A Deep Joint Sparse Non-negative Matrix Factorization Framework for Identifying the Common and Subject-specific Functional Units of Tongue Motion During Speech

Jonghye Woo\textsuperscript{a}, Fangxu Xing\textsuperscript{a}, Jerry L. Prince\textsuperscript{b}, Maureen Stone\textsuperscript{c}, Arnold D. Gomez\textsuperscript{d}, Timothy G. Reese\textsuperscript{e}, Van J. Wedeen\textsuperscript{e} and Georges El Fakhri\textsuperscript{a}

\textsuperscript{a}Gordon Center for Medical Imaging, Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA
\textsuperscript{b}Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218, USA
\textsuperscript{c}Department of Neural and Pain Sciences, University of Maryland School of Dentistry, Baltimore, MD 21201, USA
\textsuperscript{d}Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD 21218, USA
\textsuperscript{e}Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02129, USA

\textbf{ARTICLE INFO}

\textbf{Keywords:}
Tagged-MRI
Deep non-negative matrix factorization
Speech
Tongue Motion

\textbf{ABSTRACT}

Intelligible speech is produced by creating varying internal local muscle groupings—i.e., functional units—that are generated in a systematic and coordinated manner. There are two major challenges in characterizing and analyzing functional units. First, due to the complex and convoluted nature of tongue structure and function, it is of great importance to develop a method that can accurately decode complex muscle coordination patterns during speech. Second, it is challenging to keep identified functional units across subjects comparable due to their substantial variability. In this work, to address these challenges, we develop a new deep learning framework to identify common and subject-specific functional units of tongue motion during speech. Our framework hinges on joint deep graph-regularized sparse non-negative matrix factorization (NMF) using motion quantities derived from displacements by tagged Magnetic Resonance Imaging. More specifically, we transform NMF with sparse and graph regularizations into modular architectures akin to deep neural networks by means of unfolding the Iterative Shrinkage-Thresholding Algorithm to learn interpretable building blocks and associated weighting map. We then apply spectral clustering to common and subject-specific weighting maps from which we jointly determine the common and subject-specific functional units. Experiments carried out with simulated datasets show that the proposed method achieved on par or better clustering performance over the comparison methods. Experiments carried out with \textit{in vivo} tongue motion data show that the proposed method can determine the common and subject-specific functional units with increased interpretability and decreased size variability.

\textbf{1. Introduction}

Intelligible speech is produced by intricate and successful orchestration of local muscle groupings—i.e., functional units—of the extremely complex muscular architecture of the tongue (Woo et al., 2019a). The tongue is an organ that is controlled intricately by the support of its myoarchitecture, comprising an array of highly inter-digitated intrinsic and extrinsic muscles (Gaige et al., 2007). As a result, it is of great interest and need to study the intrinsic dimension-reduced structures of speech movements in order to better understand the mechanisms by which intrinsic and extrinsic muscles of the tongue coordinate to generate rapid yet accurate speech movements. To date, a great deal of work from different disciplines, including neurophysiology, biomechanics, speech and language, and medical imaging and analysis, has hypothesized and demonstrated that the control of tongue movements is governed by a reduced number of degrees of freedom (Gick and Stavness, 2013) that is associated with corresponding neuromuscular modules (Bizzi et al., 1991; Kelso, 2009), or fixed or mutable local muscle groupings (Woo et al., 2019a; Stone et al., 2004).

Medical imaging techniques, including magnetic resonance imaging (MRI), have been used to characterize functional units of speech movements (Woo et al., 2019a; Stone et al., 2004). In particular, tagged MRI allows us to non-invasively track spatiotemporally varying speech movements at the voxel level (Parthasarathy et al., 2007; Xing et al., 2017; Osman et al., 1999). More specifically, MR tagging can generate temporary grid-like patterns via a sequence of radiofrequency pulses within the tissue. This is achieved by spatially modulating longitudinal magnetization of hydrogen protons. As a result, the induced temporary grid-like tagging patterns deform alongside tongue motion and are visible perpendicular to tagging planes. Then, 2D plus time or 3D plus time velocity fields at the voxel level are typically estimated via tracking algorithms based on harmonic phase (HARP) (Parthasarathy et al., 2007; Osman et al., 1999; Xing et al., 2017).

In order to identify the “manageable” number of degrees of freedom of speech movements and better understand their spatial and temporal couplings between different parts of the tongue, various modeling approaches have been developed. Non-negative matrix factorization (NMF) (Lee and Seung, 1999) and its variants including sparse NMF are well-recognized, given that NMF is capable of examining signals derived from...
subject-specific motion tracking results from tagged MRI are first transformed into an atlas space representing a neutral tongue position. Our deep learning framework is then used to determine the common as well as subject-specific functional units. Intrinsic muscle activations that are non-negative (Ting and Chvatal, 2010). Sparse NMF (Kim and Park, 2008) is a matrix decomposition approach, where an input matrix whose entries are non-negative is expressed as a sparse linear combination of a set of building blocks. Since building blocks can be seen as an underlying anatomical basis, its associated weighting map can be used to reveal consistent and coherent sub-motion patterns. To further characterize the underlying physiology of speech movements using NMF, additional prior knowledge, such as manifold geometry of input movement data, has also been investigated (Woo et al., 2019a; Cai et al., 2010).

There are two major challenges to be addressed in this work. First, the prior approach (Woo et al., 2019a) to identify functional units using sparse NMF is based on a shallow NMF model, which may not capture the underlying tongue’s complex physiology accurately. The production of speech requires the complexity inherent in the execution, involving the activation of thousands of motor units in orthogonally oriented and interdigitated muscles. In addition, functional units are seen as 3D localized regions that show coherent displacement or related quantities, which are intermediate structures that interface between tongue muscle activation and tongue surface motion (Woo et al., 2019a). Accordingly, there is a need to develop an NMF model that can learn complex muscle coordination patterns from motion features derived from speech movements, while retaining the constraints and advantages of an NMF model which can deal with non-negative signals and offer parts-based and interpretable representations, respectively. In addition, a deep NMF is required, which can interrogate the relationship between complex muscle interdigitation and local activation and tongue surface motion (Woo et al., 2015; Stone et al., 2018). Second, because of the different motions that tongues produce during the course of speech, functional units vary substantially from one subject to another. Thus, one of the important hurdles in analyzing functional units is how to keep identified functional units across subjects comparable due to their substantial variability. Independently applying an NMF model to determining individual functional units may result in identifying building blocks and their weighting that are suboptimal, thereby yielding results that are challenging to objectively compare across subjects.

To alleviate the aforementioned challenges, we present a normalization method that can identify both the common and subject-specific functional units in a cohort of speakers in an atlas space from tagged MRI and 3D plus time voxel-level tracking by extending our prior work (Woo et al., 2020). In contrast to the prior work (Woo et al., 2020), we further describe a refined method using a deep joint sparse NMF framework to identify spatiotemporally varying functional units using a simple utterance and carry out extensive validations on both simulated and in vivo tongue motion data. Our deep joint sparse NMF framework computes a set of building blocks and both subject-specific and common weighting maps given motion quantities from tagged MRI. We then apply spectral clustering to the common and subject-specific weighting maps to jointly determine the common functional units across subjects and the subject-specific functional units for each subject.

