalue is considered as the most important node (eigenvector and eigenvalue are used in approximate sense as $\mathcal{E}_{22 \times 22}^{(1)} = W_{22 \times 22}^{(1)T} \mathcal{E}^{(0)} W_{22 \times 22}^{(1)}$ is only approximately diagonal). It is a reasonable choice as the largest diagonal value in $\mathcal{E}_{22 \times 22}^{(1)}$ is consistently much higher than the others (see figure 16). In appendix D, we explain the anatomical significance of the most important nodes.

Additionally, we also analyze the phoneme errors to show that the model in figure 14 learns representations that correlate well with place and manner of articulation. Based on the placement and manner of articulation, consonant phonemes are classified into seven groups - bilabial, labiodental, dental, alveolar, postalveolar, velar, and approximant consonants. During audible phoneme consonant classification, about 44 percent of the mistakes made by the model in figure 14 correspond to misclassifications within the same group of phonemes (for example, classification of bilabial Paa as Baa or Maa or classifying the velar Zhaa as Chaa, Shaa, or Jhaa). When such misclassifications are allowed, the average accuracy of consonant phoneme classification jumps from 0.53 (table 4) to 0.74, nearly 40 percent increase in classification accuracy. This jump is practically important as such systematic mistakes can be corrected using language models.

8. Conclusion

We present the largest open-sourced dataset of orofacial sEMG signals as individuals articulate speech. We explain the inherent structure of multivariate sEMG timeseries and demonstrate that we can build small models that can be trained with little data. We also demonstrate that we can learn meaningful representations of silently articulated speech using sEMG without need for parallel audibly articulated

sEMG data corpora or time-alignments.

① Since data collection is expensive and time-consuming, building models that can be trained with very little data increases the feasibility of silent-speech interface. ①a In the second part of the experiment, we train the models with just 13 minutes of speech data and demonstrate that it generalizes well for continuously spelled-out sentences using NATO phonetic codes. ② For the first time, we quantify and explain the domain shift of sEMG signals across individuals. ③ We demonstrate that the models presented here are interpretable and explain the correlations between the orofacial neuromuscular system and the learned weights.

These experiments and models establish a foundation for developing non-invasive speech prostheses using sEMG signals, addressing key challenges that will facilitate rapid, scalable deployment of these devices across diverse individuals.

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.

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: Mathematical formulation, concepts development, data analysis, experiment design, data collection software design, data collection, manuscript preparation.
- Zachary D. McNaughton: Data collection.
- Lee M. Miller: Concepts development and manuscript preparation.

References

- Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 12449–12460, 2020.
- Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL <https://aclanthology.org/D14-1179>.
- Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *Interspeech*, 2020.
- 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.
- G. Fairbanks. *Voice and Articulation Drillbook*. Harper, 1960. ISBN 9780060419905. URL <https://books.google.com/books?id=qN1ZAAAAMAAJ>.
- Dario Farina, Roberto Merletti, and Roger M Enoka. The extraction of neural strategies from the surface emg: an update. *Journal of applied physiology*, 117(11):1215–1230, 2014.
- 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.
- 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.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460, oct 2021. ISSN 2329-9290. doi: 10.1109/TASLP.2021.3122291. URL <https://doi.org/10.1109/TASLP.2021.3122291>.
- Zhiwu Huang and Luc Van Gool. A riemannian network for spd matrix learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.
- 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. doi: 10.1109/TASLP.2017.2738568.
- Seungwoo Jeong, Wonjun Ko, Ahmad Wisnu Mulyadi, and Heung-Il Suk. Deep efficient continuous manifold learning for time series modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

- 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.
- 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.
- Zhenhua Lin. Riemannian geometry of symmetric positive definite matrices via cholesky decomposition. *SIAM Journal on Matrix Analysis and Applications*, 40(4):1353–1370, 2019.
- Aaron Lou, Derek Lim, Isay Katsman, Leo Huang, Qingxuan Jiang, Ser Nam Lim, and Christopher M De Sa. Neural manifold ordinary differential equations. *Advances in Neural Information Processing Systems*, 33:17548–17558, 2020.
- Geoffrey S. Meltzner, James T. Heaton, Yunbin Deng, Gianluca De Luca, Serge H. Roy, and Joshua C. Kline. Silent speech recognition as an alternative communication device for persons with laryngectomy. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 25(12):2386–2398, dec 2017. ISSN 2329-9290. doi: 10.1109/TASLP.2017.2740000. URL <https://doi.org/10.1109/TASLP.2017.2740000>.
- Michael Wand, Matthias Janke, and Tanja Schultz. The emg-uka corpus for electromyographic speech processing. In *INTERSPEECH*, pages 1593–1597, 2014.

