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

Harshavardhana T. Gowda [1]   Zachary D. McNaughton [2]   Lee M. Miller [2 3 4]

Abstract

Objective. In this article, we present data and methods for decoding speech articulations using surface electromyogram (EMG) signals. EMG-based speech neuroprostheses are useful for restoring audible speech in individuals who have lost the ability to speak intelligibly due to laryngectomy, neuromuscular diseases, stroke, or trauma-induced damage (e.g., radiotherapy toxicity) to speech articulators. *Approach.* To achieve this, we collect EMG signals from the face, jaw, and neck as subjects perform orofacial movements or articulate speech. These signals are recorded using surface electrodes placed at key muscle sites involved in speech production. Furthermore, we design and evaluate models that efficiently decode EMG with limited data. *Main results.* Our findings reveal that the manifold of symmetric positive definite (SPD) matrices serves as a natural embedding space for EMG signals. Specifically, we provide an algebraic interpretation of the manifold-valued EMG data using linear transformations and analyze and quantify EMG signal distribution shifts across individuals. *Significance.* Overall, our approach demonstrates significant potential for developing parameter-efficient neural networks that can be trained with minimal data. This is particularly important for embedded sensor devices with constrained computational resources. Additionally, given the challenges associated with collecting EMG data at scale, our method potentially offers a practical solution for advancing noninvasive speech neuroprostheses.

[1]Department of Electrical and Computer Engineering, University of California, Davis [2]Center for Brain and Mind, University of California, Davis [3]Department of Neurobiology, Physiology, and Behavior, University of California, Davis [4]Department of Otolaryngology/Head and Neck Surgery, University of California, Davis. Correspondence to: Harshavardhana T. Gowda <tgharshavardhana@gmail.com>.

Data and code availability

1. Introduction

Electromyogram (EMG) signals gathered 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. EMG based speech neuroprostheses 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 present methods for decoding speech articulation using EMG, focusing on the geometric structure of the data. We demonstrate that EMG signals are naturally embedded on a manifold of SPD matrices. Furthermore, we provide an algebraic interpretation of manifold-valued EMG data and show that EMG signals can be interpreted as a linear transformation of a Euclidean space of dimension $|\mathcal{V}|$, where $|\mathcal{V}|$ denotes the number of EMG sensor nodes.

We show that all articulations from a given individual result in similar transformations of this space, such that these transformations can be equivalently described using a common approximate eigenbasis matrix. Notably, different individuals have distinct eigenbases. Thus, the distribution shift of EMG signals across individuals can be characterized as a change of basis.

Additionally, we open-source EMG data from 16 subjects performing various orofacial gestures and speech articulations, making it the largest publicly available speech EMG dataset to date.

1.1. Prior work

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

The current benchmark in silent speech interfaces is established by Gaddy & Klein; Gaddy & Klein. Using electromyogram (EMG) signals collected during *silently* articulated speech (E_S) and *audibly* articulated speech (E_A), along with corresponding audio signals (A), they develop a recurrent neural transduction model to map time-aligned features of E_A or E_S with A. In their baseline model, joint representations between E_A and A are learned during training, and the model is tested on E_S. To improve performance, a refined model aligns E_S with E_A and subsequently uses the aligned features to learn joint representations with A. The methods described above have significant shortcomings that limit their practicality for real-world deployment. They are, 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 (both E_A and E_S), and ③ the requirement of time-aligned features, which are computationally expensive and time-consuming to obtain, making near real-time implementation challenging.

Unlike the works of Gaddy & Klein; Gaddy & Klein, the goal of EMG-based neuroprostheses is to achieve EMG-to-language translation using only E_S, without requiring corresponding E_A or A, mimicking standard speech-to-text decoding (Hori et al., 2017). For example, we should be able to decode a silently articulated sentence, <IT'S KIND OF FUN>, into its corresponding phonemic sequence, <IH-T-S SPACE K-AY-N-D SPACE AH-V SPACE F-AH-N>, using only E_S. This task is particularly challenging due to the complex multivariate nature of E_S and the absence of timestamps, as the sentence is articulated silently, making it unclear when a particular word is produced. To accomplish this, we need robust representations of EMG signals. This article describes foundational methods that enable such an EMG-to-language translation. Using the methods described in this article, we demonstrate for the first time that EMG-to-language translation is possible using only E_S, as shown in Gowda et al.

Finally, none of the previous works have explored whether EMG signals are capable of distinguishing different orofacial movements underlying speech articulation (such as various positions of the tongue and jaw) or whether EMG signals can span the entire English language phonemic space. This article provides data and methods to investigate these fundamental open questions. In the following sections, we detail methods for obtaining *informative and discriminative representations* of EMG signals and demonstrate that EMG embeddings on the manifold of SPD matrices are meaningful. We show that different orofacial movements, phoneme articulations, and word articulations exhibit structured representations that enable straightforward decoding. Additionally, we analyze and quantify EMG signal distribution shifts across individuals.

2. Methods and materials

A total of 16 subjects participated in our study. Please refer to the *ethical statement* for the subject selection criteria. Below, we provide a detailed explanation of the data acquisition protocols and EMG processing methods. The same 12 subjects participated in the experiments described in *sections* 2.1.1, 2.1.2, and 2.1.3.

2.1. EMG data acquisition setup

We collect EMG signals from twenty-two muscle 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 figures 1 and 2 for electrode placement. Besides 22 data electrodes, we also have a GROUND electrode and a REFERENCE electrode. GROUND electrode is placed on the left ear lobe and the REFERENCE electrode is placed on the right ear lobe. Electrode positions and labels are identical for all participants.

Before signal acquisition, participants were briefed on the experimental protocol and seated comfortably in a chair. They were instructed to perform the articulations naturally, as they would in everyday speech. For *silent* articulations, participants were asked to speak naturally but inaudibly. Each articulation (or orofacial gesture) was performed within a 1.5-second window, during which the corresponding text was displayed on a graphical user interface (GUI). In the following sections, we provide a detailed description of the various parts of the experiment.

2.1.1. EMG OF DIFFERENT OROFACIAL MOVEMENTS

The aim of this experiment is to verify if different orofacial movements underlying speech articulation can be distinguished using EMG. 12 subjects performed 13 different orofacial movements, each repeated 10 times. The orofa-
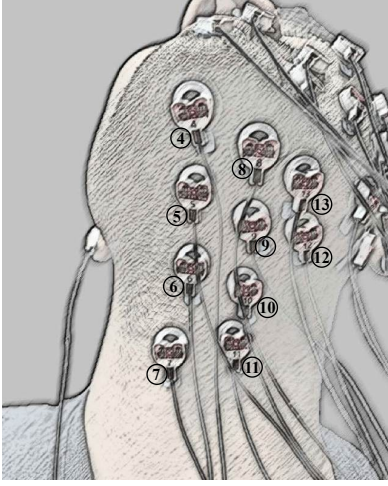
Figure 1. Placement of electrodes on the neck region.

Figure 2. Placement of electrodes on cheek and lip regions. Electrode *1* is above the upper lip and electrode *3* is below the lower lip.

cial movements are: <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>.

These movements are chosen because they encompass a broad range of muscle activations involved in natural speech production, spanning articulation mechanisms such as lip rounding, jaw positioning, and tongue placement, which are crucial for forming different phonemes. These movements correspond to critical articulatory gestures involved in both *audible* and *silent* speech. Understanding their distinct EMG signatures will help determine whether EMG-based decoding can differentiate between phonemic elements, even in the absence of acoustic output.

