Probing articulatory representation learning for phonological distinctions

Sean Foley
seanfole@usc.edu

April 16, 2024

Abstract

Speech production is hierarchical in nature, involving descending motor commands to the articulators, which are in turn modulated by higher-level cognitive representations. While a growing body of work has aimed to extract spatio-temporal units directly from articulatory data, there have been few attempts to probe the extent to which such representations capture meaningful phonological *contrasts* employed in language and to model this mapping between motor plans and phonological representations across speakers. This study employs a joint factor analysis and neural convolutive matrix factorization framework to a multispeaker rtMRI dataset of vocal tract contours. The framework generates both *gestures*, the spatio-temporal units that form a given utterance, and *gestural scores*, which detail the activation of individual gestures in time. Probing of the gestural scores shows some ability to capture phonological distinctions, suggesting that such information is encoded by the model. The gestures, however, show poor discriminability along crucial phonological dimensions, likely limited by cross-speaker spatial variability. The results highlight many of the difficulties in cross-speaker articulatory modeling, but also show some promise in using deep learning to model articulatory representations.

1 Introduction

Speech production involves precise coordination of multiple articulators in the vocal tract. Similar to other forms of motor control, it has been argued that this coordination is highly *modular*, being controlled by a finite set of functional, synergistic units (Bizzi and Cheung, 2013; Gick and Stavness, 2013). Crucially, decades of experimental work suggests that such units are modulated by higher-level cognitive information, specifically *phonological* representations that encode meaningful contrasts employed in language, implicating the hierarchical nature of speech production (Poliva et al., 2024; Goldstein et al., 2007). Speech motor control acquisition then involves stages of learning a mapping between motor control mechanisms, speech planning, and higher-level phonological representations (Parrell et al., 2019; Guenther, 2016; Hickok and Poeppel, 2007). A crucial theoretical assumption regarding these phonological representations is that they are *stable*. That is, while the vocal tract movements that are the output of a given phonological unit may vary continuously in time due to contextual and paralinguistic variation, the underlying phonological unit itself does not vary contextually.

The development of self-supervised learning within the field of deep learning has led to comparisons between the representations learned by such models and those employed in human cognitive and neural processes (Konkle and Alvarez, 2022; Martin et al., 2023). In the area of speech production, a growing body of work has aimed to decompose articulatory kinematic data into spatio-temporal units using machine learning and deep learning (Lian et al., 2022, 2023; Toutios and Narayanan, 2015; Ramanarayanan et al., 2013). While the representations learned in previous work clearly capture low-level coordination among the articulators, it is unclear whether these models develop representations that can also capture phonological contrasts across speakers. The present study aims to generate interpretable articulatory representations *across* a large set of speakers directly from articulatory data and probe the extent to which they capture crucial phonological distinctions, testing their ability to generalize from low-level articulatory coordination to abstract cognitive representations.

2 Background

2.1 Speech motor control, modularity, and articulatory gestures

A well-known problem within the field of motor control is the so-called "degrees of freedom problem" (Jordan, 2018; Faytak, 2018). That is, given the high number of degrees of freedom possible in various movements systems, e.g. reaching or speech production, how does the central nervous system (CNS) effectively and efficiently control movement patterns? It is widely believed that such control is handled by making use of a finite set of movement primitives (Bizzi et al., 2008; Bizzi and Cheung, 2013). Such primitives are coordinated, synergistic units composed of multiple muscles or organs working together to achieve some goal. In the case of speech, rather than the CNS constantly computing individual tongue or jaw movement trajectories, it makes use of primitives of combined and coordinated tongue and jaw (plus other articulator) movements to achieve a constriction in the vocal tract.

In reference to speech primitives, some have proposed viewing them as neuromuscular *modules*, linking them directly with the biomechanics underlying speech production (Gick and Stavness, 2013). In this view, an individual module activates a predefined set of muscles, with each muscle weighted proportionally to its role in the desired action. Furthermore, each module is "functionally defined" in the CNS to produce real-word bodily action (Gick, 2016). In the Articulatory Phonology (AP) framework, the atoms of speech are *articulatory gestures* (Browman and Goldstein, 1992). Gestures are defined as abstract coordinative units of articulators used to achieve constriction tasks in the vocal tract. While there may be some relation between modules and gestures (Gick and Stavness, 2013), gestures are crucially defined as dynamical regimes that can achieve a goal in a task space. That is, AP originally modeled phonological distinctions via distinctions in task-defined tract variables, including constriction location and degree. For example, the [t]/[s] contrast due to differences in constriction degree may be modeled as a distinction in full closure versus "critical" degree of constriction. Crucially, in this model gestures are not activated sequentially in time but rather are concurrently activated in accord with the demands of completing the constriction task. For example, the gestures typically associated with the segment [n] would include coordinated velum lowering and tongue tip closure gestures. This simultaneous activation is usually represented graphically using a *gestural score*, a gestural activation matrix that has vocal tract organs on the y-axis and time on the x-axis. An example of this can be seen in Figure 1 for the word "bad". Three gestures are active including bilabial and alveolar closure gestures using the lips and tongue tip respectively, and a more open ("wide") pharyngeal constriction using the tongue body for the vowel. Notice the overlap between the vowel and each consonant gesture, with such overlap being typical in CVC syllables. In accord with the Task Dynamics (TD) framework (Saltzman and Munhall, 1989), AP models the dynamics of gestures as critically damped point attractor systems within the space of vocal tract constrictions.

Evidence from neuroscience points towards the sensorimotor cortex encoding vocal tract movements into task-based units composed of multiple articulators (Chartier et al., 2018; Mugler et al., 2018). Using intra-cranial cortical recordings during speech production, (Mugler et al., 2018) found that cortical activity varied more during same phoneme produced in different contexts versus the same constriction task produced in different contexts. The authors argue that this suggests the representations encoded in the speech motor cortices are primarily gesture-based. Similarly, also using cortical recordings, (Chartier et al., 2018) showed that the sensorimotor cortex encodes vocal tract movements into task-based coordinative units, and that such an articulator-based model is a better fit of the neural activity than one based on phonemes. However, it is worth noting that while Chartier et al. (2018) found gesture-like organization, they did find considerable variance in neural activity due to contextual effects, suggesting the proposed invariance of speech primitives is represented elsewhere during speech processing. This highlights an important distinction between contrast and invariance. In frameworks such as AP/TD, invariance relates to the consistent parameters of the dynamical regimes underlying each gesture, ensuring stability across different

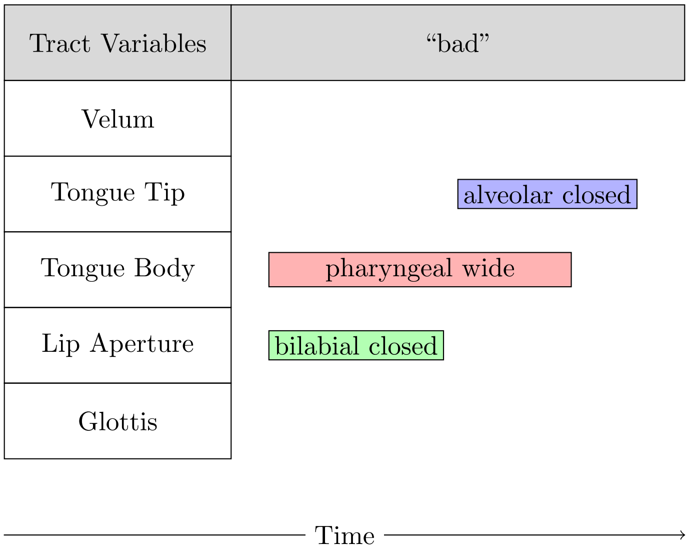

Figure 1: Gestural score for the word "bad", based on Byrd and Krivokapić (2021).

contexts. Contrast, on the other hand, emerges through distinctive modes within the distribution of constriction goals, allowing for the differentiation of phonological units.