Appendix A.

Subject number	Accuracy for model in figure 14					
	Audible phonemes			Silent phonemes		
	<i>A</i>	<i>C</i>	<i>V</i>	<i>A</i>	<i>C</i>	<i>V</i>
1	0.574±0.012	0.632±0.015	0.725±0.013	0.574±0.012	0.647±0.016	0.678±0.013
2	0.679±0.012	0.645±0.012	0.857±0.008	0.679±0.012	0.575±0.011	0.798±0.016
3	0.404±0.008	0.428±0.013	0.632±0.014	0.404±0.008	0.393±0.009	0.412±0.008
4	0.310±0.014	0.370±0.015	0.487±0.023	0.310±0.014	0.467±0.014	0.670±0.015
5	0.637±0.007	0.715±0.008	0.660±0.008	0.637±0.007	0.566±0.013	0.667±0.017
6	0.577±0.012	0.542±0.011	0.748±0.019	0.577±0.012	0.550±0.012	0.688±0.013
7	0.313±0.007	0.386±0.007	0.373±0.011	0.313±0.007	0.334±0.014	0.408±0.013
8	0.348±0.036	0.518±0.021	0.563±0.026	0.348±0.036	0.390±0.010	0.605±0.021
9	0.522±0.009	0.479±0.014	0.738±0.017	0.522±0.009	0.305±0.009	0.723±0.013
10	0.603±0.015	0.620±0.014	0.762±0.027	0.603±0.015	0.390±0.017	0.755±0.011
12	0.485±0.011	0.550±0.012	0.577±0.019	0.485±0.011	0.455±0.012	0.597±0.015
Mean	0.496	0.535	0.647	0.447	0.461	0.636

Table 4: Classification accuracy of phoneme articulation using model in figure 14. We also show classification accuracy for 23 consonant phonemes and 15 vowel phonemes separately. *A*-all phonemes, *C*-consonants only, *V*-vowels only. It is a 38-way classification problem with chance accuracy of 0.026.

Subject number	Accuracy for model in figure 15					
	Audible phonemes			Silent phonemes		
	<i>A</i>	<i>C</i>	<i>V</i>	<i>A</i>	<i>C</i>	<i>V</i>
1	0.651±0.030	0.699±0.033	0.637±0.027	0.574±0.014	0.596±0.030	0.662±0.044
2	0.659±0.023	0.625±0.026	0.815±0.046	0.590±0.017	0.516±0.020	0.780±0.031
3	0.398±0.030	0.370±0.049	0.498±0.038	0.328±0.020	0.353±0.027	0.445±0.050
4	0.383±0.019	0.401±0.038	0.490±0.033	0.380±0.025	0.405±0.027	0.558±0.037
5	0.566±0.033	0.597±0.030	0.602±0.047	0.507±0.021	0.533±0.027	0.615±0.040
6	0.591±0.030	0.608±0.037	0.710±0.048	0.521±0.020	0.558±0.041	0.630±0.052
7	0.313±0.014	0.388±0.041	0.338±0.038	0.288±0.014	0.297±0.030	0.465±0.026
8	0.419±0.030	0.532±0.050	0.473±0.052	0.418±0.020	0.376±0.039	0.587±0.041
9	0.470±0.027	0.458±0.051	0.663±0.036	0.380±0.019	0.297±0.028	0.620±0.031
10	0.633±0.033	0.613±0.041	0.787±0.039	0.426±0.029	0.400±0.031	0.667±0.039
12	0.444±0.025	0.505±0.037	0.460±0.038	0.419±0.024	0.402±0.029	0.568±0.047
Mean	0.503	0.527	0.588	0.439	0.430	0.600

Table 5: Classification accuracy of phoneme articulation using model in figure 15. We also show classification accuracy for 23 consonant phonemes and 15 vowel phonemes separately. *A*-all phonemes, *C*-consonants only, *V*-vowels only. It is a 38-way classification problem with chance accuracy of 0.026.