2.1.2. EMG OF INDIVIDUAL PHONEME ARTICULATIONS

The aim of this experiment is to verify if all English language phonemes can be decoded using EMG when articulated in both *audible* and *silent* manners. 12 subjects articulated 38 different phonemes, with each phoneme repeated 10 times in both *audible* and *silent* manners. Phonemes are broadly classified as consonants or vowels. Based on the placement of articulation, consonants are further classified into bilabial, labiodental, dental, alvelor, post vaelor, and approximant consonants. We list all the phonemes below.

Bilabial consonants: <BAA>, <PAA>, <MAA>; *Labiodental consonants*: <FAA>, <VAA>; *Dental consonants*: <THAA>, <DHAA>; *Alvelor consonants*: <TAA>, <DAA>, <NAA>, <SAA>, <ZAA>; *Post vaelor 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. In total, we have 23 consonants and 15 vowels.

Verifying that EMG signals can span the entire English language phonemic space is important because we aim to build a system in which naturally (but inaudibly) articulated speech is decoded using E_S into a corresponding phonemic sequence, similar to standard audio-to-text translation. We demonstrate such a system in Gowda et al., and this work lays the foundation for it.

2.1.3. EMG OF INDIVIDUAL WORD ARTICULATIONS

The aim of this experiment is to verify if we can decode individual words spanning the entire English language phonemic space when words are articulated in both *audible* and *silent* manners. 12 subjects articulated 36 different words that span the entire English language phonemic space, with each word repeated 10 times in both *audible* and *silent* manners. The 36 words are: <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>.

Verifying that we can decode words is important because words are composed of a sequence of phonemes (for example, the word <EAGER> can be decomposed into the phoneme sequence <IY-G-ER>) and decoding them with high fidelity (e.g., using a recurrent network) demonstrates

the potential for decoding continuously articulated sentences.

2.1.4. GENERALIZABLE LANGUAGE-SPELLING MODEL USING NATO PHONETIC CODES

In this experiment, 4 subjects articulate English language sentences using NATO phonetic codes in a *silent* manner. For instance, the word <RAINBOW> is articulated as <ROMEO-ALFA-INDIA-NOVEMBER-BRAVO-OSCAR-WHISKEY>. Although this approach does not replicate natural speech, it provides a practical mode of limited communication for individuals who have lost speech articulation capabilities. The experiment is divided into two stages.

In the first stage, subjects silently articulate each of the 26 NATO phonetic codes, repeating each code 20 times. In the second stage, subjects articulate the phonemically balanced RAINBOW (Fairbanks, 1960) and GRANDFATHER (Reilly & Fisher, 2012) passages in a spelled-out manner using NATO phonetic codes. Together, both passages contain 1968 NATO code articulations. Data from the first stage is used for training and validating the decoding models, while data from the second stage is used for testing.

During the second stage, sentences from the passages are displayed word by word, and each character in a word is articulated using the corresponding NATO phonetic code within a 1.5-second articulation window.

2.2. EMG data preprocessing

The data collection environment was carefully controlled to eliminate AC electrical interference. EMG signals undergo minimal preprocessing. Signal from the REFERENCE channel (the electrode located on the right ear lobe) is subtracted from all other EMG data channels. Then, they are bandpass filtered using a third-order Butterworth bandpass filter between 80 and 1000 Hertz and z-normalized along the time dimension for each channel. The preprocessed EMG signals are then used to construct a fully connected sensor graph, whose edge matrices are analyzed on the manifold of SPD matrices. We detail the EMG analysis methods in the following section.

2.3. EMG data analysis

EMG signals are collected by a set of sensors $\mathcal{V}$ and are functions of time t. A sequence of EMG signals E_S (or E_A) corresponding to silently (or audibly) articulated speech, is represented as $E_{S/A} = \mathbf{f}_v(t)$ for all $v \in \mathcal{V}$. Here, $\mathbf{f}_v(t)$ denotes the EMG signal captured at a sensor node v as a function of time t.

We construct a complete graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}(\tau))$ to represent the functional connectivity of EMG signals, where

$\mathcal{E}(\tau)$ denotes the set of edges over a time window $\tau = [t_{\text{START}}, t_{\text{END}}]$. The edge weight between two nodes v_1 and $v_2 \in \mathcal{V}$ in a time window is defined as $e_{12} = e_{21} = \mathbf{f}_{v_1}^T \mathbf{f}_{v_2}$, which corresponds to the covariance of the signals at those nodes during the time interval. Consequently, the edge (adjacency) matrix $\mathcal{E}(\tau)$ is symmetric and positive semi-definite. We convert semi-definite adjacency matrices into definite ones by computing $\mathcal{E} \leftarrow (1 - \eta)\mathcal{E} + \eta \text{trace}(\mathcal{E})\mathcal{I}$. Here, $\mathcal{I}$ is an identity matrix whose dimensions are the same as $\mathcal{E}$ and we empirically found that $\eta = 0.1$ suffices for all our data. Directly working with SPD matrices using affine-invariant or log-Euclidean metrics (Arsigny et al., 2007) involves computationally expensive operations, such as matrix exponential and matrix logarithm calculations. These operations make mappings between the manifold space and the tangent space, and vice versa, computationally intensive. To address this, Lin proposed methods to operate on SPD matrices using Cholesky decomposition. They established a diffeomorphism between the Riemannian manifold of SPD matrices and Cholesky space. In Cholesky space, the computational burden is significantly reduced: logarithmic and exponential computations are restricted to the diagonal elements of the matrix, making them element-wise operations. Additionally, the Fréchet mean (centroid) is derived in a closed form. For an SPD edge matrix $\mathcal{E}$, the corresponding Cholesky decomposition $\mathcal{L} = \text{CHOLESKY}(\mathcal{E})$ is such that $\mathcal{E} = \mathcal{L}\mathcal{L}^T$. A matrix $\lfloor \mathcal{L} \rfloor$ is the strictly lower triangular part of the matrix $\mathcal{L}$, and a matrix $\mathbb{D}(\mathcal{L})$ is the diagonal part of the matrix $\mathcal{L}$. In the following section, we explain in detail, the methods used to analyze SPD matrices.

2.3.1. DISTANCE BETWEEN TWO SPD MATRICES $\mathcal{E}_1$ AND $\mathcal{E}_2$

The geodesic distance between two SPD matrices $\mathcal{E}_1$ and $\mathcal{E}_2$ is same as the distance between the corresponding Cholesky matrices $\mathcal{L}_1$ and $\mathcal{L}_2$ and is calculated as

$$
\begin{aligned}
d(\mathcal{L}_1, \mathcal{L}_2) = \big\{ &||\lfloor \mathcal{L}_1 \rfloor - \lfloor \mathcal{L}_2 \rfloor||_F^2 \\
& + ||\log \mathbb{D}(\mathcal{L}_1) - \log \mathbb{D}(\mathcal{L}_2)||_F^2 \big\}^{1/2}, \quad (1)
\end{aligned}
$$

where $||.||_F$ denotes the Frobenius norm.