In addition, we incorporate both sparse and graph regularizations into our framework. First, we impose a sparsity constraint on the weighting map to obtain the optimized and simplest functional units of tongue motion, which is consistent with the notion of “gestures” in phonological theories (Ramanarayanan et al., 2013). Second, we impose a graph regularization, which allows us to discover the intrinsic geometric structure of the motion data. Since a Euclidean
distance measure via Frobenius norm is used in this work, the intrinsic and manifold geometry of the input motion data is largely ignored. By incorporating both regularizations into our formulation, we can determine a set of simplest and intrinsic sub-motion patterns by promoting the computation of distances on a manifold. As well, it is possible to identify a low-dimensional yet interpretable subspace from tongue motion data.

The contributions of the proposed method can be summarized as follows:

- The most prominent contribution of this work is to construct an atlas of the functional units—i.e., the common consensus functional units—of how tongue muscles coordinate to produce target observed motion in a healthy population from cine and tagged MRI.

- This proposed work can simultaneously yield both common as well as subject-specific functional units within a material coordinate system with reduced size variability, thereby greatly facilitating the comparison of identified functional units during speech across subjects.

- This proposed work converts NMF with sparse and graph regularizations into modular architectures by means of unfolding Iterative Shrinkage-Thresholding Algorithm (ISTA), thereby accurately capturing the sub-motion patterns through each subject’s underlying low-dimensional subspace.

- This proposed work achieves on par or better clustering performance over the comparison methods on both simulated and in vivo tongue motion datasets.

- Experiments carried out with in vivo tongue motion data show that the proposed method can determine the common and subject-specific functional units with increased interpretability and decreased size variability.

The rest of this paper is structured as follows. Section 2 reviews related work. Section 3 defines the problem and describes our proposed approach. The experimental results are shown in Section 4, and Section 5 presents a discussion. Finally, we conclude this paper in Section 6.

2. Related Work

2.1. Functional Units

Various attempts have been made to investigate functional units of tongue motion during speech using imaging and motion capture techniques. For example, Green and Wang (2003) studied functionally independent articulators within the tongue based on a correlation analysis from an x-ray microbeam database. In that work, the functional independence was assessed through movement coupling relations, demonstrating phonemic differentiation in vertical tongue motions from 20 vowel-consonant-vowel (VCV) combinations. Similarly, Stone et al. (2004) examined the functional independence of five segments within the tongue during speech using a correlation analysis from 2D plus time ultrasound and tagged MRI. That work demonstrated that adjacent segments have high correlations, while distant segments have negative correlations consistent with linguistic constraints. The present work improves upon the previous work (Stone et al., 2004) by incorporating 3D plus time tagged MRI to assess how different parts of the tongue coordinate during speech. Ramanarayanan et al. (2013) proposed a computational framework to identify linguistically interpretable tongue movement primitives of speech articulation data based on a convolutive NMF algorithm with sparsity constraints from electromagnetic articulography and synthetic data generated via an articulatory synthesizer. Our proposed work is inspired by this approach (Ramanarayanan et al., 2013), but we use far richer 3D plus time displacements from tagged MRI together with deep NMF with the addition of sparsity and intrinsic data geometry in identifying functional units. Woo et al. (2019a) presented a framework to examine functional units using a shallow graph-regularized sparse NMF model from tagged MRI and 3D plus time voxel-level tracking. Recently, Sorensen et al. (2019) investigated a functional grouping of articulators and its variability across participants from real-time MRI. All of this work, however, studied subject-specific functional units and therefore lacked an understanding of the common functional units in a healthy population. In this work, we extend our prior approaches (Woo et al., 2019a, 2020) to develop a deep joint sparse NMF framework that can co-identify common and subject-specific functional units across participants.

2.2. Deep NMF

The recent success of deep neural networks allowed many researchers to investigate “deep NMF.” For example, a deep unfolding method was developed, yielding a new formulation that can be trained using a multiplicative back-propagation method (Hershey et al., 2014). In addition, deep NMF (Le Roux et al., 2015) was proposed by unfolding the NMF iterations and untying its parameters for the application of audio source separation. Furthermore, a new architecture combining NMF with deep recurrent neural networks (Wisdom et al., 2017) was presented by unfolding the iterations of ISTA (Gregor and LeCun, 2010). In the present work, we aim to develop deep NMF with both sparse and graph regularizations by unfolding the iteration of ISTA. We note that a similar idea has been explored in the prior work (Hershey et al., 2014; Le Roux et al., 2015; Wisdom et al., 2017) described above, but this work further incorporates both sparse and graph regularizations into the deep NMF framework.

3. Methodology

3.1. Participants and MRI Data Collection

In this work, a total of 18 healthy speakers were included. Table 1 lists the characteristics of subjects. All subjects are native speakers of American English with a Maryland accent. Each speaker was trained before the MR scan to speak a simple utterance (“a souk”) in line with a periodic metronome-like sound. This word is one of several that were chosen by
**Table 1**

Characteristics of 18 healthy subjects

| Subject | Age | Gender | Subject | Age | Gender |
|---------|-----|--------|---------|-----|--------|
| 1       | 23  | M      | 10      | 26  | F      |
| 2       | 31  | F      | 11      | 22  | M      |
| 3       | 27  | F      | 12      | 43  | M      |
| 4       | 41  | F      | 13      | 27  | M      |
| 5       | 35  | M      | 14      | 42  | F      |
| 6       | 45  | F      | 15      | 59  | F      |
| 7       | 27  | F      | 16      | 52  | M      |
| 8       | 22  | F      | 17      | 54  | M      |
| 9       | 22  | F      | 18      | 27  | M      |

design to move the tongue in specific directions, while being short enough to be spoken during the 1-second recording limit imposed by tag fading. The task begins with the /s/, which positions the tongue such that the vocal tract tube has an almost uniform cross-sectional area throughout its length. The tongue moves to an anterior position for /s/ and then posteriorly into /a/ and /l/. The vowel /a/ uses a closed jaw, as do the consonants, requiring that all vocal tract shaping be done by deforming the tongue, not merely opening and closing the jaw as can happen during /s/. Thus, the word keeps the tongue high, maximizes tongue deformation, and moves posteriorly primarily.

Each speaker repeated the speech word following the periodic sound, while acquiring T2-weighted 2D tagged and cine MRI through a Siemens 3.0 T Tim Trio system (Siemens Medical Solutions, Erlangen, Germany) with a 12-channel head coil and 4-channel neck coil. Both dynamic MR images were acquired at 26 frames per second with three orthogonal directions, including coronal, axial, and sagittal directions. Then, for cine MRI, a super-resolution volume reconstruction technique (Woo et al., 2012) was used to combine three orthogonal stacks to yield a single volume with isotropic resolution.

3.2. Estimation of Subject-specific Motion Fields from Tagged MRI

For the 3D plus time motion estimation, we use a tracking method by Xing et al. (2017) that hinges on symmetric and diffeomorphic registration with HARP phase volumes to yield a sequence of voxel-level motion fields during the speech tasks from tagged MRI. In brief, 2D slices into 3D voxel locations are interpolated using cubic B-spline. Then, a HARP tracking method (Osman et al., 1999) is utilized to yield HARP phase volumes. Finally, the iLogDemons method (Mansi et al., 2011) is applied to finding symmetric and diffeomorphic transformations from a reference time frame to the target time frame. The transformations are given by

\[ \varphi_{i,j} : \Omega \rightarrow \Omega, i = 1, \ldots, N, \]

where \( N \) denotes the number of subjects (in this work, \( N=18 \)), and \( j \) denotes the time frame index (i.e., \( j = 1, \ldots, M \)), where \( M \) denotes the total number of time frames for the utterance (in this work, \( M=26 \)) in these phase volumes. Finding symmetric and diffeomorphic transformations with the volume-preserving constraint is crucial for tongue motion analysis because the tongue’s volume remains invariant, and we need to preserve the smoothness of anatomical details within the tongue in the course of transformation.