Subject number	Accuracy for model in figure 14		Accuracy for model in figure 15	
	Audible words	Silent words	Audible words	Silent words
1	0.747 \pm 0.010	0.529 \pm 0.010	0.788 \pm 0.029	0.669 \pm 0.055
2	0.883 \pm 0.009	0.765 \pm 0.008	0.865 \pm 0.013	0.745 \pm 0.025
3	0.599 \pm 0.010	0.443 \pm 0.008	0.656 \pm 0.041	0.542 \pm 0.033
4	0.340 \pm 0.011	0.423 \pm 0.010	0.305 \pm 0.026	0.381 \pm 0.034
5	0.821 \pm 0.012	0.792 \pm 0.012	0.769 \pm 0.024	0.712 \pm 0.037
6	0.807 \pm 0.011	0.692 \pm 0.014	0.767 \pm 0.034	0.674 \pm 0.027
7	0.421 \pm 0.010	0.260 \pm 0.006	0.427 \pm 0.025	0.274 \pm 0.025
8	0.694 \pm 0.012	0.578 \pm 0.012	0.755 \pm 0.038	0.661 \pm 0.025
9	0.722 \pm 0.007	0.458 \pm 0.010	0.690 \pm 0.025	0.523 \pm 0.037
10	0.803 \pm 0.010	0.722 \pm 0.011	0.846 \pm 0.026	0.738 \pm 0.021
11	0.708 \pm 0.008	0.772 \pm 0.006	0.781 \pm 0.022	0.810 \pm 0.024
12	0.634 \pm 0.012	0.570 \pm 0.014	0.681 \pm 0.035	0.606 \pm 0.036
Mean	0.681	0.584	0.694	0.611

Table 6: Classification accuracy of word articulations using models in figure 14 and 15. It is a 36-way classification problem with chance accuracy of 0.028.

Appendix B.



Figure 12: Placement of electrodes on the neck region.

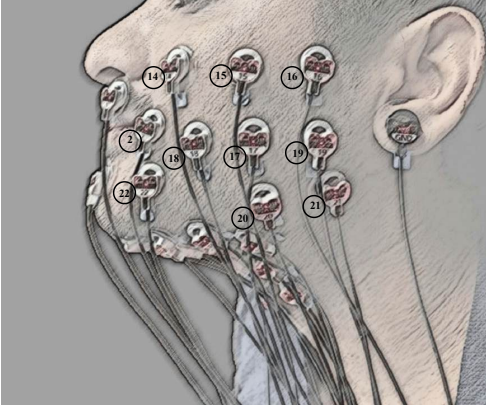


Figure 13: Placement of electrodes on cheek and lip regions. Electrode 1 is above the upper lip and electrode 3 is below the lower lip.

In figure 14, ReEig is defined as $\mathcal{E}_k = U_{k-1} \max(\epsilon I, \Sigma_{k-1}) U_{k-1}^T$ and LogEig is defined as $\mathcal{E}_k = U_{k-1} \log(\Sigma_{k-1}) U_{k-1}^T$. U_{k-1} , Σ_{k-1} are

obtained by Eigen decomposition of matrix \mathcal{E}_{k-1} and ϵ is a small constant > 0 .

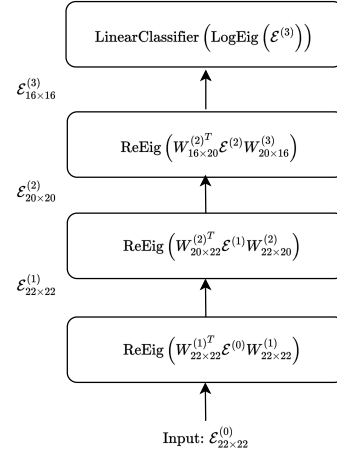


Figure 14: Neural network architecture for SPD matrix learning.

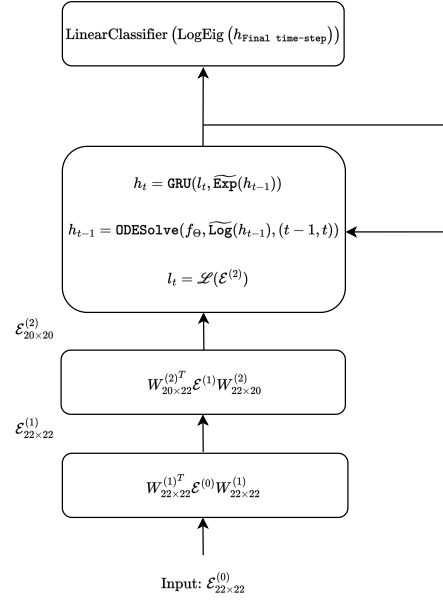


Figure 15: Neural network architecture with SPD matrix learning and manifold gated recurrent unit (GRU).