2.2 Articulatory representation learning

Recent advancements in deep learning have given risen to the sub-field known as *representation learning*. Representation learning can be defined as learning a representation of the input to the model that is highly informative or can approximate cognitive representations to some extent (Bengio et al., 2013). Typically, this involves feeding data to a model and extracting latent or "hidden" representations of that input at some level of the model. For example, Beguš and Zhou (2021) fed acoustic data to a generative adversarial network (GAN) composed of multiple convolutional layers and probed the extent to which different phonetic information is encoded at different layers. They found that some layers encode f0 and intensity to a greater extent than other layers, suggesting that as the input acoustic data is passed through the model, different layers encode different key aspects of this input.

A growing body of work has employed machine learning to learn representations directly from articulatory data. One of the more successful methods has been to treat the problem as one of matrix factorization. Matrix factorization aims to decompose a matrix into a set of basis vectors. The basis vectors can be thought of as the 'building blocks' of the input matrix. In addition to the basis vectors a set of activation weights are learned that detail the extent to which each basis vector is present in the input matrix. Ramanarayanan et al. (2013) proposed a convolutive non-negative matrix factorization (CNMF) approach to decompose electromagnetometry (EMA) data into both interpretable gesture-like units and activation matrices that show the activation of each unit in time. Additionally, sparseness constraints were imposed on the activation matrix such that only a small set of units are active at any given moment. Later work also showed that these sparse activation matrices could be used to distinguish phones (Ramanarayanan et al., 2016), though no direct comparison was done of the gestural units themselves nor was context incorporated into the model. Recently, this earlier work has been reformulated within a deep learning framework (Lian et al., 2022, 2023). Lian et al. (2022) proposed a convolutive neural matrix factorization approach using an autoencoder (*neural* indicating the use of a neural network), where an encoder takes in EMA data and outputs activation matrices and a decoder outputs a set of basis vectors used to reconstruct the input using the encoder activation matrices. This method generated both interpretables activation matrices, akin to *gestural scores* in AP, and gesture-like units. Lian et al. (2023) extended this to rtMRI data using a joint factor analysis and CNMF method. Using a corpus of data from 8 speakers, this method also produced interpretable gestural units, although no cross-speaker comparisons were made and no quantitative analysis done on the gestural units themselves.

The current study extends previous work in articulatory representation learning. While earlier work has shown clearly that it is possible to extract interpretable representations directly from articulatory kinematic data, it is unclear whether such models can learn generalizations from this data that are indicative of higher-level cognitive units employed in language. Additionally, previous work has been limited in the extent to which representations have been learned *across* speakers, likely due to difficulties in cross-speaker comparisons attributed to differences in vocal tract morphology. In the current study, the aim is two-fold: 1) extract interpretable gesture-like units and gestural scores directly from articulatory data across a cohort of speakers; 2) probe extent to which the learned representations encode meaningful distinctions in vocal tract actions, e.g. tongue tip constriction versus tongue dorsum constriction. Regarding this latter goal, such distinctions are most meaningful in certain contexts, for example in the difference between the words "tap" and "gap", where this contrast is held by the different places of articulation of the initial consonants. While the underlying compositional structure may exist elsewhere, context makes such distinctions more salient.

To account for the role of context, context features will be incorporated into the models to probe not only whether the learned representations can distinguish phonemes or constriction tasks, but whether context aids in making these distinctions. If the learned representations truly capture phonological contrasts, it should be the case that the representations are more informative in a given context. To probe this, classifiers will be tasked with making phonological distinctions using the learned articulatory representations with and without context features. If the representation learning model is encoding phonological information, it is predicted that context features will significantly improve classification performance using the learned representations. Performance is not expected to be as strong without contextual information, given that contextual variance can obscure differences between constriction tasks.

3 Method

3.1 Problem Formulation

Given a set of vocal tract contours $\mathbf{X} \in \mathbb{R}^{2p \times t}$, where p is the number of vertices and t the number of frames, we aim to decompose $\mathbf{X}$ into a set of gestures $\mathbf{G} \in \mathbb{R}^{K \times D \times 2p}$, where K is the kernel size and D is the number of gestures, and gestural scores $\mathbf{H} \in \mathbb{R}^{D \times t}$. The gestures represent the spatio-temporal units employed to produce the given utterance, while the gestural scores are the activation matrix indicating when each gesture is active in time.

3.2 Model

We follow the joint factor analysis and neural convolutive matrix factorization framework initially proposed by Lian et al. (2022, 2023). This framework is a two-step process, including 1) guided factor analysis (GFA) and 2) neural convolutive matrix factorization (NCMF) (Figure 2).

3.2.1 Guided Factor Analysis

The GFA aims to decompose the input vocal tract contours $\mathbf{X} \in \mathbb{R}^{2p \times t}$ into articulator specific *factors* $\mathbf{F} = [\mathbf{F}_{jaw}|\mathbf{F}_{tongue}|\mathbf{F}_{lips}|\mathbf{F}_{velum}|\mathbf{F}_{larynx}] \in \mathbb{R}^{2p \times q}$ and *factor scores* $\mathbf{Y} = [\mathbf{Y}_{jaw}|\mathbf{Y}_{tongue}|\mathbf{Y}_{lips}|\mathbf{Y}_{velum}|\mathbf{Y}_{larynx}] \in \mathbb{R}^{t \times q}$, such that $\mathbf{X} = \mathbf{F}\mathbf{Y}^T$. The factors characterize the spatial variation in the shape and position of each articulator, while the factor scores capture temporal variation (Sorensen et al., 2019; Toutios and Narayanan, 2015).

The factors are initially extracted from the contours using Principal Component Analysis (PCA) on the articulator specific vertices. First, the jaw component $\mathbf{F}_{jaw}$ is extracted by setting all non-jaw vertices to zero and performing PCA on the resulting contours, which outputs $\mathbf{\Sigma}_{jaw}$ as in (1), where $\mathbf{Q}_{jaw}$ and $\mathbf{\Lambda}_{jaw}$ are the eigenvector matrix and diagonal variance matrix respectively. The final

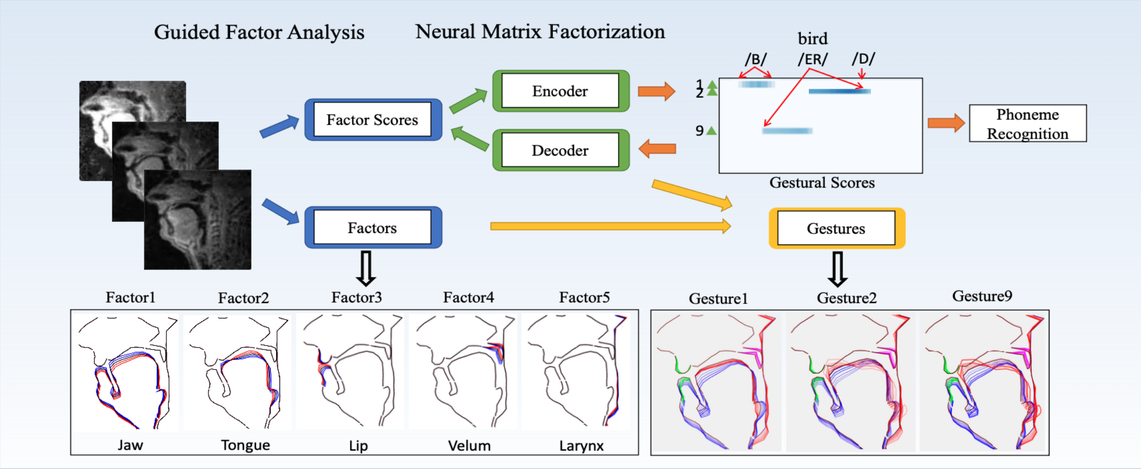

Figure 2: The joint guided factor analysis (GFA) and neural convolutive matrix factorization (NCMF) architecture as originally proposed in Lian et al. (2023). Note: the current study does not incorporate the phoneme recognition module.