3.3. Identification of Subject-specific Functional Units via a Deep Sparse NMF Framework

Assume that the tongue is comprised of \( K \) distinct clusters—i.e., functional units—in the course of a given phoneme of interest, each of which corresponds to a characteristic motion from which muscle coordinations and interactions occur. In this work, we opt to use graph-regularized sparse NMF to identify functional units for the following reasons. First, in order to accurately characterize each functional unit, it is necessary to project the high-dimensional and complex 3D plus time voxel-level tracking into a \( K \) low-dimensional subspace in which each axis corresponds to a particular sub-motion pattern. In addition, it is natural that functional units comprising a subset of intrinsic and extrinsic muscles are not entirely independent of each other; and there could be some overlaps among them. Furthermore, since it is assumed that functional units are the result of an additive mixture of the underlying muscle activations, the linear combination coefficients—i.e., weighting maps—need to take non-negative values only.

3.3.1. Extraction of Motion Features

We extract motion quantities from the 3D plus time motion estimation stated above to identify the functional units (Woo et al., 2019a): the magnitude and angle of each trajectory given by

\[ M^p_l = \sqrt{(x_{t+1}^p - x_l^p)^2 + (y_{t+1}^p - y_l^p)^2 + (z_{t+1}^p - z_l^p)^2} \quad (2) \]

\[ Z^p_l = \frac{x_{t+1}^p - x_l^p}{\sqrt{(x_{t+1}^p - x_l^p)^2 + (y_{t+1}^p - y_l^p)^2}} + 1 \quad (3) \]

\[ X^p_l = \frac{y_{t+1}^p - y_l^p}{\sqrt{(y_{t+1}^p - y_l^p)^2 + (z_{t+1}^p - z_l^p)^2}} + 1 \quad (4) \]

\[ Y^p_l = \frac{z_{t+1}^p - z_l^p}{\sqrt{(z_{t+1}^p - z_l^p)^2 + (x_{t+1}^p - x_l^p)^2}} + 1 \quad (5) \]

where \( M^p_l \) is the magnitude of each point trajectory and \( Z^p_l \), \( X^p_l \), and \( Y^p_l \) represent the cosine of the angle after projecting two consecutive adjacent point trajectories in the z, x, and y axes plus one to make sure that all values are non-negative, respectively.
The block diagram shows learned ISTA by unfolding the iteration of ISTA for sparse NMF (2 times in this figure).

```
\[ U \rightarrow V \]
```

3.3.2. Deep Graph-regularized Sparse NMF

Mathematically, the objective of NMF is to factorize the non-negative matrix \( U = [u_1, \ldots, u_n] \in \mathbb{R}^{n \times m} \) into the non-negative matrix \( V = [v_{ik}] \in \mathbb{R}^{m \times k} \), the building blocks, and the non-negative matrix \( W = [w_{ik}] \in \mathbb{R}^{k \times n} \), the weighting map, that minimizes the following objective function:

\[
E_1 = \frac{1}{2} \| U - VW \|_F^2,
\]

where \( \| \cdot \|_F \) represents the matrix Frobenius norm defined as

\[
\| V \|_F = \sqrt{\text{Tr}(V^T V)} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} v_{ij}^2}.
\]

Here, \( \text{Tr}(\cdot) \) denotes the trace of a matrix. Among other divergence measures (Lee and Seung, 1999; Cichocki et al., 2008; Sra and Dhillon, 2006), we focus on the Frobenius norm to compute dissimilarity between the non-negative input data matrix \( U \) and its approximation \( VW \). The objective function of graph-regularized sparse NMF can be defined as:

\[
E_2 = \frac{1}{2} \| U - VW \|_F^2 + \lambda \| W \|_1 + \beta \text{Tr}(WLW^T),
\]

where \( \lambda \) and \( \beta \) denote the balancing parameters of the sparsity and graph regularizations, respectively, and \( L \in \mathbb{R}^{n \times n} \) represents the graph Laplacian matrix. The graph Laplacian is defined as \( L = D - Q \), where \( Q \) is a heat kernel weighting function associated with the input matrix \( U \) and the degree matrix \( D \) is a diagonal matrix whose entries are \( D_{ii} = \sum_j Q_{ij} \). Minimizing the graph regularization term, \( \text{Tr}(WLW^T) \), serves as a smoothing operator. In this work, the building block, \( V \), and the initial weighting map, \( W^{(0)} \), are first initialized using the work by Cai et al. (2010). The ISTA method is then used to solve Eq. (9) for \( W \) as in Fig. 2 and Algorithm 1:

```
Algorithm 1: ISTA to solve Eq. (9)

Input: motion feature matrix \( U \), building block \( V \) and initial weighting map \( W^{(0)} \)
for \( h = 1 \) to \( H \) do
    \[ S^{(h)} = W^{(h-1)} + \frac{1}{c} V^T (U - VW^{(h-1)}) - \frac{\beta}{c} W^{(h-1)} L \]
    \[ W^{(h)} = \text{Soft}_{\lambda/c}(S^{(h)}) \]
Return \( W^{(H)} \)
```

Here \( 1/c, h \), and \( \text{Soft}_{\lambda/c}(z) \) denote the step size, the ISTA iteration index, and the soft thresholding function with a threshold value \( \lambda/c \), respectively. \( \text{Soft}_{\lambda/c}(z) \) is given by

\[
\text{Soft}_{\lambda/c}(z_n) = \text{sign}(z_n) \max(|z_n| - \lambda/c, 0). \tag{10}
\]

Because of the non-negative constraint imposed on \( W \), the soft-thresholding operation can be seen as a rectified linear unit (ReLU) activation function. It is worth noting that this minimization is equivalent to a fully connected layer, followed by ReLU activation, which bears structural similarity with the current deep neural network models.

3.3.3. Spectral Clustering

Once we obtain the weighting map from the ISTA method, we carry out clustering on the weighting map to partition the weighting map into disjoint subsets with high intra-cluster similarity, while maintaining low inter-cluster similarity via the eigen-structure of a data affinity graph. First, we construct an affinity matrix from the weighting map, which can be given by

\[
A(i, j) = \exp \left( -\frac{\| w(i) - w(j) \|_2}{\sigma} \right), \tag{11}
\]

where \( w(i) \) denotes the \( i \)-th column vector of the weighting map \( W \), and \( \sigma \) represents the scale factor. Then, spectral clustering (Shi and Malik, 2000) is carried out on the affinity matrix, followed by color-coding of each voxel within the tongue for visualization.