Appendix C.

Minimum distance to mean algorithm on the manifold is defined as below:

Given M classification classes and N training samples, SPD matrices in the training set $\{\mathcal{E}_n^m\}$, where $n \in \{1, 2, \dots, N\}$ and $m \in \{1, 2, \dots, M\}$ are used to construct centroids for each of the M classes such that the centroid of class m is,

$$\mathcal{C}^m = \mathbb{E}_F(\{\mathcal{L}(\mathcal{E}^m)\}), \quad (10)$$

where the Fréchet mean is calculated according to equation 3. Given a test dataset of SPD matrices $\{\mathcal{T}\}$, $T \in \mathcal{T}$ is assigned to that class whose centroid is nearest to $\mathcal{L}(T)$. That is, the class of T is

$$\arg \min_m d_{\mathcal{L}_c^+}(\mathcal{L}(T), \mathcal{C}^m),$$

where $d_{\mathcal{L}_c^+}$ is the Riemannian geodesic distance.

Appendix D.

Various orofacial muscles roughly correspond to the following electrode locations. Muscles in the upper and posterior cheek area, including **masseter** and **temporalis**, which move the jaw; and muscles in the lower cheek region, including **hyoglossus**, **palatoglossus**, and **styloglossus** which help tongue movements roughly correspond to electrodes 17, 19, 20, and 21 in figure 13; **Zygomaticus** muscle which helps movement of the upper lip corresponds to region approximately around the electrode nodes 22, 18, and 15 in figure 13. Electrodes under the jaw overlie muscles that move the tongue and those that coordinate the tongue with the soft palate, including **genioglossus** (region approximately around the electrode nodes 8 and 9 in figure 12), as well as those that lower the mandible such as **digastric**. Electrodes near the larynx overlie muscles that control hyoid bone and larynx height (region approximately around the electrode nodes 6, 7, 10, and 11 in figure 12) which help in vowel and pitch production (e.g. **sternohyoid**) as well as jaw and tongue movement (**digastric**, **stylohyoid**).

As we see in figure 17, nodes that are most important to decoding are mostly different across subjects. This is consistent with the observation that the approximate eigenbasis vectors that define the articulations are different for different subjects (figure 6). Such differences manifest due to varying

physiology as well as idiosyncratic speaking styles of individuals. However, electrode nodes underlying **hyoglossus**, **palatoglossus**, and **styloglossus** appear as most important nodes in many individuals.

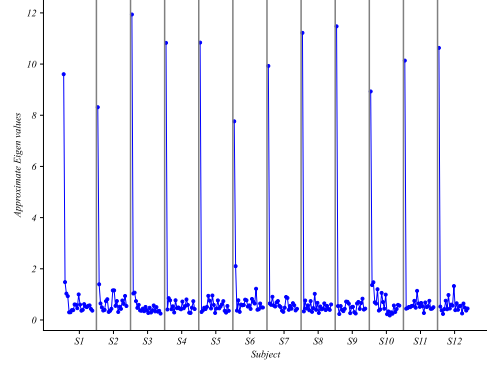


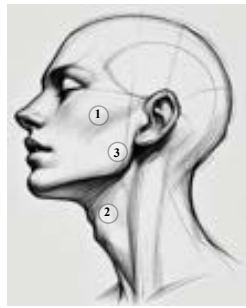
Figure 16: Diagonal values of $\mathcal{E}_{22 \times 22}^{(1)}$ averages over all 360 trails of audible word articulations in table 7. The 22 diagonal values shall be seen as approximate eigenvalues of matrices $\mathcal{E}_{22 \times 22}^{(0)}$. Note that the largest eigenvalue is much higher than others.