$\mathbf{F}_{jaw}$ is extracted as in (2), where $\mathbf{\Lambda}_{union}$ derives from the covariance matrix of concomitant tongue, jaw, and lip motion $\mathbf{X}_{union}^T\mathbf{X}_{union} = \mathbf{Q}_{union}\mathbf{\Lambda}_{union}\mathbf{Q}_{union}^{-1}$, when $union \in \{tongue, jaw, lips\}$. The intuition is that this factor then captures tongue and lip movement that accompanies jaw movement.

$$\begin{aligned}
\mathbf{\Sigma}_{jaw} &= \mathbf{Q}_{jaw}\mathbf{\Lambda}_{jaw}\mathbf{Q}_{jaw}^{-1} \qquad (1)\\
\mathbf{F}_{jaw} &= \mathbf{\Sigma}_{union}\mathbf{Q}_{jaw}\mathbf{\Lambda}_{jaw}^{-1/2} \qquad (2)
\end{aligned}$$

Following extraction of the jaw factor, for the other articulators, where $other \in \{tongue, lips, velum, larynx\}$, the jaw factor is used to extract factors that are independent of the jaw motion. First, the projection matrix $\tilde{\mathbf{X}}$ is obtained as in (3) and (4), where '+' is the Moore-Penrose pseudoinverse.

$$\begin{aligned}
\tilde{\mathbf{X}} &= \mathbf{X}(\mathbf{I} - \mathbf{F}_{jaw}\mathbf{F}_{jaw}^{+}) \qquad (3)\\
&= \mathbf{X} - \hat{\mathbf{X}} \qquad (4)
\end{aligned}$$

PCA is performed on the projection matrix for each articulator in *other* after setting the vertices for all other articulators to zero, resulting in $\tilde{\mathbf{\Sigma}}_{other}$ via (5). The final factors are obtained as in (6). In total five factors are obtained, two for the tongue and one for each of the other articulators.

$$\begin{aligned}
\tilde{\mathbf{\Sigma}}_{other} &= \mathbf{Q}_{other}\mathbf{\Lambda}_{other}\mathbf{Q}_{other}^{-1} \qquad (5)\\
\mathbf{F}_{other} &= \tilde{\mathbf{\Sigma}}_{other}\mathbf{Q}_{other}\mathbf{\Lambda}_{other}^{-1/2} \qquad (6)
\end{aligned}$$

Finally, after factors are obtained for each articulator, factor scores are calculated as in (7).

$$\mathbf{Y} = \mathbf{X}^T\mathbf{F}^{+} = \mathbf{X}^T[\mathbf{F}_{jaw}|\mathbf{F}_{tongue}|\mathbf{F}_{lips}|\mathbf{F}_{velum}|\mathbf{F}_{larynx}]^{+} \qquad (7)$$

The operation in (7) essentially generates a set of coefficients that state how much each factor is active during each point in time in the input. The variance of these coefficients captures temporal variability in the vocal tract shape.

3.2.2 Neural Convolutive Matrix Factorization

The NCMF method employs a sparse autoencoder. A typical autoencoder involves an encoder and a decoder. The encoder function $\mathbf{h} = f(\mathbf{x})$ maps the input to a latent representation and the decoder function $\mathbf{r} = g(\mathbf{h})$ produces a reconstruction of the input using the encoder output, with reconstruction loss used to train the model. A sparse autoencoder adds a sparsity penalty to the

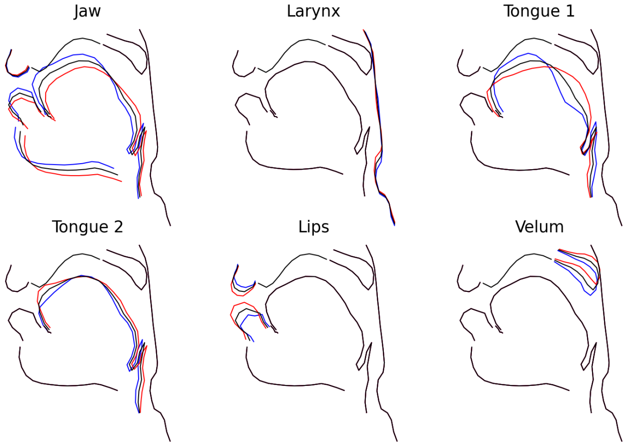

Figure 3: Factors from a single subject showing spatial variation in the five articulators during one utterance.

output of encoder $\mathbf{\Omega}(\mathbf{h})$, as in (8). Sparsity encourages the model to use a sparse set of features to reconstruct the input (Goodfellow et al., 2016). In the context of speech production, sparsity is in line with the notion that only a few gestures are active at a time, allowing also for some degree of overlap between gestures at their temporal edges (Saltzman and Munhall, 1989).

$$\mathbf{L}(\mathbf{x}, g(f(\mathbf{x}))) + \mathbf{\Omega}(\mathbf{h}) \tag{8}$$

The NCMF step aims to parameterize the input factor scores as in (9) (Lian et al., 2022, 2023).

$$\mathbf{Y} \approx \hat{\mathbf{Y}} = \sum_{i=0}^{K-1} \mathbf{H}_i^T \mathbf{W}_i \tag{9}$$

The input to the encoder is the factor scores $\mathbf{Y}$, which then outputs gestural scores $\mathbf{H} \in \mathbb{R}^{D \times t}$, where D is output size of the encoder. The gestural scores indicate which gesture is active when during the input utterance. The decoder takes as input $\mathbf{H}$ and reconstructs the input factor scores as $\hat{\mathbf{Y}}$. $\mathbf{W} \in \mathbb{R}^{K \times D \times q}$ are the decoder weights, with a kernel of size K which acts as the duration of the learned gestures.

The matrix factorization is implemented using an auto-encoder. The encoder consists of two 1D convolutional layers with (in, out, $kernel\ size$) shape of (6, 64, 3) and (64, D, 3) respectively. The first layer acts as a form of “upconvolution”, in which the dimensionality of the input is expanded (from 6 to 64), and the second layer involves reducing the dimensionality from 64 to D. The decoder is comprised of one 1D convolutional layer with shape (D, 6, 21). The encoder and decoder weights are initialized using the Kaiming initialization (He et al., 2015).