3.4. Deep Joint GS-NMF to Co-identify the Common and Subject-specific Functional Units

3.4.1. Construction of an average intensity and motion field atlas for a reference state from cine and tagged MRI

An average intensity and four-dimensional (4-D) motion field atlas (Woo et al., 2019c) is built for a reference time frame from cine and tagged MRI. Due to large variability in speech movements across subjects, even for the same speech task, putting all the data into an atlas space is crucial to facilitate the comparison of subjects by standardizing varying tongue shape and size, and motion field for each subject. Towards this goal, a symmetric diffeomorphic registration using a cross-correlation (CC) similarity metric is used to construct the average intensity atlas (Avants et al., 2011). Let
where \( \phi_i : \Omega_A \rightarrow \Omega_i \) denote the diffeomorphic transformation between the volume of the \( i \)-th subject and the atlas volume. Then, all the motion tracking results are mapped to the atlas space using the following transformation:

\[
\tilde{\phi}_{ij} = \phi_i \circ \phi_{ij} \circ \phi_i^{-1},
\]

where \( \tilde{\phi}_{ij} \) represents a motion field from the reference time frame to the \( j \)-th time frame of the \( i \)-th subject transformed in the atlas space. This Lagrangian configuration allows us to anchor the root of the motion fields in the material coordinates, thereby allowing for all motion features to be mapped back to the static anatomy (Woo et al., 2019c). Further, in order to achieve accurate time alignment across subjects, we visually identify the critical time instants (\( /a/, /s/, /u/, \) and \( /l/ \)) of all the subjects from their imaging data and align those instants at the same time positions. The data between those critical time positions are then interpolated based on the original imaging data and spread over an evenly distributed time grid.

### 3.4.2. Deep Joint GS-NMF

Once we establish the atlas and put all the data in the atlas space, we form a feature matrix to identify functional units (Woo et al., 2019a). Let \( U_i \) denote an input non-negative feature matrix of the \( i \)-th subject, consisting of the magnitude and angle derived from displacements as in Eq. (6). The objective function of GS-NMF for identifying individual functional units is defined as

\[
\mathcal{E}_3 = \frac{1}{2} \| U_i - V_i W_i \|_F^2 + \lambda \| W_i \|_1 + \beta \text{Tr}(W_i L_i W_i^T),
\]

where \( V_i, W_i, \) and \( L_i \) represent building blocks, their weighting map, and the graph Laplacian matrix for the \( i \)-th subject, respectively, and \( \lambda \) and \( \beta \) are weighting parameters for sparsity and graph regularizations, respectively. In order to identify the common weighting map, the following loss function, a measure of disparity between each weighting map and the common weighting map, is defined as

\[
\mathcal{E}_4 = \frac{1}{2} \| W_i - W^* \|_F^2,
\]

where \( W^* \) represents the common weighting map.

The overall objective function to find the building blocks, subject-specific weighting maps, and common weighting map is then defined as

\[
\mathcal{O}_2 = \sum_{i=1}^N \mathcal{E}_3 + \sum_{i=1}^N \gamma \mathcal{E}_4
\]

s.t. \( \forall i \in \{1, \cdots, N\}, V_i, W_i, W^* \geq 0 \)

where \( N \) denotes the number of subjects and \( \gamma \) represents a weighting parameter between the GS-NMF reconstruction error and the disparity term incorporating the common weighting map.

#### Algorithm 2: ISTA to solve Eq. (15)

**Input:** motion feature matrix \( U_i \), building block \( V_i \), initial weighting map \( W_i^{(0)} \) initialized by Cai et al. (2010), and initial common weighting map initialized by Eq. (18)

**Output:** \( W_{i}^{(H)}, W^* \)

for \( i = 1 \) to \( N \) do

for \( h = 1 \) to \( H \) do

Solving for \( S_i^{(h)} \) using Eq. (16) by fixing \( V_i \) and \( W^{(h)} \)

Solving for \( W_i^{(h)} \) using Eq. (17) by fixing \( V_i \) and \( W^{(h)} \)

Solving for \( W^* \) using Eq. (18) by fixing \( W_i^{(h)} \)

Return \( W_{i}^{(H)}, W^* \)

The objective function is optimized via an iterative and alternative ISTA update scheme as in Algorithm 2:

\[
S_i^{(h)}(z) = W_i^{(h-1)} + \frac{1}{c} \left[ V_i^T (U_i - V_i W_i^{(h-1)}) - \gamma (W_i^{(h-1)} - W^*) - \frac{\beta}{c} W_i^{(h-1) L_i} \right]
\]

\[
W_i^{(h)} = \text{Soft}_{\lambda/c}(S_i^{(h)}(z))
\]

\[
W^* = \frac{\sum_{i=1}^N a_i W_i^{(h)}}{\sum_{i=1}^N a_i}
\]

where \( 1/c, h, \) and \( \text{Soft}_{\lambda/c}(z) \) denote the step size, the ISTA iteration index, and the soft thresholding function with a threshold value \( \lambda/c \) as in Eq. (10), respectively.

#### 3.4.3. Spectral Clustering

The final subject-specific and common functional units are then obtained by applying spectral clustering to the weighting map for each subject and the common weighting map as described in Sec. 3.3.3.

#### 3.5. Complexity Analysis

In this subsection, we discuss the time complexity of our approach in comparison to the prior works. For simplicity, we assume that the time complexity of multiplication of two matrices—e.g., a \( p \times q \) matrix and a \( q \times r \) matrix—is \( O(pqr) \) (Cormen et al., 2009). The time complexity required for evaluating the recursive formula of our approach is \( O(N(Hmnk + n^2m)) \), where \( N \) is the number of subjects and \( H \) is the ISTA iteration number. The time complexities for GS-NMF-S and ISTA-S-NMF-S per subject are \( O(tpmnk + n^2m) \) and \( O(Hmnk) \), respectively, where \( t_p \) is the iteration number for the multiplicative updates. Thus, our approach has a time complexity similar to the prior works, while achieving superior clustering performance.
4. Experimental Results

In this section, we validate our approach described above on 2D plus time and 3D plus time synthetic motion data first because of the lack of ground truth in the in vivo tongue motion data. We then use in vivo tongue motion data obtained by tagged MRI to determine the common and subject-specific functional units. The clustering results are typically evaluated by comparing the label of each data point computed by methods against the ground truth label. Both the accuracy (AC) and the normalized mutual information (NMI) have been widely used to evaluate the clustering performance (Cai et al., 2010; Ghasedi Dizaji et al., 2017). Specifically, given points defined within the tongue \( x_i \), let \( s_i \) and \( g_i \) be the label identified by each method and the ground truth label, respectively. Then, AC is defined as:

\[
AC = \frac{\sum_{i=1}^{n} \delta(g_i, \text{map}(s_i))}{n} \times 100,
\]

where \( n \) denotes the total number of points, \( \text{map}(s_i) \) represents the mapping function that aligns each cluster label \( s_i \) with the ground truth label \( g_i \) via the Kuhn-Munkres algorithm (Lovász and Plummer, 2009), and \( \delta(a, b) \) is given by

\[
\delta(a, b) = \begin{cases} 
1 & \text{if } a = b \\
0 & \text{otherwise.} 
\end{cases}
\]

Let \( A \) and \( B \) denote the set of clusters computed from each method and ground truth, respectively, and the mutual information metric is given by

\[
\text{MI}(A, B) = \sum_{a_i \in A, b_j \in B} p(a_i, b_j) \cdot \log_2 \left( \frac{p(a_i, b_j)}{p(a_i) \cdot p(b_j)} \right),
\]

where \( p(a_i) \) and \( p(b_j) \) represent the probabilities that a point from the whole points belongs to the clusters \( a_i \) and \( b_j \), respectively, and \( p(a_i, b_j) \) represents the joint probability that the selected point simultaneously belongs to the clusters \( a_i \) and \( b_j \). Then, NMI is defined as follows:

\[
\text{NMI}(A, B) = \frac{\text{MI}(A, B)}{\max(H(A), H(B))} \times 100,
\]

where \( H(A) \) and \( H(B) \) denote the entropies of \( A \) and \( B \), respectively. Note that the range of NMI value is from 0 to 100, where 100 means the two set of clusters are the same.