2.3.2. FRÉCHET MEAN (CENTROID) OF SPD MATRICES

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

$$\mathcal{F}_{\text{CHOLESKY}} = \frac{1}{n} \sum_{i=1}^{n} \lfloor \mathcal{L}_i \rfloor \quad +$$
$$\exp \left(\frac{1}{n} \sum_{i=1}^{n} \log(\mathbb{D}(\mathcal{L}_i)) \right). \quad (2)$$

2.3.3. k-MEDOIDS CLUSTERING ALGORITHM

Given a set of (n) SPD edge matrices, $\mathcal{E}$, from different orofacial movements, if the edge matrices $\mathcal{E}$ corresponding to the same movement cluster together while those from different movements remain separate, we can conclude that treating EMG signals as graphs and representing covariance edge matrices $\mathcal{E}$ on the manifold of SPD matrices provides a robust signal representation. This would suggest that such an embedding space is meaningful for naturally distinguishing different orofacial movements. In the *results* section, we test whether this is the case for different orofacial movements described in *section* 2.1.1.

To identify such natural clusters in an unsupervised manner, we implement the classic k-medoids algorithm (Kaufman & Rousseeuw, 1990) using partitioning around medoids (PAM) heuristic by replacing the Euclidean distance with the geodesic distance in equation 1.

Whenever we use k-medoids clustering algorithm, edge matrices $\mathcal{E}(\tau)$ are created using the entire articulation duration of 1.5 seconds.

2.3.4. MINIMUM DISTANCE TO MEAN ALGORITHM (MDM)

Given a set of (n) SPD edge matrices, $\mathcal{E}$, from different speech articulations, these matrices may exhibit a structured representation on the manifold of SPD matrices. However, such a structure may not be sufficiently distinct for unsupervised algorithms like k-medoids clustering to effectively discern. In such cases, supervised algorithms like MDM may be more useful.

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

$$\mathcal{C}^j = \mathbb{E}(\{\text{CHOLESKY}(\mathcal{E}^j)\}), \quad (3)$$

where the Fréchet mean $\mathbb{E}$ is calculated according to equation 2. Given a test dataset of SPD matrices $\{\mathcal{T}\}$, $T \in \mathcal{T}$ is assigned to that class whose centroid is nearest to CHOLESKY(T). That is, the class of T is

$$\arg \min_j d(\text{CHOLESKY}(T), \mathcal{C}^j), \quad (4)$$

where $d(.)$ is calculated according to equation 1.

Whenever we use MDM, edge matrices $\mathcal{E}(\tau)$ are created using the entire articulation duration of 1.5 seconds.

In all cases, when using MDM, we apply the following *train-test* data split: out of the 10 total articulation instances of each word or phoneme, 6 are used for training the model, and 4 are used for testing. The articulations in the test set are not present in the training set.

2.3.5. t-SNE FOR EMG DATA VISUALIZATION

We use the t-SNE (t-Stochastic Neighbor Embedding) algorithm, adapted from Van der Maaten & Hinton, for data visualization. Unlike standard t-SNE, which takes vectors as input and uses Euclidean distance, we input edge matrices $\mathcal{E}$ and employ the distance defined in equation 1.

2.3.6. A NEURAL NETWORK FOR SPD MATRIX LEARNING

We use k-medoids (defined in *section* 2.3.3) and MDM (defined in *section* 2.3.4) to test whether EMG signals, when interpreted as a graph arising from the underlying functional connectivity of the neuromuscular system, give rise to edge matrices, $\mathcal{E}$, that are naturally distinguishable on the embedding manifold of SPD matrices (see the *results* section). However, to learn more expressive representations of the edge matrices $\mathcal{E}$, we use a neural network as defined in Huang & Van Gool. This neural network takes SPD matrices $\mathcal{E}$ as input and outputs a classification of the matrices. We describe this neural network in detail in appendix A.1. Succinctly, the neural network learns a new matrix, $\mathcal{E}^{(1)}$, from $\mathcal{E}^{(0)}$ via the transformation $\mathcal{E}^{(1)} = W^T \mathcal{E}^{(0)} W$, where W is a full-rank, semi-orthogonal matrix learned by the neural network (when $\mathcal{E}^{(0)}$ is an SPD matrix, W is an orthogonal matrix such that $W^T = W^{-1}$).

Given a set of SPD edge matrices, $\{\mathcal{E}^{(0)}\}$ (corresponding, for instance, to different speech articulations), if the neural network learns a transformation W such that the transformed set of matrices, $\{\mathcal{E}^{(1)}\}$, are all approximately diagonal, then we can conclude that all SPD matrices in $\{\mathcal{E}^{(0)}\}$ result in a similar linear transformation of the space $\mathbb{R}^{|\mathcal{V}|}$. That is, this transformation can be equivalently expressed in the corresponding approximate eigenbasis W such that any $\mathcal{E}^{(0)}$ can be expressed as

$$W \mathcal{E}^{(1)} W^T,$$

where $\mathcal{E}^{(1)}$ is approximately diagonal. Therefore, this decomposition can be interpreted as an approximate eigendecomposition, with W representing the approximate eigenbasis vectors and $\mathcal{E}^{(1)}$ representing the corresponding approximate eigenvalues.

In the *results* section, we test whether this holds true. If so, all articulations from a given individual would share

an approximate common eigenbasis. Furthermore, if such an eigenbasis is distinct from those of other individuals, it would suggest that the shift in the EMG signal distribution across individuals arises from a transformation in the underlying *basis*.

Whenever we use this method, edge matrices $\mathcal{E}(\tau)$ are created using the entire articulation duration of 1.5 seconds.

For phoneme articulations (as described in *section* 2.1.2) and word articulations (as described in *section* 2.1.3), when using this method, we apply the following *train-test* data split: out of the 10 total articulation instances of each word or phoneme, 6 are used for training the model, and 4 are used for testing. The articulations in the test set are not present in the training set.

For the generalizable language-spelling articulations using NATO phonetic codes (as described in *section* 2.1.4), we apply the following *train-validation-test* data split when using this method.

- *Train set:* A subset of the articulations from the first stage of the experiment, where each of the 26 NATO phonetic codes was articulated 20 times. We use 16 repetitions of each of the 26 codes, resulting in a total of 416 training instances.

- *Validation set:* A subset of the articulations from the first stage of the experiment, where each of the 26 NATO phonetic codes was articulated 20 times. We use 4 repetitions (not included in the train set) of each of the 26 codes, resulting in a total of 104 validation instances.

- *Test set:* The second stage of the experiment, which consists of 1968 NATO code articulations from both the RAINBOW and GRANDFATHER passages.

The model is trained on the train set, and the weights that achieve the best accuracy on the validation set are used for testing on the test set.

2.3.7. A RECURRENT NEURAL NETWORK FOR SPD MATRIX LEARNING

The ultimate goal is to translate continuously articulated speech in a *silent* manner by afflicted individuals (such as those who have undergone a laryngectomy) into *audible* speech using EMG signals (E_S). This is very similar to audio-to-text translation, with a *key* difference: while audio is a univariate signal that can be modeled as the application of a time-varying filter to a time-varying source signal, EMG signals are multivariate and result from a purely additive combination of muscle activation patterns. Here, we test whether EMG signals contain informative features at a fine-scale resolution of 30 *ms* by training a GRU to classify

various phonemes (as described in *section* 2.1.2), words (as described in *section* 2.1.3), and a generalizable language-spelling model using NATO phonetic words (as described in *section* 2.1.4). This serves as a foundational verification paradigm for the full-fledged EMG-to-language translation framework, which we describe in Gowda et al.