	Gestures	Number of gesture repetitions	Individual articulation duration
OROFACIAL MOVEMENTS	Cheeks - puff out, cheeks - suck in, jaw - dropdown, jaw - move backward, jaw - move forward jaw - move left, jaw - move right, lips - pucker, lips - smile, lips - tuck as if blotting, tongue - back of lower teeth, tongue - back of upper teeth tongue - the roof of the mouth	13 gestures are repeated 10 times each (5 + 5 in two different sessions).	1.5 seconds
PHONEMES	Bilabial consonants: Baa, Paa, Maa Labiodental consonants: Faa, Vaa Dental consonants: Thaa, Dhaa Alveolar consonants: Taa, Daa, Naa, Saa, Zaa Post alveolar consonants: Chaa, Shaa, Jhaa, Zhaa Velar consonants: Kaa, Gaa, NGaa Approximant consonants: Yaa, Raa, Laa, Waa Vowels and Diphthongs: OY as in bOY, OW as in nOW, AO as in OUght, AA as in fAther, AW as in cOW, AY as in mY, AE as in At, EH as in mEt, EY as in mAte, IY as in mEET, IH as in It, AH as in HUt, UW as in fOOD, ER as in hER, UH as in hOOD	38 articulations are repeated 10 times each (5 + 5 in two different sessions) in an audible manner AND 38 articulations are repeated 10 times each (5 + 5 in two different sessions) in a silent manner.	1.5 seconds
WORDS	Eager, lift, eight, edge, cap, matted, tub, box, rune, rook, folder, block, fun, mop, pod, very, went, throat, this, tango, doubt, not, pretty, xerox, rodent, limb, batch, jeep, ship, beige, yes, echo, gold, sing, Uh-oh, hiccup	36 articulations are repeated 10 times each (5 + 5 in two different sessions) in an audible manner AND 38 articulations are repeated 10 times each (5 + 5 in two different sessions) in a silent manner.	1.5 seconds
Rainbow Passage	Entire passage is enunciated naturally in voiced and unvoiced manners.	Passage is repeated 2 times in an audible manner AND The passage is repeated 2 times in a silent manner.	Each sentence is articulated in a time window of 5 seconds.

Table 7: Orofacial and articulatory cues used in the first part of the experiment.

	Gestures	Number of gesture repetitions	Articulation duration
NATO phonetic alphabets	Alfa, Bravo, Charlie, Delta, Echo, Foxtrot, Golf, Hotel, India, Juliette, Kilo, Lima, Mike, November, Oscar, Papa, Quebec, Romeo, Sierra, Tango, Uniform, Victor, Whiskey, X-ray, Yankee, Zulu	26 words are repeated 20 times each ($5 + 5 + 5 + 5$ in four different sessions) in a silent manner.	1.5 seconds
Grandfather passage	Entire passage is enunciated character-by-character using NATO phonetic alphabets (silent manner).	The passage is enunciated once.	Each character is enunciated in a time window of 1.5 seconds. A word is articulated continuously character-by-character followed by a break after each word.
Rainbow passage	Entire passage is enunciated character-by-character using NATO phonetic alphabets (silent manner).	The passage is enunciated once.	Each character is enunciated in a time window of 1.5 seconds. A word is articulated continuously character-by-character followed by a break after each word.
Grandfather passage	Entire passage is enunciated naturally in a silent manner.	The passage is enunciated 6 times.	Each sentence is enunciated in a time window of 4 seconds

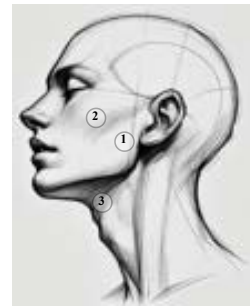
Table 8: Articulatory cues used in the second part of the experiment.



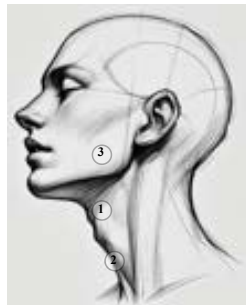
Subject 1



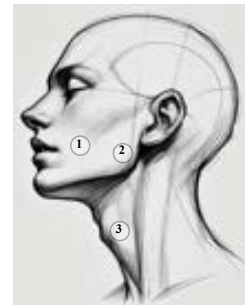
Subject 2



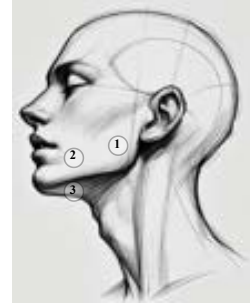
Subject 3



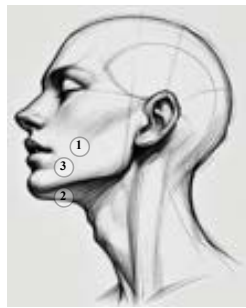
Subject 4



Subject 5



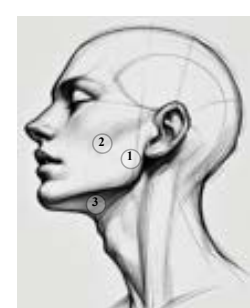
Subject 6



Subject 7



Subject 8



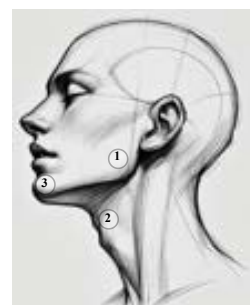
Subject 9



Subject 10



Subject 11



Subject 12

Figure 17: Three most important sEMG electrode locations for each subject.