Table 2: Comparison methods and their characteristics

| Method                      | Key characteristics |
|-----------------------------|---------------------|
| G-NMF-S (Cai et al., 2010)  | Shallow NMF, sparse regularization |
| GS-NMF-S (Woo et al., 2019a)| Shallow NMF, both sparse and graph regularizations |
| ISTA-S-NMF-S (a variant of Gregor and LeCun (2010)) | Deep NMF, sparse regularization |
| ISTA-GS-NMF-S (proposed)    | Deep NMF, both sparse and graph regularizations |

Figure 3: Illustration of 2D synthetic tongue motion simulation: (A) 2D displacement field, (B) the result using GS-NMF-S, (C) the result using ISTA-S-NMF-S, and (D) the result using our proposed approach. The different color represents different class labels.

4.1. Experiments Using Synthetic Tongue Motion Data

Our strategy for quantitative evaluation was to use the proposed method and the comparison methods (see Table 2), including graph-regularized NMF + spectral clustering (G-NMF-S), graph-regularized sparse NMF + spectral clustering (G-NMF-S) (Woo et al., 2019a), and ISTA for sparse NMF + spectral clustering (ISTA-S-NMF-S) to extract groupings from a synthetic displacement field composed of known areas representative of functional units. We then analyzed the difference between the grouping as outputted by the methods against the known distribution. We constructed simulated 2D and 3D displacement fields based on a tongue geometry derived from a vocal tract atlas previously developed (Woo et al., 2015; Stone et al., 2018). The 2D displacement fields were based on the areas illustrated in Fig. 3 and included Lagrangian displacements of heterogeneous magnitude representative of vertical, horizontal movement, and rotations in one deformed configuration. Table 3 lists numerical comparisons between G-NMF-S, GS-NMF-S (Woo et al., 2019a), ISTA-S-NMF-S, and the proposed approach (ISTA-GS-NMF-S). The results indicated that our approach surpassed the comparison methods in our 2D experiments. In our experiments, we chose \( \lambda=500, c=100, \beta=0, H=10, \) and \( \sigma=0.07 \) for ISTA-S-NMF-S and \( \lambda=500, c=100, \beta=0.05 \),
$H=10$, and $\sigma=0.07$ for our approach. These parameters were chosen to empirically maximize the clustering performance.

The 3D displacement fields included two temporal sequences of Lagrangian motion across 11 time frames each. The first dataset included spatially heterogeneous displacement fields as displayed in Fig. 4. The displacements were distributed based on the location of the verticalis (V), superior longitudinal (SL), and Transverse (T), which were defined using the vocal tract atlas for each time frame. Thus, we had a total of four ground truth labels in our quantitative evaluation. In addition, the V and SL muscles were rotated downward and upward, respectively, while the T muscle was translated upward in the course of 11 time frames (see Fig. 4). The second dataset also had composite Lagrangian displacement fields from 11 time frames as displayed in Fig. 5. We used the composite displacement field of genioglossus (GG), T, and geniohyoid (GH), which also were defined using the vocal tract atlas. We note that in the second dataset, the GG and T interdigitated with each other, and therefore we had a total of four ground truth labels in our quantitative evaluation. The GG and T muscles were rotated downward and upward, respectively, while the GH muscle was translated upward in the course of 11 time frames (see Fig. 5). The clustering outcomes using different methods are listed in Table 4, demonstrating that our approach achieved an accuracy level comparable or better than the comparison methods. In our experiments, for the first dataset, we chose $\lambda=890$, $c=55$, $\beta=0.03$, $H=49$, and $\sigma=0.05$ for ISTA-S-NMF-S and $\lambda=890$, $c=55$, $\beta=0.03$, $H=49$, and $\sigma=0.05$ for our approach. For the second dataset, we chose $\lambda=800$, $c=100$, $\beta=0.0$, $H=50$, and $\sigma=0.03$ for ISTA-S-NMF-S and $\lambda=800$, $c=100$, $\beta=0.05$, $H=50$, and $\sigma=0.03$ for our approach. These parameters were chosen to empirically maximize the clustering performance as shown in Figs. 6 and 7. In Fig. 6, in the case of the first dataset, for $H$, there was a local maxima, but in the case of the second dataset, in Fig. 7, after 20 iterations, our proposed approach converged to a global maxima, the perfect score, which appears to be a special case. The effects of $\lambda$ and $\beta$ on the clustering performance of 3D tongue simulation data are shown in Table 5. Our simulation study using 2D data shows that our approach using both regularizations achieved better performance, whereas our simulation study using 3D data shows that our approach performed on par with ISTA-S-NMF-S.

4.2. Experiments Using In Vivo Tongue Motion Data

We applied our proposed framework to a cohort of healthy subjects with a simple word “a souk” to identify both the common and subject-specific functional units in the atlas space. We first transformed all the motion fields into the atlas space. Second, we extracted the motion quantities, including the magnitude and angle of the motion trajectories and constructed an input spatiotemporal matrix containing 18 healthy subjects. Finally, we scaled the matrix, which was then inputted into our deep joint sparse NMF framework described above. The F-test was used to compare the variability of the sizes of the identified functional units from different approaches with a level of significance set at $p<0.05$. In addition, to test the normality of the sizes of the identified functional units, we performed the Anderson-Darling test (Scholz and Stephens, 1987). In all the experiments below, we chose $\lambda=800$, $c=600$, $\gamma=20$, $H=100$, and $\beta=0.03$ that are consistent with our experiments using the 3D tongue simulator.

Figs. 8 and 9 show two and three unit cluster representations of the common functional units, subject-specific functional units (subject 7 in Table 1) identified using the prior work (Woo et al., 2019a), and subject-specific functional units identified using our proposed approach of three distinct phonemes, including (1) /l/-/s/, (2) /l/-/ul/, and (3) /ul/-/k/ from “a souk.” For the transition of /l/ to /s/, the two functional units in Fig. 8(A) showed that the tip and base of the tongue are clustered together, which represents forward/upward motion, while the posterior tongue was clustered as a separate unit, which represents forward motion. The results from the proposed
Table 3
Clustering Performance of 2D Tongue Simulation Data

| 2D Tongue (%) | G-NMF-S | GS-NMF-S | ISTA-S-NMF-S | ISTA-GS-NMF-S |
|---------------|---------|----------|--------------|---------------|
| AC            | 85.74   | 86.15    | 91.23        | 97.53         |
| NMI           | 89.30   | 89.16    | 94.30        | 97.84         |

Table 4
Clustering Performance of 3D Tongue Motion Simulation Data

| Data 1 (%)  | G-NMF-S | GS-NMF-S | ISTA-S-NMF-S | ISTA-GS-NMF-S |
|-------------|---------|----------|--------------|---------------|
| AC          | 98.57   | 98.58    | 99.92        | 99.95         |
| NMI         | 96.70   | 95.73    | 99.58        | 99.72         |

| Data 2 (%)  | G-NMF-S | GS-NMF-S | ISTA-S-NMF-S | ISTA-GS-NMF-S |
|-------------|---------|----------|--------------|---------------|
| AC          | 99.37   | 99.36    | 100          | 100           |
| NMI         | 98.00   | 97.99    | 100          | 100           |