L2 loss is used for the reconstruction loss. Sparsity loss is also used over both the time and channel dimensions, such that the model is penalized for encoder activation matrices that are not sparse across the column vectors, as defined in Hoyer (2004). The intuition is that only a few gestures should be active in time, allowing also for some degree of overlap between gestures at their edges. The channel sparsity limits any one gesture from being active for too long. Sparsity loss is defined as in (9) and (10), where $\mathbf{s}(\mathbf{H}_i)$ is vector-wise sparsity. This loss is added to the reconstruction loss using a weighting factors λ_1 and λ_2. Additionally, an entropy constraint (12) is added to force the encoder to use a diverse set of activations during learning, with weighting factor

λ_3. The total loss function is shown in (13).

$$\mathbf{S}_1(\mathbf{H}) = \frac{1}{D}\sum_{i=1}^{D}\mathbf{s}(\mathbf{H}_i) \tag{10}$$

$$\mathbf{S}_2(\mathbf{H}) = \frac{1}{D}\sum_{i=1}^{t}\mathbf{s}(\mathbf{H}_i^T) \tag{11}$$

$$\mathbf{E}(\mathbf{H}) = \frac{1}{D}\sum_{i=1}^{D}(\frac{-\mathbf{S}_1(\mathbf{H}_i)}{\sum_{i=1}^{D}\mathbf{S}_1(\mathbf{H}_i)}\log(\frac{\mathbf{S}_1(\mathbf{H}_i)}{\sum_{i=1}^{D}\mathbf{S}_1(\mathbf{H}_i)})) \tag{12}$$

$$\mathbf{L} = \mathbb{E}_{\mathbf{X}}[\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 + \lambda_1\mathbf{S}_1(\mathbf{H}) + \lambda_2\mathbf{S}_2(\mathbf{H}) + \lambda_3\mathbf{E}(\mathbf{H})] \tag{13}$$

Following the training of the autoencoder, the decoder weights $\mathbf{W}$ are used to obtain the final gestures $\mathbf{G}$ as in (11).

$$\mathbf{G} = \mathbf{W}\mathbf{F}^T \tag{14}$$

The intuition is that the sparse decoder weights encode a sparse set of temporal units, which, when multiplied with the original factors, combine the factors of multiple articulators into interpretable spatio-temporal units.

3.2.3 Phoneme Classification

The output of the NCMF pipeline is a set of gestures $\mathbf{G}$ pooled across speakers, with each gesture representing a spatiotemporal unit of vocal tract coordination, and a set of encoder activation matrices $\mathbf{H}$, which are akin to gestural scores. Given the goal of assessing the extent to which the model is encoding meaningful phonological contrasts, both the gestures and the gestural scores will be submitted to phoneme classifiers. Each classifier is tasked with assessing the extent to which the gestures and gestural scores can be used to discriminate between phonemes. Additionally, context labels will be added to the models as a feature to test if contextual information aids in phonemic discriminability, suggesting the model better encodes phonemic information in certain contexts. For the gestural scores, multinomial logistic regression is used, while for the gestures a convolutional neural network (CNN) classifier is used.

The CNN classifier is comprised of two 1D convolutional layers, each with dropout ($p = 0.3$) and ReLU activation in-between. Following this, two fully connected (FC) layers are used, with the last layer having an output the size of the number of phonemes for each the vowel and consonants sets. The configurations for each layer are shown in Table 1. Phoneme classification is done separately for vowels and consonants.

Layer	(in, out, kernel size)
Conv1	(230,128,3)
Conv2	(128,64,3)
FC1	(64,128)
FC2	(128, # classes)

Table 1: CNN Classifier Architecture

4 Experiments

4.1 Dataset

The dataset consists of 10 native American English speakers (5F, 5M) subset from a larger 75 speaker rtMRI dataset (Lim et al., 2021). Data from one speaker was excluded due to the distribution of their factors and factor scores falling beyond two standard deviations of the group mean and standard deviation, leaving 9 speakers (5F, 4M). Speakers of other varieties of English were excluded to control for potential differences in language-specific articulatory settings. All speakers read from the same set of stimuli, which were comprised of phonetically rich sentences, with each sentence repeated twice. Spontaneous speech was not included so as to balance the data across speakers. The vocal tract was imaged in the midsagittal plane. In addition, simultaneous audio was collected at a sampling rate of 20kHz. The rtMRI video frames were segmented by the major articulators, including the tongue, jaw, lips, velum, and larynx, using the segmentation method proposed in Anonymous (2024), with 83 frames per second. The resulting vocal tract contours are comprised of $2p = 230$ points for each frame. Data from all 9 speakers was used during the training and test phases. The grandfather passage was used for the test set and all other stimuli used for training. We used MFA to extract phoneme level alignments from the audio (McAuliffe et al., 2017). Before the guided factor analysis, the vocal tract contours were split into segments of 100 frames, resulting in each sample being of size (230,100).

4.2 Implementation Details

The overall pipeline has three steps: (1) guided factor analysis to obtain factors and factor scores per segment; (2) neural convolutive matrix factorization to train the decoder weights and obtain the final gestures; and (3) gesture analysis and phoneme classification using multinomial logistic regression and the CNN classifier.

4.2.1 Neural Convolutive Matrix Factorization

We follow the general NCMF implementation as in Lian et al. (2022, 2023). The guided factor analysis is first done offline to obtain both the factors and factor scores for each input segment. The input to the factor analysis is the segmented rtMRI contours, with the contours centered on zero on a by-item basis, i.e. the mean for each item is subtracted from the data. The resulting factor scores are fed as input to the encoder. The autoencoder is trained for 1,000 epochs with the Adam optimizer (Kingma and Ba, 2014), an initial learning rate of 1e-5, and a batch size of 4. The parameter D, which determines the size of both the encoder and decoder outputs, is set at 15. The sparsity weighting factors are set at $\lambda_1 = 0.05$ and $\lambda_2 = 0.04$ and the entropy weighting factor is set at $\lambda_3 = 0.1$. The learning rate is decayed by a rate of 0.95 every 10 epochs. The model was trained using an NVIDIA A40 GPU and took roughly one hour. The resulting encoder weights are extracted and used to compute the final gestures. For each speaker, we use the phoneme level alignments from MFA to segment the encoder activation matrices for each segment into phoneme level activations. The channel sparsity is used to determine which gestures are active during that phoneme, resulting in gesture and phoneme label pairs for all phonemes across the test set of speakers. To allow for cross-speaker comparisons, all of the gestures are projected onto the mean vocal tract shape across all of the test set contours by adding this mean shape to each individual gesture[1]. An example can be seen in Figure 5, where the original vocal tract shape is in A and the mean shape is in B. The example gesture appears to show a coronal constriction. The constriction is maintained in version B, though the movement of the tongue tip is actually clearer.

4.2.2 Gesture Analysis

The gestural analysis uses two sets of features: 1) indices of the gestural score rows that are active during any given phoneme and 2) the gestural units. For the gestural units, sparsity is used to indicate which rows are active during that phoneme. All of the gestures corresponding to highly activated rows are extracted and concatenated across the time axis, where each gesture is of size (230,21), 230 is the number of points and 21 is time. Highly active rows are defined as those with an average sparsity below 0.3 during the relevant phoneme interval[2]. This set of concatenated gestures

[1] The mean vocal tract shape here is the mean of the 230 contour points across all speakers, including the points corresponding to the tongue, jaw, lips, palate, velum, and pharyngeal wall, i.e. all points shown in Figures 4 and 5.

[2] The value of 0.3 was determined by visually inspecting the gestural scores and experimenting with different values to see which rows are extracted. A sparsity of 0.3 was found to extract rows that are visually active in the gestural scores.