Standard speech-to-text models (Hori et al., 2017) use recurrent neural networks that process vectorized representations of audio signals. Here, we adapt a recurrent neural network to accept SPD matrices $\mathcal{E}$ as inputs instead of vectors. Specifically, we use the gated recurrent unit (GRU) model described in Jeong et al. This method modifies the standard Euclidean GRU to accept SPD matrices as input by incorporating techniques from Chen et al. and Lou et al.

We provide a detailed description of the recurrent model in appendix A.2.

Whenever we use this method, edge matrices $\mathcal{E}(\tau)$ are generated by slicing the EMG signals using a sliding window with a context size of 150 *ms* and a step size of 30 *ms*. For an articulation lasting 1.5 seconds, this results in a sequence of 46 edge matrices, which are treated as 46 time steps in the input sequence to the GRU.

The *train-validation-test* split is same as described in *section* 2.3.6.

3. Results

Here, we present results demonstrating that EMG signals exhibit structured representations on the manifold of SPD matrices. We show that embeddings of different orofacial movements, words, and phonemes on this manifold are meaningful, such that EMG signals corresponding to different articulations naturally form distinct clusters. Furthermore, we apply the methods described in *section* 2.3.6 and *section* 2.3.7 to the generalizable language-spelling corpus using NATO phonetic codes (see *section* 2.1.4) and demonstrate that high-fidelity decoding of speech articulation is achievable using EMG data.

3.1. EMG shows structured representation on the manifold of SPD matrices

When EMG signals are interpreted as graphs defined by the underlying functional connectivity of the orofacial neuromuscular system, the corresponding graph edge matrices exhibit structured representations on the manifold of SPD matrices. Orofacial movements underlying critical articulatory gestures involved in both *audible* and *silent* speech articulations demonstrate highly structured clustering, which is discernible even using unsupervised algorithms such as *k*-medoids clustering (*section* 2.3.3). While complex articulations, such as phonemes and words, are more challenging

to distinguish using unsupervised methods, they still exhibit meaningful representations and achieve high decoding accuracy using MDM (*section* 2.3.4). Decoding accuracy using both *k*-medoids and MDM is significantly higher than chance levels, as shown in the following sections.

3.1.1. DECODING OROFACIAL GESTURES

We decode 13 orofacial gestures, as described in *section* 2.1.1, using the *k*-medoids algorithm (*section* 2.3.3) and present the average decoding accuracy across all 12 subjects in table 1. A detailed subject-wise decoding accuracy is provided in appendix B.1.

Average decoding accuracy of orofacial gestures using *k*-medoids algorithm
0.736

Table 1. Different orofacial gestures are naturally distinguishable on the manifold of SPD matrices. Classification accuracy of 13 orofacial movements described in *section* 2.1.1. Chance accuracy is merely $\frac{1}{13} = 0.077$.

3.1.2. VISUALIZING OROFACIAL GESTURES USING *t*-SNE

Here, we visualize SPD edge matrices corresponding to different orofacial gestures using *t*-SNE (see *section* 2.3.5). As shown in figure 3, different gestures are naturally separated on the manifold.
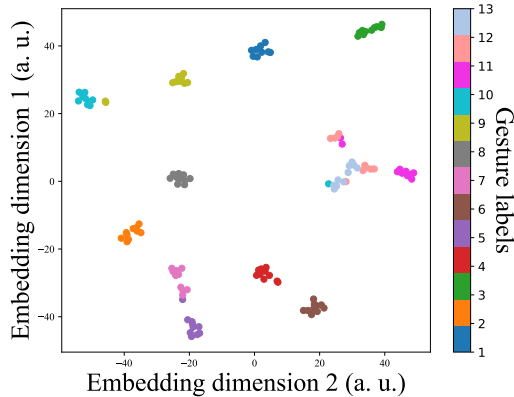

Figure 3. Different orofacial gestures are naturally distinguishable on the manifold of SPD matrices. *t*-SNE of edge matrices of various orofacial movements described in *section* 2.1.1 for subject *1*. Embedding is colored according to gestures (*a.u.* - arbitrary units).

3.1.3. DECODING INDIVIDUAL PHONEME AND WORD ARTICULATIONS

Individual phoneme articulations (as described in *section* 2.1.2) and individual word articulations (as described in *section* 2.1.3) exhibit a structured representation that is insufficient for *k*-medoids clustering but distinguishable using MDM. This is because speech articulations are more complex compared to orofacial movements. We report the average decoding accuracy using MDM across all subjects for both *silent* and *audible* articulations in tables 2 and 3, respectively. A detailed subject-wise decoding accuracy is provided in appendix B.2.1.

Furthermore, we demonstrate that phonemes and words can be classified with higher accuracy using neural networks described in *sections* 2.3.6 and 2.3.7, with results presented in appendices B.2.2 and B.2.3, respectively.

Train-test split for MDM is outlined in *section* 2.3.4.

Decoding accuracy of *audible* phonemes			Decoding accuracy of *silent* phonemes		
A	C	V	A	C	V
0.36	**0.36**	**0.45**	**0.36**	**0.34**	**0.49**

Table 2. Phoneme articulations are naturally distinguishable on the manifold of SPD matrices. Classification accuracy of phoneme articulations using MDM. It is a 38-way classification problem with chance accuracy of $\frac{1}{38} = 0.026$. We also show classification accuracy for 23 consonant phonemes and 15 vowel phonemes separately. *A*-all phonemes, *C*-consonants only, *V*-vowels only.

Decoding accuracy of *audible* words	Decoding accuracy of *silent* words
0.544	**0.439**

Table 3. Word articulations are naturally distinguishable on the manifold of SPD matrices. Classification accuracy of word articulations using MDM. It is a 36-way classification problem with chance accuracy of $\frac{1}{36} = 0.028$.

3.2. Generalizable language-spelling paradigm

Now, we present the results for the generalizable language-spelling paradigm described in *section* 2.1.4. Here, we train the neural networks described in *sections* 2.3.6 and 2.3.7 and report the results. Notably, our *training* set consists of only about 10 minutes of data, yet we test on a much larger test set and still achieve decoding accuracies significantly higher than chance level.

We train the models on 416 NATO phonetic code articulations, validate them on 104 NATO phonetic code articulations, and test them on 1968 articulations from the RAINBOW and GRANDFATHER passages. The top-*5* decoding accuracies are presented in table 4. As observed, these

accuracies are significantly higher than the chance-level top-*5* decoding accuracy. The recurrent network described in *section* 2.3.7 outperforms the network in *section* 2.3.6. This improved performance can be attributed to the recurrent network's ability to model articulations as sequences. A detailed subject-wise decoding accuracy is provided in appendix B.3.

Decoding accuracy using model in *section* 2.3.6	Decoding accuracy using model in *section* 2.3.7
0.631	**0.773**

Table 4. Top-*5* decoding accuracy of spelled-out RAINBOW and GRANDFATHER passage articulations using NATO phonetic codes. It is a 26-way classification problem with chance top-*5* accuracy of 0.178.

3.3. EMG signal distribution shift across individuals

EMG signals are significantly affected by distributional shifts across individuals due to inter-subject differences in neural drive and muscle properties (Farina et al., 2014), (Farina et al., 2004). Factors such as subcutaneous fat thickness, the spatial distribution of muscle fibers, variations in muscle fiber conduction velocity (Farina et al., 2014), and contextual elements like electrode placement (which we have ensured is consistent across all subjects to the extent allowed by anatomical and physiological constraints) contribute to this variability. Additionally, neural properties, such as the discharge characteristics of the neural drive, further amplify inter-individual differences in EMG signals (Farina et al., 2014).