Table 5
Impact of Each Parameter on the Clustering Performance of 3D Tongue Motion Simulation Data

| Data 1 (%) | $\lambda=0$ | $\beta=0$ | $\lambda=0$ and $\beta=0$ |
|------------|--------------|------------|---------------------------|
| AC         | 36.85        | 99.92      | 81                        |
| NMI        | 0.62         | 99.58      | 74.1                      |

| Data 2 (%) | $\lambda=0$ | $\beta=0$ | $\lambda=0$ and $\beta=0$ |
|------------|--------------|------------|---------------------------|
| AC         | 37.36        | 100        | 99.38                     |
| NMI        | 1.78         | 100        | 98.05                     |

Notably, the functional units identified using our approach need to be interpreted in relation to the common functional units. More specifically, in Fig. 8(B) and Fig. 9(B), except for /s/-/l/, the functional units identified using the proposed approach look similar to the common functional units in Fig. 8(A) and Fig. 9(A). The functional units identified using the previous approach in Fig. 8(C) and Fig. 9(C), however, looks quite different. Fig. 8(C) and Fig. 9(C) appear to use a different strategy from the common functional units and the

approach in Figs. 8(B) and 9(B) showed clearer divisions between the units, thereby yielding more interpretable results in relation to the common functional units than the previous approach as visually assessed. The three functional units in Fig. 9(A) created clear divisions between the tongue base, tip, and posterior tongue. For the transition of /s/ to /u/, the two functional units (Fig. 8(A)) showed divisions between the anterior and posterior tongue. The three functional units (Fig. 9(A)) further formed clear divisions between the anterior, base, and posterior tongue. For the transition of /u/ to /k/, the upper tongue was clustered, since the tongue body was elevated, while the base and the body of the tongue were divided into separate units.
pressed the tongue anteriorly-to-posteriorly, which elevates the posterior surface (red) up toward the velum. This compression (green) would reflect the line of action of the inferior longitudinal (IL) muscle. The /u/-/k/ motion in Fig. 8(C) and Fig. 9(C) shows a single unit for the upper anterior surface (red) and tip and a single unit for the posterior and root region (green). This could reflect a single unit for GG posterior and GH, whose muscles are parallel in the anterior-to-posterior direction. The GG posterior shortening pulls the root forward, and elevates the upper tongue. The GH shortening elevates the entire tongue as a unit.

Fig. 10 illustrates the comparisons of the sizes of the identified functional units across subjects. Of note, all the data exhibited a normal distribution (Anderson-Darling test, \( p > 0.05 \)). For the transition of /s/ to /u/, the standard deviations of the sizes of the identified functional units from the previous approach and our approach were 10.4% and 6.8% for the two units \( (p < 0.05) \), respectively. For the three units, the standard deviations of the sizes of the identified functional units from the previous approach and our approach were 7.9% and 2.2% for unit 1 \( (p < 0.05) \), 8.1% and 2.3% for unit 2 \( (p < 0.05) \), and 8.8% and 3.3% for unit 3 \( (p < 0.05) \), respectively. The results indicated that our approach yielded reduced variability of the sizes of the functional units in terms of standard variations and that our approach and the previous approach showed significant statistical difference for all the units.

For the transition of /s/ to /u/, the standard deviations of the sizes of the functional units from the previous approach and our approach were 6.7% and 4.4% for the two units \( (p < 0.05) \), respectively (see Fig. 10). For the three units, the standard deviations of the sizes of the functional units from the previous approach and our approach were 6.9% and 3.3% for unit 1 \( (p < 0.05) \), 5.0% and 3.6% for unit 2 \( (p = 0.1) \), and 8.0% and 2.1% for unit 3 \( (p < 0.05) \), respectively. The results indicated that our approach yielded reduced variability of the sizes of the functional units in terms of standard variations, while our approach and the previous approach showed significant statistical difference except for unit 2 from three functional units.

For the transition of /u/ to /k/, the standard deviations of the sizes of the functional units from the previous approach and our approach were 10.2% and 3.7% for the two units \( (p = 0.45) \), respectively. For the three units, the standard deviations of the sizes of the functional units from the previous approach and our approach were 9.9% and 6.5% for unit 1 \( (p < 0.05) \), 7.1% and 5.9% for unit 2 \( (p = 0.23) \), and 9.3% and 6.7% for unit 3 \( (p = 0.09) \), respectively. We note that the results in Fig. 8(C) and Fig. 9(C) were identified by the previous approach \( (\text{Woo et al., 2019a}) \) in the atlas space, while the results in Fig. 8(B) and Fig. 9(B) were co-identified with the common functional units in Fig. 8(A) and Fig. 9(A). The results indicated that our approach yielded reduced variability of the sizes of the functional units in terms of standard variations and that our approach and the previous approach showed significant statistical difference except for units 2 and 3 from three functional units.

Functional units in Fig. 8(B) and Fig. 9(B). For example, in the /s/-/s/ motion. Fig. 8(C) and Fig. 9(C) have a bilateral difference in the upper tongue; the midline (green) is different from the lateral (blue), reflecting the development of a midline groove as the tongue moves into /s/. In Fig. 8(B) and Fig. 9(B), the groove development may be more subtle, i.e., smaller, or already present in /s/. In the /s/-/u/ motion, Fig. 8(C) and Fig. 9(C) still show 3 sections organized similar to the common functional units, but rougher in that the grouping is less crisp at the edges. In the /u/-/k/ motion, Fig. 8(C) and Fig. 9(C) again show that the division of functional units is less crispily segmented, compared with the functional units in Fig. 8(B) and Fig. 9(B). In Fig. 8(A) and Fig. 9(A) and Fig. 8(B) and Fig. 9(B), there is likely a compression/shortening of the tip-to-root region (green), compressing the tongue anteriorly-to-posteriorly, which elevates the posterior surface (red) up toward the velum. This compression (green) would reflect the line of action of the inferior longitudinal (IL) muscle. The /u/-/k/ motion in Fig. 8(C) and Fig. 9(C) shows a single unit for the upper anterior surface (red) and tip and a single unit for the posterior and root region (green). This could reflect a single unit for GG posterior and GH, whose muscles are parallel in the anterior-to-posterior direction. The GG posterior shortening pulls the root forward, and elevates the upper tongue. The GH shortening elevates the entire tongue as a unit.

First Author et al.: Preprint submitted to Elsevier
5. Discussion

The quest for identifying intrinsic “dimension-reduced modular structures”—i.e., functional units—has been central to research on speech production, including motor control theories from different perspectives. Early findings (Öhman, 1967; Mermelstein, 1973) indicated that the tongue is separated into tip and body carrying out “quasi-independent” motions. A recent study (Stone et al., 2004) suggested that the tongue could be further divided into the anterior, dorsal, middle, and posterior regions carrying out “quasi-independent” motions. Additionally, there is a great deal of work investigating factor analytic models, including Principal Component Analysis (PCA) (Slud et al., 2002; Stone et al., 1997, 2014; Xing et al., 2016) and NMF (Ramanarayanan et al., 2013; Woo et al., 2019a), to represent tongue motions as linear combinations of the basic factors. Our study furthers this underlying framework via a data-driven approach in which any different size, shape, and region of the tongue can constitute this modular structure according to the task at hand. This is made possible, in part, owing to recent technological advancements in MR imaging and analysis and machine learning that allow us to examine both tongue structure and function at an unprecedented resolution and accuracy.