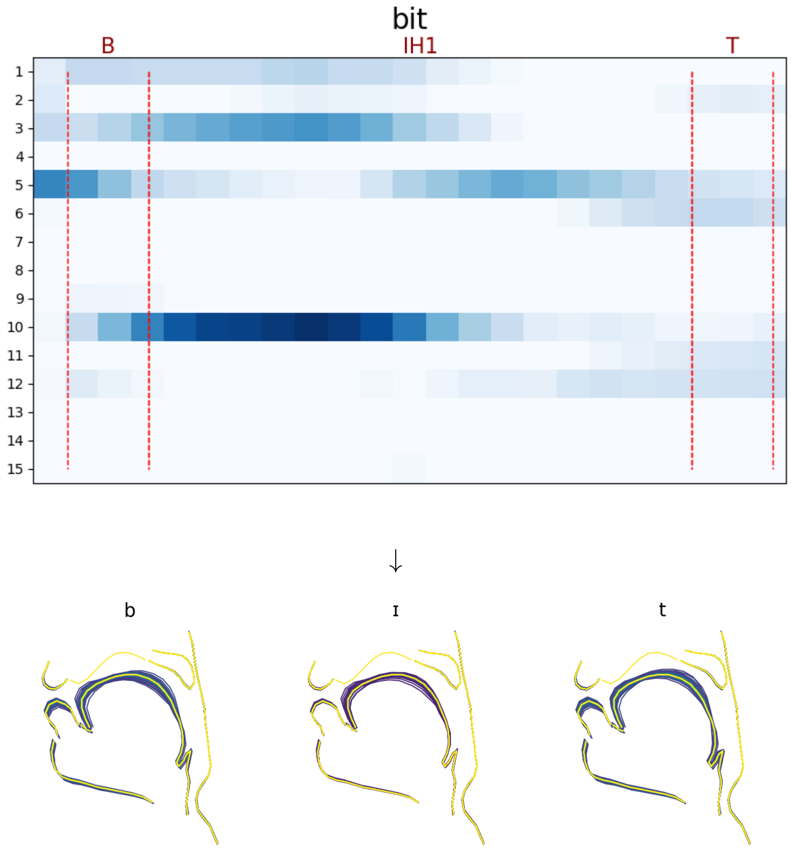

Figure 4: One example encoder activation matrix, akin to a gestural score, for the word "bit" (top). Concatenated gestural units extracted for each phoneme in the word "bit" after being projected onto the mean vocal tract shape (bottom).

are added to a matrix of shape (230,315) (where 315 is the maximum length for the time dimension, i.e. 15 gestures × a kernel size of 21), where non-active time frames are set to zero. This acts as a form of padding to keep all items of the same size before being fed to the classifier. An example can be seen in Figure 4 with the word "bit" from one speaker. The gestural score **H** is on the left with red boundaries indicating phoneme alignments. Darker colors indicate higher activation, e.g. rows 3, 5, and 10 are highly activated during the vowel. On the right are the concatenated gestural units extracted for each of the three phonemes, where each of these is a 230 × 315 matrix. For example, the gesture for [b] contains the individual gestural units at indices 1,2,5, and 10, as these rows are active during the interval for this phoneme.

The CNN classifier was trained using the Adam optimizer, an initial learning rate of 1e-3, and a batch size of 16. The learning rate is decayed by a rate of 0.90 every 20 epochs. Early stopping was implemented using the validation loss with patience set to 20 epochs. The metric used for assessment of phoneme classification is F1[3]. 5 speakers were used for training, 2 for validation, and 2 for the test set, with F1 scores reported on the held-out test set.

In addition to analysis using the actual gestures, the encoder weight indices were used to assess if the model encodes a given phoneme similarly across different contexts and within contexts versus other phonemes. For this, multinomial logistic regression models were fit to predict phoneme label given an array of binary activation indices. For example, given 15 possible active rows in the matrix

[3] F1 score better accounts for imbalance among the class distributions compared to a simple accuracy metric and is more suitable given some phonemes are more frequent than others. Further details can be found here: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

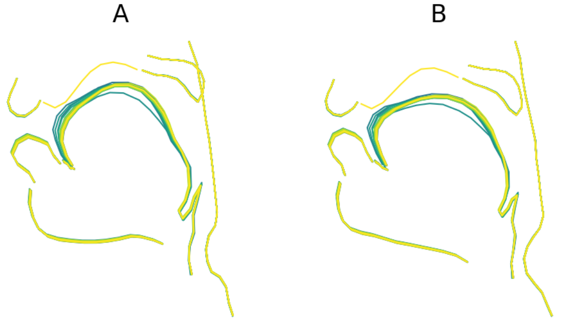

Figure 5: One gestural unit (A) before and (B) after being projected onto the mean vocal tract shape for the test set. The before shape is the mean for that item.

at any given point in time, indices for rows that show high activation during that time are coded as 1, while inactive rows are coded as zero, resulting in a 15-dimensional array. The surrounding context is also used as a feature to the model, with models being fit with and without this feature to determine if phonemes are encoded similarly in certain contexts. Both left and right context were coded by whether the target phoneme was adjacent to a consonant, vowel or silence, e.g. C_C or #_C, where # indicates silence.

5 Experiment Results

The results from the models are described below, organized by gestural scores (encoder activation matrices) and the gestures (the spatio-temporal units).

5.1 Gestural Scores

The results from the multinomial logistic regression models are shown in Table 2, with both accuracy and ROC/AUC scores reported. ROC/AUC is a better overall metric of performance as it accounts for class distributions, where 0.5 indicates chance performance. Performances are shown for models fit with and without context as a feature and for models fit with random features, where binary 15-D vectors were randomly generated. For vowels (9 classes), we see that the model fit without context is not much better than random, while the model with context features shows a much larger improvement. For consonants (24 classes), we see a similar story, though the model without context features shows a sizeable improvement in performance. To determine if the difference between models fit with and without context as a feature is significant ten-fold cross-validation was performed on both models to create a model performance array. Following this, paired t-tests were performed on the arrays to determine if they differ significantly. For both the consonant and vowel models, the improvement when using context was significant ($p < 0.001$). Interestingly, the vowel models without context were not significantly better than using random features, while for the consonant models they were.

Given that fitting a model on all of the phoneme classes presents a potentially overly difficult multiclass problem given the large number of classes (e.g. $k = 24$ for consonants), models were also fit using a smaller subset of labels using natural classes. For consonants includes labial, coronal, and dorsal, and for vowels high vs. low and front vs. back. The results of these models are reported in Table 3. For vowels, both sets of models show better performance when using context, though not much of an improvement. For consonants, while the model without context shows better accuracy it has a lower ROC/AUC score, suggesting the model without context may be over-relying on predicting the majority class (coronals). Similarly, paired t-tests were performed on cross-validated performance arrays to assess if model performances were significantly different. As with models fit with larger classes, both consonant models were significantly better than the model fit on random features, and the improvement with context added was also significant ($p < 0.05$). For vowels, only the models fit with context were significantly better than random ($p < 0.01$).

Model	Context	Accuracy	ROC/AUC	Better than random?
Vowels	Random	0.27	0.48	
	Without	0.28	0.58	N
	With	0.35	0.65	Y
Consonants	Random	0.09	0.51	
	Without	0.16	0.61	Y
	With	0.23	0.76	Y

Table 2: Accuracy and ROC/AUC comparisons for models fit with and without context as a feature and using random features for all vowel and consonant classes. The far right column indicates if the model performed significantly better than a model fit with random features.