Here, we demonstrate that the aforementioned factors lead to covariate signal distribution shifts across individuals, which can be quantified as a change of basis. Specifically, any edge matrix $\mathcal{E}$ (corresponding to any articulation) from a given individual can be expressed using a common approximate eigenbasis. That is, $\mathcal{E}$ can be decomposed as $\mathcal{E} = W\sigma W^T$, where σ is approximately diagonal.

We validate this by training an SPD matrix learning neural network separately for each individual (see *section* 2.3.6 and appendix A.1). For instance, we train the neural network to classify different words articulated in an *audible* manner (*section* 2.1.3: 36 different words spanning the entire phonemic space of the English language, with each word articulated 10 times) and analyze the learned weights that yield the best accuracy on the *test* set. The input to the neural network consists of edge matrices corresponding to different word articulations, denoted as $\mathcal{E}^{(0)}$. In the first layer, the network learns a transformation matrix $W^{(1)}$ such that

$$\mathcal{E}^{(1)} = W^{(1)^T}\mathcal{E}^{(0)}W^{(1)}$$

is approximately diagonal for all $\mathcal{E}^{(0)}$. Consequently, any $\mathcal{E}^{(0)}$ can be decomposed as

$$\mathcal{E}^{(0)} = W^{(1)}\mathcal{E}^{(1)}W^{(1)^T},$$

where $\sigma = \mathcal{E}^{(1)}$ is approximately diagonal. This demonstrates that all SPD matrices can be represented using an approximate common eigenbasis. However, such an approximate eigenbasis varies across individuals. In figure 4, we show that matrices $\mathcal{E}^{(1)}$ are more diagonal compared to $\mathcal{E}^{(0)}$ and in figure 5, we show that the transformation matrix $W^{(1)}$ learned by the neural network is different across subjects.
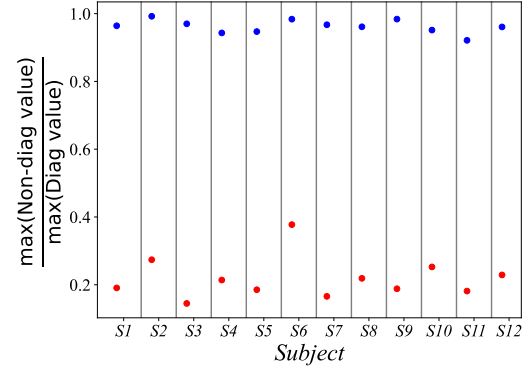
Figure 4. All edge matrices within an individual can be approximately diagonalized. Blue: average value of $\frac{\max(\text{ABS}((\text{NON DIAG}(\mathcal{E}^{(0)}))}{\max(\text{DIAG}(\mathcal{E}^{(0)}))}$ for all word articulations. Red: average value of $\frac{\max(\text{ABS}((\text{NON DIAG}(\mathcal{E}^{(1)}))}{\max(\text{DIAG}(\mathcal{E}^{(1)}))}$ for all word articulations. As we can see, $\mathcal{E}^{(1)}$ are approximately diagonal compared to $\mathcal{E}^{(0)}$.

3.3.1. VISUALIZING EMG DATA DISTRIBUTION SHIFT USING *t*-SNE

Here, we visualize SPD edge matrices corresponding to words articulated in an *audible* manner (*section* 2.1.3) across different subjects using *t*-SNE (*section* 2.3.5). We observe that, instead of similar words clustering together - for example, the word <EAGER> from different subjects does not form a distinct cluster - instead, each subject's embeddings exhibit a stronger within-subject similarity. Specifically, the embedding of the word <EAGER> from subject *A* is more similar to other words articulated by subject *A* (e.g., <LIFT> or <EIGHT>) than to the same word <EAGER> articulated by any other subject.

We discuss the implications of this observation on the generalization capability of neural networks in the *discussion* section. Furthermore, in appendix C, we demonstrate that this effect persists even for naturally articulated continuous speech when analyzed at a fine phonemic temporal resolution of 100 *ms*. This finding has significant implications for the architectural design of EMG-to-language transla-
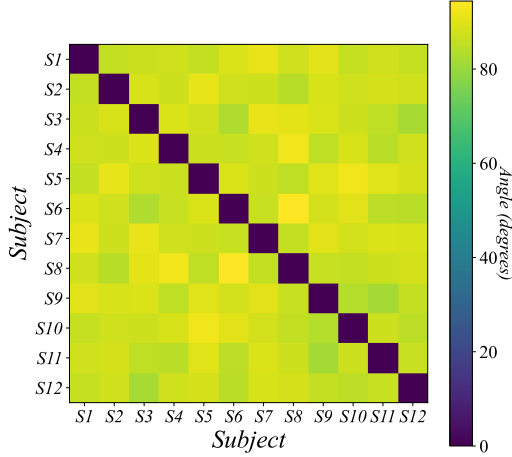
Figure 5. EMG embeddings of speech articulations on the manifold differ substantially across individuals. Approximate eigenbasis vectors ($Q = W^{(1)}$) are different for different individuals.
$$\theta = \cos^{-1}\left(\frac{\texttt{trace}(Q_i Q_j^T)}{\sqrt{\texttt{trace}(Q_i Q_i^T)}\sqrt{\texttt{trace}(Q_j Q_j^T)}}\right)$$
between approximate eigenbasis matrices Q_i and Q_j of different individuals i and j.

tion models, which we further elucidate in the *discussion* section.

In figure 6, we show the *t*-SNE embedding of *audible* word articulations across all 12 subjects. As observed, the embeddings cluster together according to subjects rather than words.
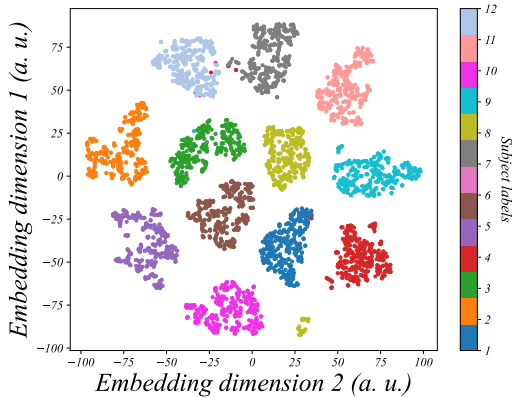

Figure 6. Articulations from different subjects cluster separately on the manifold of SPD matrices. *t*-SNE of edge matrices of word articulations. Embedding is colored according to subjects (*a.u.* - arbitrary units).

4. Observations and discussions

① We show that EMG signals, when interpreted as a graph arising from the underlying neuromuscular system, give rise to edge matrices that exhibit a structured representation on the manifold of SPD matrices. We demonstrate this structured representation for orofacial movements underlying speech articulations, as well as for individual word and phoneme articulations (using *k*-medoids and MDM algorithms).

Given that EMG signals exhibit such a structured representation on the manifold, we analyze classification accuracy using more expressive manifold neural networks, including recurrent networks, as described in *sections* 2.3.6 and 2.3.7. Using these methods, we show that it is possible to develop limited communication speech neuroprostheses, where sentences are spelled out using NATO codes. These models are easy to train and require only a small amount of data, laying the foundation for a more generalizable EMG-to-language spelling model.

② Building on the methods discussed here, we demonstrate for the first time that EMG-to-language translation is possible at a phonemic level with a fine-scale resolution of 20 *ms*, akin to standard speech-to-text translation using only EMG signals collected during *silent* speech articulations, as shown in Gowda et al., highlighting the significance of this work.