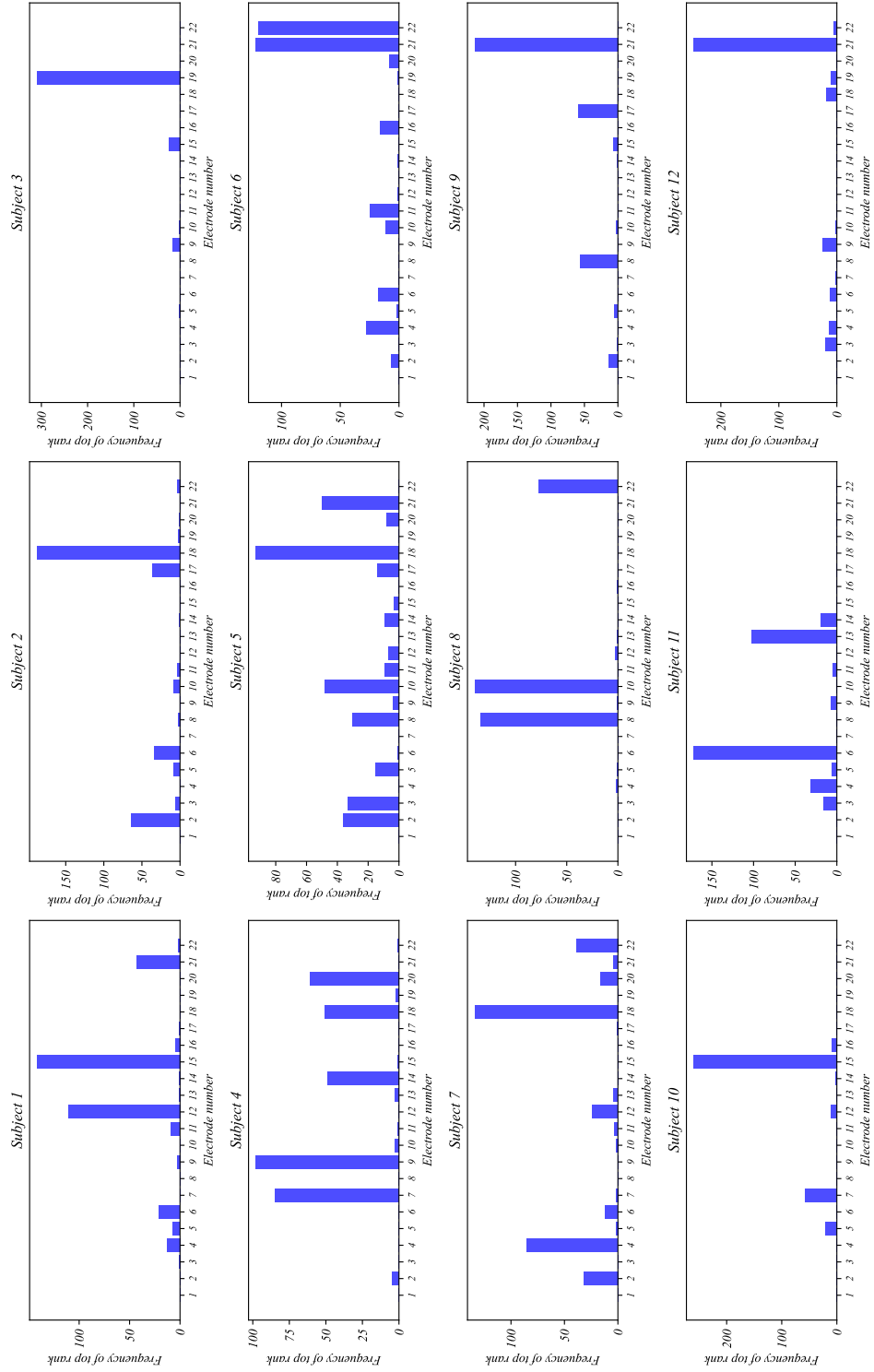


Figure 18: Frequency of top-rank for each of the 22 electrodes across 360 trials.