Model		Context	Accuracy	ROC/AUC	Better than random?
Vowels	High vs. Low	Random	0.55	0.49	
		Without	0.61	0.58	N
		With	0.62	0.61	Y
	Front vs. Back	Random	0.47	0.46	
		Without	0.61	0.56	N
		With	0.64	0.59	Y
Consonants		Random	0.46	0.49	
		Without	0.65	0.60	Y
		With	0.63	0.67	Y

Table 3: Accuracy and ROC/AUC comparisons for models fit with and without context as a feature and using random features for vowel and consonant major feature classes. The far-right column indicates if the model performed significantly better than a model fit with random features.

The logistic regression results suggest that to some extent the patterns of row activations of the encoder do encode phonological distinctions, particularly when given contextual information. To get a sense of what each row in the gestural score is encoding when active, for each phoneme the relative proportion of each row being active during that phoneme is plotted for vowels (Figure 6) and consonants (Figure 7) in a fixed context. The vowel context is C_C and the consonants context is V_V. If certain rows encode particular articulatory movements that are shared across certain phonemes, it is expected that those phonemes would show similar activation patterns. Figure 6 shows the same plot twice, with the plot on the left coding vowels by backness, and the plot on the right coding them by height. No clear, discernable trend can be seen across the different rows in terms of vowels involving similar constrictions having similar patterns of activation. The plot coded by backness (left) shows similar patterning between [ʌ]/[u] and [ɑ]/[ɔ] in rows 3 and 6 and 5 and 14 respectively. While the latter suggests encoding of pharyngeal constriction in these rows (though [ʌ] would also be expected to be here), the former is not as clear. Consonant patterns of activation are shown in Figure 7 and consonants are coded by coronal, labial, and dorsal constrictions[4]. For consonants, there is not much consistency. Some degree of separation can be seen between coronals and labials in rows 5 and 12, suggesting potentially tongue front movements being more likely to be encoded by these rows, though higher frequency during [k] here refutes this. Across the vowels and consonants, it can be seen that rows 4,7,8,9, and 15 show very little frequency activation, implicating that the model is primarily only using the remaining 10 rows during learning.

5.1.1 Gestures

The results from the CNN classifiers fit using the gestural units are shown in Table 4. We can see for both vowels and consonants the models fit without context features perform worse than random, and models that use context show a sizeable proportional increase in performance. To assess if the models fit using context features were a significant improvement over the models using random

[4] [h] is included with dorsals for convenience.

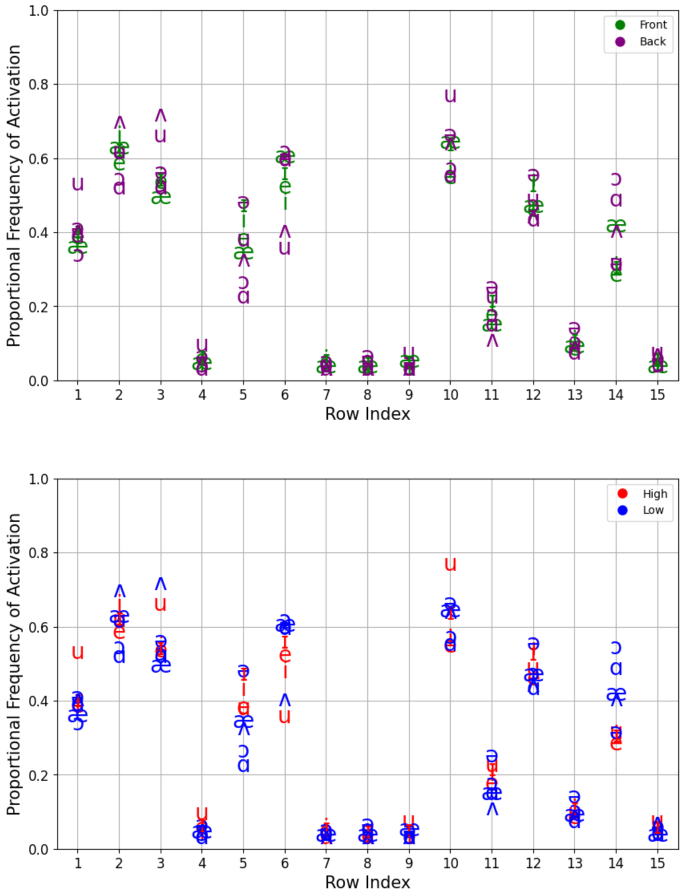

Figure 6: Proportional frequency of a given encoder row being active during each vowel phoneme. Vowels are coded by backness (top) and height (bottom).

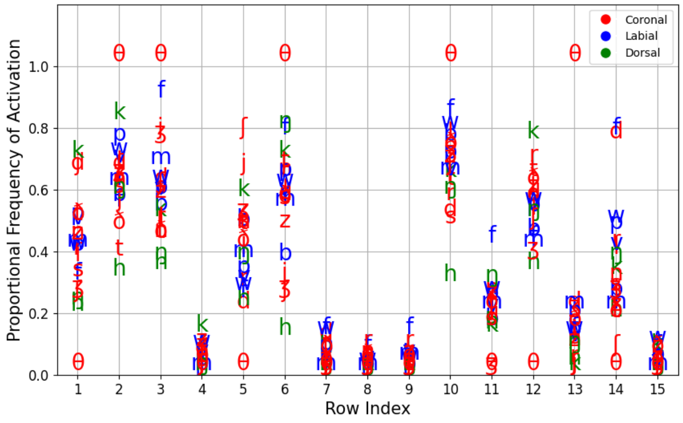

Figure 7: Proportional frequency of a given encoder row being active during each consonant phoneme. Consonants are coded as coronal, labial, and dorsal.

features the bootstrap sampling method was used. In this method, 50 random predicted labels were taken from the two models being compared and the ground truth labels, and F1 is computed for each model for this small sample. This was done for 1,000 iterations with the F1 scores collected for each model. Paired t-tests were conducted on the performance arrays to test for significant differences between the two performance distributions. For both the vowel and consonant models, the models fit using context performed significantly better than the models fit with random features ($p < 0.05$). Nonetheless, despite being better than random the F1 scores are still quite poor. Figure 8 shows a confusion matrix for the vowel model fit using context features. As is clear, while the model may show better than random performance, it mostly relies on predicting the majority class [ɪ]. It performs slightly better with the vowel [i] than the others. This demonstrates that the recovered gestures poorly discriminate vowels; the confusion matrix for the consonant model (not shown) is very similar, with the more frequent phonemes [t] and [n] being over-predicted.

Model	**Context**	**F1**	**Better than random?**
Vowels	Random	0.09	
	Without	0.07	N
	With	0.14	Y
Consonants	Random	0.03	
	Without	0.02	N
	With	0.06	Y

Table 4: F1 score comparisons for models fit with and without context as a feature and using random features for all vowel and consonant classes. The far right column indicates if the model performed significantly better than a model fit with random features.

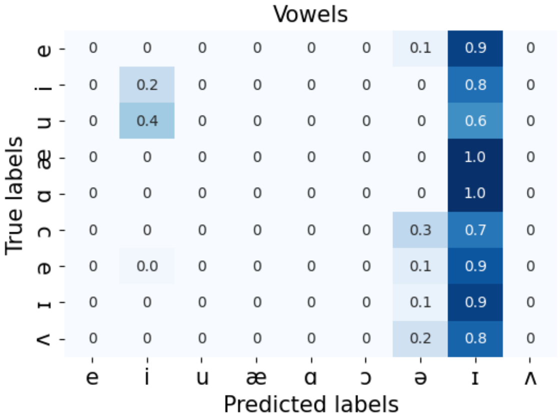

Figure 8: Confusion matrix for the large vowel model when using context.