③ While EMG-based translation operates on multichannel biosignals reflecting neuromuscular activity, speech processing relies on a fundamentally different signal modality - audio. Audio signals are univariate, with all articulatory content captured in a single channel. Therefore, they can be viewed as functions sampled on a one-dimensional temporal grid, making deep convolutional neural networks well-suited for learning audio features (Baevski et al., 2020; Hsu et al., 2021), as they capture hierarchical representations using varying receptive fields.

In contrast, EMG signals are multivariate, as they capture muscle activations across multiple sensor nodes. Unlike audio, which is well-represented as a one-dimensional sequence, EMG data is better interpreted as a transformation of the space $\mathbb{R}^{|\mathcal{V}|}$, where $|\mathcal{V}|$ denotes the number of sensors. This distinction arises because EMG captures spatially distributed bioelectrical activity rather than a single waveform. Consequently, instead of treating EMG as a simple sequence, it is more naturally analyzed using structured representations, such as those found on the manifold of SPD matrices.

We believe this fundamental difference between audio and EMG modalities has important implications for feature extraction and model design, necessitating approaches that go beyond standard convolutional architectures. While direct comparisons for articulatory gestures are lacking - since this work serves as the first such benchmark - previous research (Gowda & Miller) has demonstrated that decoding

hand gestures using EMG on the SPD manifold outperforms deep learning counterparts defined in the Euclidean domain. Furthermore, by employing the manifold methods described here, we demonstrate that EMG collected from silently articulated speech (E_S) can be translated at a phonemic level without relying on corresponding E_A (EMG collected during audible speech) or A (audio) in Gowda et al. This represents a significant improvement in the training paradigm compared to previous works by Gaddy & Klein; Gaddy & Klein which required both E_S and E_A for model training.

④ We analyzed the representations learned by the neural network model in *section* 2.3.6 to determine whether the model had learned meaningful representations with respect to the placement and manner of articulation. To this end, we examined phoneme errors to assess whether the model's learned representations correlated well with place and manner of articulation.

Based on placement and manner of articulation, consonant phonemes are classified into seven groups: bilabial, labiodental, dental, alveolar, post-velar, velar, and approximant consonants. During audible phoneme consonant classification, approximately 44 percent of the errors made by the model in *section* 2.3.6 involved misclassifications within the same phoneme group (for example, classifying the bilabial <PAA> as <BAA> or <MAA>, or misclassifying the velar <ZHAA> as <CHAA>, <SHAA>, or <JHAA>).

When such misclassifications are considered acceptable, the average consonant phoneme classification accuracy increases from 0.53 (*section* 2.1.2) to 0.74, representing a nearly 40 percent improvement. This increase is practically significant, as such systematic errors can be corrected using language models.

⑤ Given that the transformation of the space $\mathbb{R}^{|\mathcal{V}|}$ induced by the edge matrices of EMG from different individuals varies - which can be quantified as a change of basis transformation - and since $\mathbb{R}^{|\mathcal{V}|}$ can be represented by infinitely many such bases, zero-shot generalization, where a model trained on one group of individuals generalizes to unseen individuals, appears challenging. In future work, we would like to explore if we can design efficient zero-shot or few-shot learning strategies.

5. Conclusion

We present an efficient data representation for multivariate orofacial EMG signals and demonstrate that the manifold of SPD matrices provides a meaningful embedding space. Furthermore, we present methods to decode speech articulations using EMG. The open-sourced data and code presented here constitute the largest such dataset to date, establishing a strong foundation for further advancements in the field.

Several open questions remain to be addressed, including the generalizability of models across individuals - whether such models can be built at all, and if so, what the most efficient strategy for building them would be.

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

Arsigny, V., Fillard, P., Pennec, X., and Ayache, N. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29(1):328–347, 2007. doi: 10.1137/050637996. URL https://doi.org/10.1137/050637996.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Diener, L., Felsch, G., Angrick, M., and Schultz, T. Session-independent array-based emg-to-speech conversion using convolutional neural networks. In *Speech Communication; 13th ITG-Symposium*, pp. 1–5, 2018.

Fairbanks, G. *Voice and Articulation Drillbook*. Harper, 1960. ISBN 9780060419905. URL https://books.google.com/books?id=qN1ZAAAAMAAJ.

Farina, D., Merletti, R., and Enoka, R. M. The extraction of neural strategies from the surface emg. *Journal of applied physiology*, 96(4):1486–1495, 2004.

Farina, D., Merletti, R., and Enoka, R. M. The extraction of neural strategies from the surface emg: an update. *Journal of applied physiology*, 117(11):1215–1230, 2014.

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

Gaddy, D. and Klein, D. 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)*, pp. 175–181, 2021.

Gowda, H. T. and Miller, L. M. Topology of surface electromyogram signals: hand gesture decoding on riemannian manifolds. *Journal of Neural Engineering*.

Gowda, H. T., Rahimi, F., and Miller, L. M. Non-invasive electromyographic speech neuroprosthesis: a geometric perspective. *arXiv preprint arXiv:2502.05762*, 2025.

Hori, T., Watanabe, S., and Hershey, J. R. Joint ctc/attention decoding for end-to-end speech recognition. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 518–529, 2017.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

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

Janke, M. and Diener, L. 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.

Jeong, S., Ko, W., Mulyadi, A. W., and Suk, H.-I. Deep Efficient Continuous Manifold Learning for Time Series Modeling . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 46(01):171–184, January 2024. ISSN 1939-3539. doi: 10.1109/TPAMI.2023.3320125. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2023.3320125.

Jou, S.-C., Schultz, T., Walliczek, M., Kraft, F., and Waibel, A. Towards continuous speech recognition using surface electromyography. In *Ninth International Conference on Spoken Language Processing*, 2006.

Kapur, A., Sarawgi, U., Wadkins, E., Wu, M., Hollenstein, N., and Maes, P. Non-invasive silent speech recognition in multiple sclerosis with dysphonia. In *Machine Learning for Health Workshop*, pp. 25–38. PMLR, 2020.

Kaufman, L. and Rousseeuw, P. J. Partitioning around medoids (program pam). In *Wiley Series in Probability and Statistics*, pp. 68–125. John Wiley & Sons, Inc., Hoboken, NJ, USA, March 8 1990. ISBN 978-0-470-31680-1. doi: 10.1002/9780470316801.ch2. URL https://doi.org/10.1002/9780470316801.ch2. Retrieved 2021-06-13.

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

Lou, A., Lim, D., Katsman, I., Huang, L., Jiang, Q., Lim, S. N., and De Sa, C. M. Neural manifold ordinary differential equations. *Advances in Neural Information Processing Systems*, 33:17548–17558, 2020.

Meltzner, G. S., Heaton, J. T., Deng, Y., De Luca, G., Roy, S. H., and Kline, J. C. Development of semg sensors and algorithms for silent speech recognition. *Journal of neural engineering*, 15(4):046031, 2018.

Reilly, J. and Fisher, J. L. Sherlock holmes and the strange case of the missing attribution: A historical note on "the grandfather passage". *Journal of Speech, Language, and Hearing Research*, 55(1):84–88, Feb 2012. doi: 10.1044/1092-4388(2011/11-0158).

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

Toth, A. R., Wand, M., and Schultz, T. Synthesizing speech from electromyography using voice transformation techniques. In *Interspeech 2009*, pp. 652–655, 2009. doi: 10.21437/Interspeech.2009-229.