The successful speech movement requires the orchestration of a highly flexible configuration of intrinsic and extrinsic muscles of the tongue and the vocal tract articulators. The cortical control of articulation is known to be carried out by the ventral sensorimotor cortex Bouchard et al. (2013). The production of intelligible speech arises from a coordinated motor pattern by means of a set of primitive or modular representations (Browman and Goldstein, 1992; Galantucci et al., 2006). To mine such a modular structure in the tongue inherent in speech movements using NMF, Woo et al. (2019a) proposed to incorporate two additional constraints, including sparsity and manifold geometry about the motion patterns, to determine a set of optimized and geometrically meaningful structures. This graph-regularized sparse NMF formulation allows computing a low-dimensional yet interpretable subspace, followed by identifying subject-specific functional units via spectral clustering. More recently, Woo et al. (2020) investigated the use of the same sparse NMF framework in a groupwise setting to co-identify the common and subject-specific functional units to increase interpretability due to large variability in the identified functional units across subjects. In the present work, we further proposed a joint deep graph-regularized sparse NMF and spectral clustering to co-identify the common and subject-specific functional units. This, in turn, increased interpretability and decreased size variability in the identified functional units compared with the previous approach (Woo et al., 2019a). In addition, the identified subject-specific functional units are jointly obtained alongside the common functional units,
Figure 9: Illustration of (A) the common functional units (3 units) identified using our proposed approach, (B) three functional units identified using the prior work (Woo et al., 2019a), and (C) three functional units identified using our proposed method for the transitions of /æ/-/s/, /s/-/u/, and /u/-/k/, respectively.

thereby greatly facilitating the comparison of each subject with another.

To achieve deep NMF, we converted the standard NMF with sparse and graph regularizations into modular architectures using unfolding ISTA to learn building blocks and associated weighting map. The deep NMF using unfolding ISTA (Gregor and LeCun, 2010) has been studied previously, but it is worth noting that, to our knowledge, this is the first attempt at incorporating both sparse and graph regularizations into the ISTA framework. In addition, we further introduced a common low-dimensional subspace that can learn the common weighting map jointly with subject-specific weighting maps across subjects.

The use of a deep variant of NMF based on ISTA is uniquely important to decompose the complex muscle coordination patterns into modular components or functional units over other alternative statistical techniques of matrix factorization, such as shallow NMF, PCA, Independent Component Analysis (ICA), or factor analysis. First and foremost, shallow NMF and its variants have been widely used for research on muscle synergies due to their great interpretability (Shourijeh et al., 2016; Ting and Macpherson, 2005; Torres-Oviedo and Ting, 2007; Bruton and O’Dwyer, 2018). The shallow model, however, learns functional units or synergies by directly mapping the internal tongue motion to its underlying subspace. Successful tongue movement hinges on the orchestration of a complex set of neural activations of numerous intrinsic and extrinsic tongue muscles. As such, considering the complex tongue structure and function, it is highly likely that the mapping between the internal tongue motion and its underlying low-dimensional subspace contains rather complex hierarchical information, which may not be captured by shallow NMF-based approaches. Second, PCA has been the most widely used method applied to kinematic data (Wang et al., 2013), while NMF-based approaches are the best suited to studying muscle synergies because of their ability to handle the non-negative nature of muscle activation signals (Bruton and O’Dwyer, 2018). More specifically, while both NMF and PCA learn low-dimensional building blocks and their weighting map, the non-negative constraints imposed on the decomposition process in NMF lead to a marked difference between the two methods. For example, the obtained building blocks from NMF are independent. If the distribution of the data is Gaussian, by contrast, then the obtained building blocks from PCA are orthogonal and independent. Otherwise, these building blocks will merely be uncorrelated, not necessarily independent. Therefore, NMF-based approaches learn interpretable and parts-based representations in that a set of components is combined to form a whole in a non-subtractive manner. In contrast, PCA represents each data as a linear combination of a limited number of building blocks.
Figure 10: The comparison of the sizes of the three functional units using our approach and the previous approach (Woo et al., 2019a) for the transitions of /æ/-/s/, /s/-/u/, and /u/-/k/.

that explain the maximal amount of variance. Since there are no non-negative constraints in PCA, the linear combination may mix the elements of the building blocks and weighting map, which can cancel each other out. Consequently, there is a lack of physical meaningfulness of the building blocks. (Woo et al., 2019a)

There are a few limitations in this work. First, quantitative evaluation of the proposed work in the context of in vivo tongue motion is a challenging task. The notion of accuracy within our unsupervised learning setting is ill-posed, as accurate validation is impossible due to the lack of ground truth other than simulation studies and visual assessment with a thorough knowledge of tongue structure and function. In the present work, a tongue motion simulator based on a vocal tract atlas (Woo et al., 2015) was used to generate Lagrangian tongue motion. With this simulator alongside the ground truth, we were able to validate our method, showing superior performance over the comparison methods. Second, the ideal method to measure motor control is electromyography (EMG) because it records the activation of muscles. However, the muscles of the tongue are orthogonal and almost entirely interdigitated, making it almost impossible to disambiguate one direction of fiber activation from another. Moreover, unlike typical skeletal muscles, the muscles of the tongue are not designed to move a bone around a joint, but to deform its surface to shape the vocal tract tube. Only imaging methods, including MRI and ultrasound, can measure speech motor control, while not interfering with speech. Ultrasound has also been used to study tongue motion, but it only records the tongue surface, not internal motion, and cannot see the structures beyond the tongue surface as the sound does not penetrate beyond the first tissue interface. Therefore, tagged MRI is the only modality that can be used to study speech motor control to the best of our knowledge.

Third, in this work, we chose the number of clusters without a principled approach. In addition, there were a few parameters that we tuned with the help of the 3D tongue simulator. The development of a new approach to determine the optimal number of clusters in conjunction with optimization parameters is a subject for future research. Finally, in our simulation studies, while we used representative simulation datasets to test our approach, the number of sample size is too small to compute statistical significance. In our future work, we will increase the number of datasets to compute statistical difference between different approaches.

There are a few ways to expand on this work. First, the human tongue consists of numerous intrinsic and extrinsic muscles, each of which has distinct roles to compress and expand tissue points. For example, GG has a muscular architecture that locally activates different parts of the muscle, from GG anterior to GG posterior (Miyawaki, 1975; Stone et al., 2004). As such, identifying such fine-grained local functional units within a single muscle or a subset of muscles in a hierarchical manner would reveal new insights into the mechanisms of how different elements of muscular architecture interact with each other. In order to accurately localize the internal muscles, structural MRI or diffusion MRI is needed as they can provide the location of the internal muscles or fiber architecture, respectively. Accurate registration (Woo et al., 2014) is needed to put the imaging data into correspondence for both controls and patients (Liu et al., 2021). Second, various intra-subject variabilities in speech articulation is not fully explored in this work. In our future work, we will investigate functional units of a range of motion patterns having intra-subject variability to see the central tendency and its variability in the identified functional units. Finally, our framework can be applied to patient populations, such as those with amyotrophic lat-
eral sclerosis (Xing et al., 2018; Lee et al., 2018) or tongue cancer with speech or swallowing impairments (Woo et al., 2019b); assessing how local functional units adapt after a variety of treatments can potentially advance therapeutic, rehabilitative, and surgical procedures. For example, our prior work (Xing et al., 2019) found an increased correlation between the floor-of-mouth muscle group and internal tongue muscle group for tongue cancer patients compared with healthy subjects to compensate for their post-surgery function loss. Our functional units analysis will further shed light on how patients adapt their speech movements depending on different tumor size and location as well as treatment methods.