The gestures were also fit on classifiers that used a smaller number of class labels, using the same labels as in the gestural score models. The results are reported in Table 5. Interestingly, across both the vowel and consonants models, the F1 scores are nearly the same as the models fit on random features. Only the vowel height model with context shows a noticeable improvement over the model with random features. This is surprising given that some of the models tasked with classifying more labels showed significant improvement over the random models. However, as Figure 8 highlights, this performance improvement may rely on the model showing some improvement on one or two phonemes, while this second group of models shows that the gestures truly lack discriminability. In addition, the bootstrap sampling revealed that none of the secondary models showed a significant improvement over the random models.

6 Discussion

This study attempted to decompose kinematic data of the vocal tract into a set of gestural units and gestural scores and to then assess the extent to which both of these capture meaningful phonological distinctions employed in American English. The discussion will focus on the ability of the model

Model		Context	F1	Better than random?
Vowels	High vs. Low	Random	0.42	
		Without	0.42	N
		With	0.51	N
	Front vs. Back	Random	0.44	
		Without	0.44	N
		With	0.44	N
Consonants		Random	0.26	
		Without	0.11	N
		With	0.26	N

Table 5: F1 score comparisons for models fit with and without context as a feature and using random features for vowel and consonant major feature classes. The far right column indicates if the model performed significantly better than a model fit with random features.

to produce gestural scores and gestures that encode phonological distinctions. Limitations of the approach, particularly in the model choice and architecture and data preprocessing, along with potential directions for future research are discussed as well.

6.1 Encoding of phonological information

The results from the models fit on the active row indices suggest that to some extent phonological contrasts are encoded in the model. The regression models fit using these features and using context were significantly better than models that used random features. This suggests that, using the temporal information present in the factor scores, the model can identify some re-occurring patterns that distinguish vocal tract tasks in certain contexts. In the models that were fit to predict phoneme labels, it is worth pointing out that the consonant model outperformed the vowel model, despite having to predict nearly three times as many classes (9 vs. 24). This is unsurprising given the well-known difficulty in distinguishing vowels by tongue position (Esling, 2005; Smith, 2018). It is also likely that small differences in global tongue shaping during vowels were not precisely captured within the factors.

The models that were tasked with predicting the smaller set of labels, i.e. height and backness for vowels and primary place of constriction for consonants, also showed significantly better performance when using active rows indices and context as features than when using random features. However, looking at the ROC/AUC scores the latter group of models in the comparison to the first group, one can notice that there isn't much of an improvement, hinting that the active row features offers minimal phonological information beyond random features. If this were otherwise, we would expect a proportional improvement in model performance relative to the multiple-fold decrease in classes. Nonetheless, given that the model was only given coefficients indicating temporal variation, the results do suggest that at some level such temporal information does encode differences between vocal tract tasks and there is some consistency in this variation across speakers. Additionally, the fact that context improved model performance indicates that such patterns are more distinguishable when performed in similar contexts, e.g. C_C for consonants. Naturally, phonological contrast is expected to exist in cases when the task varies meaningfully when performed in the same context.

Given the somewhat promising results from the gestural score models, an attempt was made to assess if certain articulatory aspects were encoded at individual rows within the encoder. Impressionistically, no concrete, discernible patterns were found, as seen in Figures 5 and 6. Some back vowels have a tendency to pattern together, as do labials and coronals in some rows, though this is not entirely consistent. This indicates that the model does not encode individual articulatory movements at the level of individual rows and that rather it uses the combination of multiple rows' activations in reconstructing the factor scores. This differs from what was shown in earlier versions of this model where it was shown across tokens that the same row was active for vowels and labials/coronals (Lian et al., 2022, 2023). This may be due to the previous analyses only showing a few tokens, without carrying out quantitative analysis of the row activations across various contexts and across multiple speakers, highlighting the necessity of doing such quantitative analyses and probing of these models more thoroughly.

The results from the models fit using the gestural units as features showed quite poor performance. Surprisingly, while the models tasked with predicting more labels showed significantly better performance than when using random features, the models using fewer labels did not. This likely results from the models' over-reliance on predicting the majority class and only being somewhat more accurate in predicting a couple other classes. This poor performance highlights the difficulty in modeling both spatial and temporal variation across speakers. Crucially, the autoencoder was only fit using the temporal information, while the factors, which capture spatial variation, are computed on a by-item basis, introducing considerable variation into the final gestural units. The guided factor analysis in its current form does not incorporate cross-speaker variation and as such, variations in constriction location and implementation likely introduce a degree of noise into the final gestures that obscures meaningful contrasts. Note that the factors are of dimension $2p = 230$, where each coefficient indicates variation in that point during the 100-frame sequence. While the exact points that correspond to the different articulators are the same across speakers, the exact points which correspond to tongue sub-units, e.g. the tongue tip or the tongue body will differ. For example, a tongue tip constriction for one speaker may be captured primarily in points 3-6, while for another they may be shifted more anterior to points 5-8, and similarly for other tongue regions. As a result, distinctions in constriction tasks are obscured heavily by this variability.

6.2 Limitations

There are a number of limitations in the current approach. First, given the joint nature of the entire pipeline, training the entire model is slightly inefficient, as the factors and factor scores have to be trained offline with the CNMF then being trained separately (Lian et al., 2023). This makes model adjustments, e.g. the number of frames used as input, difficult to implement as one has to re-do the entire pipeline as opposed to adjusting a single parameter of the CNMF model once. Second, the guided factor analysis output creates a number of complications for training an autoencoder. Typically, CNN-based architectures rely on the training data being normalized beforehand, otherwise model performance suffers. The factors and factor scores cannot be normalized effectively as to do so completely removes the variability within them, making their values almost meaningless. Additionally, while the contours are centered on zero before going through the PCA on a by-item basis, normalizing the contours across speakers removes too much variability that the resulting factors and factor scores show little to no variation in articulator movements.

Third, the number of frames given to the model is likely much too large, as 100 frames equates to nearly 1 second of articulator movements during read speech. The factors won't be able to capture enough of the variation in these movements to effectively be able to distinguish gestures on an appropriate timescale. An ideal number of frames is likely closer to 20-30. However, using a smaller number of frames introduces greater variation into the factors and thus the cross-speaker variability becomes too great for the CNMF model to learn effectively. As mentioned above, normalization of the factors is not feasible here to reduce this variability. Using a large number of frames results in the factor scores not capturing more fine-grained temporal variation as well. For example, in a sequence with coronal and dorsal consonants, the lingual variation in a palatal constriction, i.e. during [i], may be obscured, resulting in low activation for the tongue factor at the time of the [i] in the factor score.

Fourth, the current GFA implementation fits a PCA on a by-item basis, introducing variability within and across speakers. A more ideal approach is a 3-way factor analysis (Harshman et al., 1977), where speaker variation is also included in the model. Such models have been shown to be viable approaches to cross-speaker articulatory modeling (Serrurier et al., 2019; Kochetov et al., 2023). However, crucially, such models rely solely on modeling spatial variation and the resulting components capture spatial variation in the same way the factors in the current approach do, excluding temporal variation. Fitting an autoencoder with such factors could result in the model learning a distribution over these spatial components, though this would not result in interpretable gestural units that show constrictions being formed.