Van der Maaten, L. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

A. Neural networks for learning SPD matrices

Here, we describe in detail the neural networks presented in *sections* 2.3.6 and 2.3.7.

A.1. A neural network on Stiefel manifold

The neural network based on the methods described by Huang & Van Gool is shown in figure 7. A brief description of this network was provided in *section* 2.3.6.

$$\text{LinearClassifier}\left(\text{LogEig}\left(\mathcal{E}^{(3)}\right)\right)$$

$$\mathcal{E}^{(3)}_{16\times16}$$

$$\text{ReEig}\left(W^{(2)^T}_{16\times20}\mathcal{E}^{(2)}W^{(3)}_{20\times16}\right)$$

$$\mathcal{E}^{(2)}_{20\times20}$$

$$\text{ReEig}\left(W^{(2)^T}_{20\times22}\mathcal{E}^{(1)}W^{(2)}_{22\times20}\right)$$

$$\mathcal{E}^{(1)}_{22\times22}$$

$$\text{ReEig}\left(W^{(1)^T}_{22\times22}\mathcal{E}^{(0)}W^{(1)}_{22\times22}\right)$$
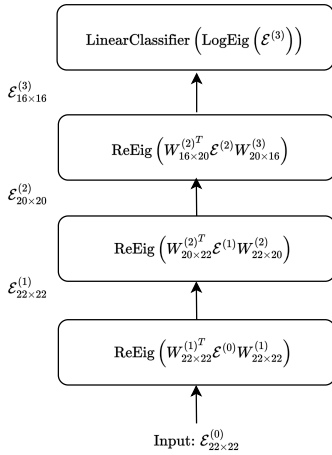
$$\text{Input: } \mathcal{E}^{(0)}_{22\times22}$$

Figure 7. Neural network architecture for SPD matrix learning.

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

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

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 & Van Gool for backpropagation formulae through non-linear and tangent space mapping layers.

A.2. Manifold recurrent neural network

The recurrent neural network based on the methods described by Jeong et al. is integrated with the methods given by Huang & Van Gool and is shown in figure 8. A brief description of this network was provided in *section* 2.3.7.

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

Gates of GRU as defined by Jeong et al. are given below.

Update-gate z_τ at time-step τ is

$$z_\tau = \text{SIGMOID}(w_z \lfloor \mathcal{L}_\tau \rfloor + u_z \lfloor h_{\tau-1} \rfloor + b_z) + \\ \text{SIGMOID}(b_{z'}[\exp(w_{z'} \log(\mathbb{D}(\mathcal{L}_\tau)) + \\ u_{z'} \log(\mathbb{D}(h_{\tau-1}))]), \quad (5)$$

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

Reset-gate r_τ at time-step τ is

$$r_\tau = \text{SIGMOID}(w_r \lfloor \mathcal{L}_\tau \rfloor + u_r \lfloor h_{\tau-1} \rfloor + b_r) + \\ \text{SIGMOID}(b_{r'}[\exp(w_{r'} \log(\mathbb{D}(\mathcal{L}_\tau)) + \\ u_{r'} \log(\mathbb{D}(h_{\tau-1}))]), \quad (6)$$

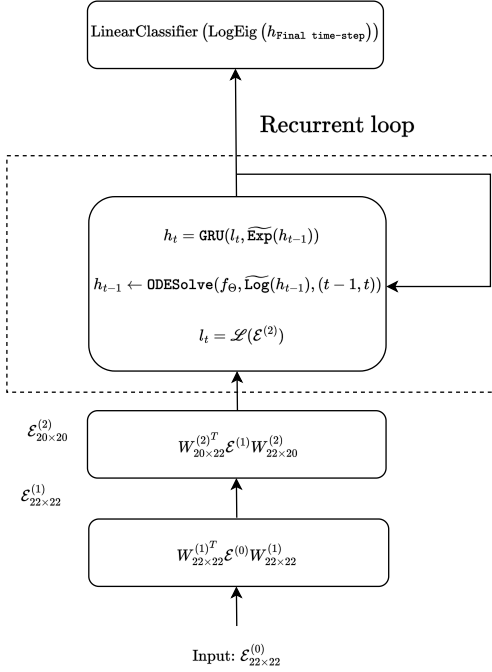where w_r, u_r, b_r, $w_{r'}$, and $u_{r'}$ are real weights and $b_{r'}$ is a real positive weight.

Figure 8. Recurrent neural network architecture for SPD matrix learning.

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

$$
\hat{h}_\tau = \text{TANH}(w_h \lfloor \mathcal{L}_\tau \rfloor + u_h(\lfloor r_\tau \rfloor * \lfloor h_{\tau-1} \rfloor) + b_h) +
$$
$$
\text{SOFTPLUS}(b_{h'} \exp(w_{h'} \log(\mathbb{D}(\mathcal{L}_\tau))
$$
$$
+ u_{h'} \log(\mathbb{D}(r_\tau) * \mathbb{D}(h_{\tau-1})))), \quad (7)
$$

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

Output vector h_τ is

$$
h_\tau = (1 - \lfloor z_\tau \rfloor) * \lfloor h_{\tau-1} \rfloor + \lfloor z_\tau \rfloor * \lfloor \hat{h}_\tau \rfloor +
$$
$$
\exp((1 - \mathbb{D}(z_\tau)) * \log(\mathbb{D}(h_{\tau-1})) +
$$
$$
\mathbb{D}(z_\tau) * \log(\mathbb{D}(\hat{h}_\tau))). \quad (8)
$$

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

We define an implicit layer solved using neural ordinary differential equations. The dynamics f of EMG data is modeled by a neural network with parameters Θ. The output state h_τ is updated as,

$$
h_{\tau-1} \leftarrow \text{ODESOLVE}(f_\Theta, \widetilde{\text{LOG}}(h_{\tau-1}), (\tau - 1, \tau))
$$
$$
h_\tau = \text{GRU}(\mathcal{L}_\tau, \widetilde{\text{EXP}}(h_{\tau-1})), \quad (9)
$$

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 in-

verse operation as defined by Lin. GRU is a gated recurrent unit whose gates are given by equations 5 - 8.

B. Detailed results

Here, we present detailed subject-wise results.

B.1. Orofacial gesture decoding

In table 5, we present the subject-wise decoding accuracy for 13 different orofacial gestures (described in *section* 2.1.1) using the *k*-medoids algorithm, as described in *section* 2.3.3. As observed, the edge matrices corresponding to different gestures exhibit a structured representation on the manifold of SPD matrices across subjects.

Subject number	Decoding accuracy of orofacial gestures using *k*-medoids algorithm
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 5. Different orofacial gestures are naturally distinguishable on the manifold of SPD matrices. Classification accuracy of 13 orofacial movements described in *section* 2.1.1. Chance accuracy is merely $\frac{1}{13} = 0.077$.

B.2. Phoneme and word articulation decoding

Here, we present the decoding accuracy of phoneme and word articulations using methods described in *sections* 2.3.4, 2.3.6, and 2.3.7.

B.2.1. PHONEMES AND WORDS DECODING USING MDM

In table 6, we present the decoding accuracy of phoneme articulations (described in *section* 2.1.2), and in table 7, we present the decoding accuracy of word articulations (described in *section* 2.1.3), using MDM (described in *section* 2.3.4). As observed, the edge matrices corresponding to different word and phoneme articulations exhibit a structured representation on the manifold of SPD matrices across subjects. Phoneme articulations from subject 11 were corrupted and are unavailable.