To the best of our knowledge, this is the first report identifying common and subject-specific functional units from cine and tagged MRI. The atlas constructed from cine MRI was used as a reference anatomical configuration for subsequent analyses to identify and visualize the functional units of the internal motion patterns during speech. In this way, it was possible to contrast and compare the identified functional units across subjects that were not biased by each subject’s anatomical characteristics. In addition, the proposed work furthered this underlying concept in which constructing the atlas of functional units was carried out in a low-dimensional subspace, since correspondences across subjects in the low-dimensional subspace were guaranteed through the reference material coordinate system. Therefore, the proposed work holds promise to provide a link between internal tongue motion and underlying low-dimensional subspace, thereby advancing our understanding of the inner workings of the tongue during speech. In addition, the identified common and subject-specific functional units could offer a unique resource in the scientific research community and open new vistas for functional studies of the tongue.

6. Conclusion

In this work, we presented a new method to jointly identify common and subject-specific functional units. To address the limitations of shallow NMF and identify comparable and interpretable functional units across subjects, a deep joint NMF framework incorporating sparse and graph regularizations was proposed. Our proposed method was extensively validated on synthetic and real data to demonstrate the benefit of its novel features. Our results show that our method can determine the common and subject-specific functional units with increased interpretability and decreased size variability.

Acknowledgments

This work is partially supported by NIH R01DC014717, R01DC018511, R01CA133015, R21DC016047, R00DC012575, P41EB022544 and NSF 1504804 PoLS.

References

Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C., 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 54, 2033–2044.

Rizzi, E., Mussa-Ivaldi, F.A., Giszter, S., 1991. Computations underlying the execution of movement: a biological perspective. Science 253, 287–291.

Bouchard, K.E., Mesgarani, N., Johnson, K., Chang, E.F., 2013. Functional organization of human sensorimotor cortex for speech articulation. Nature 495, 327–332.

Browman, C.P., Goldstein, L., 1992. Articulatory phonology: An overview. Phonetica 49, 155–180.

Bruton, M., O’Dwyer, N., 2018. Synergies in coordination: a comprehensive overview of neural, computational, and behavioral approaches. Journal of neurophysiology 120, 2761–2774.

Cai, D., He, X., Han, J., Huang, T.S., 2010. Graph regularized nonnegative matrix factorization for data representation. IEEE transactions on pattern analysis and machine intelligence 33, 1548–1560.

Cichocki, A., Lee, H., Kim, Y.D., Choi, S., 2008. Non-negative matrix factorization with $\alpha$-divergence. Pattern Recognition Letters 29, 1433–1440.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C., 2009. Introduction to algorithms. MIT Press.

Gaige, T.A., Benner, T., Wang, R., Wedeen, V.J., Gilbert, R.J., 2007. Three dimensional myoarchitecture of the human tongue determined in vivo by diffusion tensor imaging with tractography. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine 26, 654–661.

Galantucci, B., Fowler, C.A., Turvey, M.T., 2006. The motor theory of speech perception reviewed. Psychonomic bulletin & review 13, 361–377.

Ghasedi Dizaji, K., Herandi, A., Deng, C., Cai, W., Huang, H., 2017. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization, in: Proceedings of the IEEE international conference on computer vision, pp. 5736–5745.

Gick, B., Stavness, I., 2013. Modularizing speech. Frontiers in psychology 4, 977.

Green, J.R., Wang, Y.T., 2003. Tongue-surface movement patterns during speech and swallowing. The Journal of the Acoustical Society of America 113, 2820–2833.

Gregor, K., LeCun, Y., 2010. Learning fast approximations of sparse coding, in: Proceedings of the 27th International Conference on International Conference on Machine Learning, pp. 399–406.

Hershey, J.R., Roux, J.L., Weninger, F., 2014. Deep unfolding: Model-based inspiration of novel deep architectures. arXiv preprint arXiv:1409.2574.

Kelso, J.S., 2009. Synergies: atoms of brain and behavior, in: Progress in motor control. Springer, pp. 83–91.

Kim, J., Park, H., 2008. Sparse nonnegative matrix factorization for clustering. Technical Report. Georgia Institute of Technology.

Le Roux, J., Hershey, J.R., Weninger, F., 2015. Deep NMF for speech separation, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 66–70.

Lee, D.D., Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791.

Lee, E., Xing, F., Ahn, S., Reese, T.G., Wang, R., Green, J.R., Atassi, N., Wedeen, V.J., El Fakhri, G., Woo, J., 2018. Magnetic resonance imaging based anatomical assessment of tongue impairment due to amyotrophic lateral sclerosis: A preliminary study. The Journal of the Acoustical Society of America 143, EL248–EL254.

Liu, X., Xing, F., Yang, C., Kuo, C.C.J., ElFakhri, G., Woo, J., 2021. Symmetric-constrained irregular structure inpainting for brain mr imaging with tumor pathology. arXiv preprint arXiv:2101.06775.

Lovász, L., Plummer, M.D., 2009. Matching theory. Volume 367. American Mathematical Soc.

Mansi, T., Penneck, X., Sermesant, M., Delingette, H., Ayache, N., 2011. iLogDemons: A demons-based registration algorithm for tracking incompressible elastic biological tissues. International journal of computer vision 92, 92–111.

Merelstein, P., 1973. Articulatory model for the study of speech production. The Journal of the Acoustical Society of America 53, 1070–1082.

Miyawaki, K., 1975. A preliminary report on the electromyographic study of the tongue during speech. In addition, the identified common and subject-specific functional units from cine and tagged MRI. The atlas constructed from cine MRI was used as a reference anatomical configuration for subsequent analyses to identify and visualize the functional units of the internal motion patterns during speech. In this way, it was possible to contrast and compare the identified functional units across subjects that were not biased by each subject’s anatomical characteristics. In addition, the proposed work furthered this underlying concept in which constructing the atlas of functional units was carried out in a low-dimensional subspace, since correspondences across subjects in the low-dimensional subspace were guaranteed through the reference material coordinate system. Therefore, the proposed work holds promise to provide a link between internal tongue motion and underlying low-dimensional subspace, thereby advancing our understanding of the inner workings of the tongue during speech. In addition, the identified common and subject-specific functional units could offer a unique resource in the scientific research community and open new vistas for functional studies of the tongue.

6. Conclusion

In this work, we presented a new method to jointly identify common and subject-specific functional units. To address the limitations of shallow NMF and identify comparable and interpretable functional units across subjects, a deep joint NMF framework incorporating sparse and graph regularizations was proposed. Our proposed method was extensively validated on synthetic and in vivo tongue motion data to demonstrate the benefit of its novel features. Our results show that our method can determine the common and subject-specific functional units with increased interpretability and decreased size variability.

Acknowledgments

This work is partially supported by NIH R01DC014717, R01DC018511, R01CA133015, R21DC016047, R00DC012575, P41EB022544 and NSF 1504804 PoLS.

References

Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C., 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 54, 2033–2044.