6.3 Future directions

There are a number of ways that the current work can be extended. The current study has highlighted many of the difficulties in incorporating articulatory data from multiple speakers into a generative model. There are two ways forward in light of this. First, future work could focus more on using data from a single speaker, as was done in Lian et al. (2022) using the EMA mngu0 corpus (Richmond et al., 2011). Using a single speaker allows one to control for cross-speaker differences in vocal morphology, implementation, and within-speaker variability that may obscure meaningful distinctions. Second, the most ideal model is likely one that is trained directly on the articulatory data, whether it is EMA, ultrasound, or rtMRI kinematic data, as opposed to training on the output of some intermediate model. Training with such data would tie the model's objective function directly to its ability to reproduce articulatory data that is closer to the actual vocal tract movements, allowing for more transparent representations to be learned. Third, other variations of autoencoders or other generative models could be employed here. Some success has been found in using generative adversarial networks (GAN) to generate articulatory kinematics, though it has not yet been tested whether the generator network in this models develops meaningful representations of the input (Beguš et al., 2023b,a). Masked autoencoders have been highly successful in computer vision and may be applicable here (He et al., 2022). In these models, parts of the input are masked before being fed to the encoder, forcing the decoder to then reconstruct the original input by predicting the masked portions. In speech production, coarticulation and gestural blending are robust, and as such, a masked autoencoder would have to learn to predict such behavior, potentially leading to highly informative representations.

7 Conclusion

This study employed a joint GFA and NCMF approach to decompose rtMRI articulatory data into a set of gestural scores and gestures, and then probed the extent to which these representations captured phonological distinctions used in language. It was found that the gestural scores did somewhat encode phonological information and that this encoding was highly dependent on contextual information. The gestures that were learned were minimally phonologically informative and this is attributed to considerable cross-speaker variability obscuring such distinctions. The results present some promise in using deep learning for learning meaningful articulatory representations directly from kinematic data and suggestions for future directions in this area were offered.

References

Anonymous (2024). Multimodal segmentation for vocal tract modeling. Preprint.

Beguš, G., Lu, T., Zhou, A., Wu, P., and Anumanchipalli, G. K. (2023a). Ciwagan: Articulatory information exchange. *arXiv preprint arXiv:2309.07861*.

Beguš, G. and Zhou, A. (2021). Interpreting intermediate convolutional layers of cnns trained on raw speech. *CoRR abs/2104.09489. URL: https://arxiv. org/abs/2104.09489*.

Beguš, G., Zhou, A., Wu, P., and Anumanchipalli, G. K. (2023b). Articulation gan: Unsupervised modeling of articulatory learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Bizzi, E. and Cheung, V. C. (2013). The neural origin of muscle synergies. *Frontiers in computational neuroscience*, 7:51.

Bizzi, E., Cheung, V. C., d'Avella, A., Saltiel, P., and Tresch, M. (2008). Combining modules for movement. *Brain research reviews*, 57(1):125–133.

Browman, C. P. and Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, 49(3-4):155–180.

Byrd, D. and Krivokapić, J. (2021). Cracking prosody in articulatory phonology. *Annual Review of Linguistics*, 7(1):31–53.

Chartier, J., Anumanchipalli, G. K., Johnson, K., and Chang, E. F. (2018). Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex. *Neuron*, 98(5):1042–1054.

Esling, J. H. (2005). There are no back vowels: The larygeal articulator model. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 50(1-4):13–44.

Faytak, M. D. (2018). *Articulatory uniformity through articulatory reuse: insights from an ultrasound study of Sūzhōu Chinese*. University of California, Berkeley.

Gick, B. (2016). Ecologizing dimensionality: Prospects for a modular theory of speech production. *Ecological psychology*, 28(3):176–181.

Gick, B. and Stavness, I. (2013). Modularizing speech.

Goldstein, L., Pouplier, M., Chen, L., Saltzman, E., and Byrd, D. (2007). Dynamic action units slip in speech production errors. *Cognition*, 103(3):386–412.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Guenther, F. H. (2016). *Neural control of speech*. Mit Press.

Harshman, R., Ladefoged, P., and Goldstein, L. (1977). Factor analysis of tongue shapes. *The Journal of the Acoustical Society of America*, 62(3):693–707.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

Hickok, G. and Poeppel, D. (2007). The cortical organization of speech processing. *Nature reviews neuroscience*, 8(5):393–402.

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(9).

Jordan, M. I. (2018). Motor learning and the degrees of freedom problem. In *Attention and performance XIII*, pages 796–836. Psychology Press.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kochetov, A., Savariaux, C., Lamalle, L., Noûs, C., and Badin, P. (2023). An mri-based articulatory analysis of the kannada dental-retroflex contrast. *Journal of the International Phonetic Association*, pages 1–37.

Konkle, T. and Alvarez, G. A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. *Nature communications*, 13(1):491.

Lian, J., Black, A. W., Goldstein, L., and Anumanchipalli, G. K. (2022). Deep neural convolutive matrix factorization for articulatory representation decomposition. *arXiv preprint arXiv:2204.00465*.

Lian, J., Black, A. W., Lu, Y., Goldstein, L., Watanabe, S., and Anumanchipalli, G. K. (2023). Articulatory representation learning via joint factor analysis and neural matrix factorization. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Lim, Y., Toutios, A., Bliesener, Y., Tian, Y., Lingala, S. G., Vaz, C., Sorensen, T., Oh, M., Harper, S., Chen, W., et al. (2021). A multispeaker dataset of raw and reconstructed speech production real-time mri video and 3d volumetric images. *Scientific data*, 8(1):187.

Martin, K., Gauthier, J., Breiss, C., and Levy, R. (2023). Probing self-supervised speech models for phonetic and phonemic information: a case study in aspiration. *arXiv preprint arXiv:2306.06232*.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Mugler, E. M., Tate, M. C., Livescu, K., Templer, J. W., Goldrick, M. A., and Slutzky, M. W. (2018). Differential representation of articulatory gestures and phonemes in precentral and inferior frontal gyri. *Journal of Neuroscience*, 38(46):9803–9813.

Parrell, B., Lammert, A. C., Ciccarelli, G., and Quatieri, T. F. (2019). Current models of speech motor control: A control-theoretic overview of architectures and properties. *The Journal of the Acoustical Society of America*, 145(3):1456–1481.

Poliva, O., Venezia, J., Brodbeck, C., and Hickok, G. (2024). Phoneme processing.

Ramanarayanan, V., Goldstein, L., and Narayanan, S. S. (2013). Spatio-temporal articulatory movement primitives during speech production: Extraction, interpretation, and validation. *The Journal of the Acoustical Society of America*, 134(2):1378–1394.

Ramanarayanan, V., Van Segbroeck, M., and Narayanan, S. S. (2016). Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories. *Computer speech & language*, 36:330–346.

Richmond, K., Hoole, P., and King, S. (2011). Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. In *Twelfth Annual Conference of the International Speech Communication Association*.

Saltzman, E. L. and Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological psychology*, 1(4):333–382.

Serrurier, A., Badin, P., Lamalle, L., and Neuschaefer-Rube, C. (2019). Characterization of inter-speaker articulatory variability: A two-level multi-speaker modelling approach based on mri data. *The Journal of the Acoustical Society of America*, 145(4):2149–2170.

Smith, C. M. (2018). *Harmony in gestural phonology*. PhD thesis, University of Southern California.

Sorensen, T., Toutios, A., Goldstein, L., and Narayanan, S. (2019). Task-dependence of articulator synergies. *The Journal of the Acoustical Society of America*, 145(3):1504–1520.

Toutios, A. and Narayanan, S. S. (2015). Factor analysis of vocal-tract outlines derived from real-time magnetic resonance imaging data. In *ICPhS*.