Subject number	Decoding accuracy of *audible* phonemes			Decoding 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 6. Phoneme articulations are naturally distinguishable on the manifold of SPD matrices. Classification accuracy of phoneme articulations using MDM. It is a 38-way classification problem with chance accuracy of $\frac{1}{38} = 0.026$. We also show classification accuracy for 23 consonant phonemes and 15 vowel phonemes separately. *A*-all phonemes, *C*-consonants only, *V*-vowels only.

Subject number	Decoding accuracy of *audible* words	Decoding 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 7. Word articulations are naturally distinguishable on the manifold of SPD matrices. Classification accuracy of word articulations using MDM. It is a 36-way classification problem with chance accuracy of $\frac{1}{36} = 0.028$.

B.2.2. PHONEMES AND WORDS DECODING USING METHOD IN *section* 2.3.6

While we use MDM (described in *section* 2.3.4) to demonstrate that phoneme and word articulations exhibit a structured representation on the manifold of SPD matrices, we can leverage the neural network described in *section* 2.3.6 to learn more expressive and distinguishable features. As shown here, decoding accuracy using the neural network is higher than that of MDM for both word and phoneme articulations. Table 8 presents the decoding accuracy of phoneme articulations, while table 9 reports the decoding accuracy of word articulations. Phoneme articulations from subject 11 were corrupted and are unavailable.

Subject number	Decoding accuracy of phoneme articulations using model in *section* 2.3.6					
	Audible phonemes			*Silent* phonemes		
	A	C	V	A	C	V
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 8. Classification accuracy of phoneme articulations using model in *section* 2.3.6. 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 $\frac{1}{38} = 0.026$. Accuracies are averaged over 10 random seeds.

Subject number	Decoding accuracy of word articulations using model in *section* 2.3.6	
	Audible words	*Silent* words
1	0.747±0.010	0.529±0.010
2	0.883±0.009	0.765±0.008
3	0.599±0.010	0.443±0.008
4	0.340±0.011	0.423±0.010
5	0.821±0.012	0.792±0.012
6	0.807±0.011	0.692±0.014
7	0.421±0.010	0.260±0.006
8	0.694±0.012	0.578±0.012
9	0.722±0.007	0.458±0.010
10	0.803±0.010	0.722±0.011
11	0.708±0.008	0.772±0.006
12	0.634±0.012	0.570±0.014
Mean	**0.681**	**0.584**

Table 9. Classification accuracy of word articulations using model in section 2.3.6. It is a 36-way classification problem with chance accuracy of $\frac{1}{36} = 0.028$. Accuracies are averaged over 10 random seeds.

B.2.3. PHONEMES AND WORDS DECODING USING METHOD IN *section* 2.3.7

While we use MDM (described in *section* 2.3.4) to demonstrate that phoneme and word articulations exhibit a structured representation on the manifold of SPD matrices, we can leverage the neural network described in *section* 2.3.7 to learn more expressive and distinguishable features. As shown here, decoding accuracy using the neural network is higher than that of MDM for both word and phoneme articulations. Table 10 presents the decoding accuracy of phoneme articulations, while table 11 reports the decoding accuracy of word articulations. Phoneme articulations from subject 11 were corrupted and are unavailable.

Subject number	Decoding accuracy of word articulations using model in *section* 2.3.7					
	Audible phonemes			*Silent* phonemes		
	A	C	V	A	C	V
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 10. Classification accuracy of phoneme articulations using model in *section* 2.3.7. 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 $\frac{1}{38} = 0.026$. Accuracies are averaged over 10 random seeds.

Subject number	Decoding accuracy of word articulations using model in *section* 2.3.7	
	Audible words	*Silent* words
1	0.788±0.029	0.669±0.055
2	0.865±0.013	0.745±0.025
3	0.656±0.041	0.542±0.033
4	0.305±0.026	0.381±0.034
5	0.769±0.024	0.712±0.037
6	0.767±0.034	0.674±0.027
7	0.260±0.006	0.274±0.025
8	0.755±0.038	0.661±0.025
9	0.690±0.025	0.523±0.037
10	0.846±0.026	0.738±0.021
11	0.781±0.022	0.810±0.024
12	0.681±0.035	0.606±0.036
Mean	**0.694**	**0.611**

Table 11. Classification accuracy of word articulations using model in section 2.3.7. It is a 36-way classification problem with chance accuracy of $\frac{1}{36} = 0.028$. Accuracies are averaged over 10 random seeds.

B.3. NATO phonetic code articulations

Here, we present the subject-wise decoding accuracies of the generalizable language-spelling paradigm described in *section* 2.1.4. Four subjects articulated the RAINBOW and GRANDFATHER passages character by character using NATO phonetic codes in a *silent* manner. The decoding accuracies are presented in table 12.

Subject number	Decoding accuracy using model in *section* 2.3.6	Decoding accuracy using model in *section* 2.3.7
1	0.697	0.8154
2	0.632	0.7837
3	0.638	0.7337
4	0.556	0.7606
Mean	**0.631**	**0.773**

Table 12. Top-*5* decoding accuracy of spelled-out RAINBOW and GRANDFATHER passage articulations using NATO phonetic codes. It is a 26-way classification problem with chance top-*5* accuracy of 0.178.

C. EMG data distribution shift

In *section* 3.3, we discussed the EMG data distribution shift across individuals. We showed that instead of similar words clustering together, each subject's embeddings exhibit a stronger within-subject similarity. To demonstrate this, we created graph edge matrices from EMG signals recorded during the *audible* articulation of words. These edge matrices, denoted as $\mathcal{E}$, were constructed using the entire articulation duration of 1.5 seconds.

To further verify whether a similar distribution shift occurs on a finer temporal scale - corresponding to individual phonemes or co-articulated phonemes - we had 12 subjects articulate the RAINBOW passage *audibly*, as one would naturally speak. The entire passage was articulated over a duration of 215 seconds, and SPD edge matrices were constructed using a time window of 100 *ms*. Consequently, we obtained 2150 edge matrices for the entire passage duration, capturing speech at a phonemic level. We visualize these 2150 edge matrices from each of the 12 subjects using *t*-SNE (*section* 2.3.5) in figure 9.

Similar to our observations in *section* 3.3, we found that articulations from different subjects exhibit strong within-subject similarity. That is, we did not observe any clustering based on articulation content - for example, similar phonemes from different subjects did not cluster together.

This finding suggests that, fundamentally, the functional connectivity of orofacial neuromuscular signals differs across individuals.
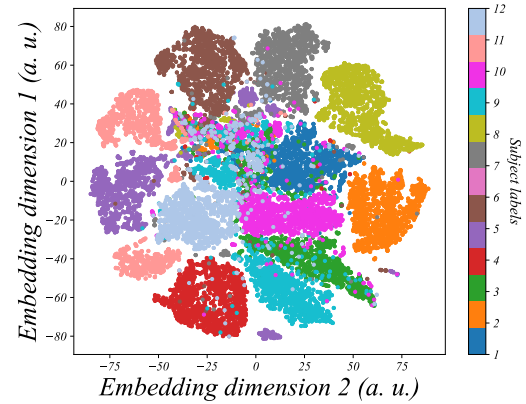
Figure 9. Articulations from different subjects cluster separately on the manifold of SPD matrices. *t*-SNE of edge matrices of RAINBOW passage articulations. Embedding is colored according to subjects (*a.u.* - arbitrary units).