Automatic Data-Driven Learning of Articulatory Primitives from Real-Time MRI Data using Convolutive NMF with Sparseness Constraints

Vikram Ramanarayanan, Athanasios Katsamanis, and Shrikanth Narayanan

Signal Analysis and Interpretation Lab (SAIL), Ming Hsieh Department of Electrical Engineering,
University of Southern California, Los Angeles, CA

`vramanar@usc.edu, <nkatsam,shri>@sipi.usc.edu`

Abstract

We present a procedure to automatically derive interpretable dynamic articulatory primitives in a data-driven manner from image sequences acquired through real-time magnetic resonance imaging (rt-MRI). More specifically, we propose a convolutive Nonnegative Matrix Factorization algorithm with sparseness constraints (cNMFsc) to decompose a given set of image sequences into a set of basis image sequences and an activation matrix. We use a recently-acquired rt-MRI corpus of read speech (460 sentences from 4 speakers) as a test dataset for this procedure. We choose the free parameters of the algorithm empirically by analyzing algorithm performance for different parameter values. We then validate the extracted basis sequences using an articulatory recognition task and finally present an interpretation of the extracted basis set of image sequences in a gesture-based Articulatory Phonology framework.

Index Terms: real-time MRI, gestures, Nonnegative Matrix Factorization, sparse representations.

1. Introduction

Extracting interpretable representations from raw articulatory data is critical for better understanding, modeling and synthetic reproduction of the human speech production process. If we view the speech planning and execution mechanism in humans as a control system, we would like to understand the properties and characteristics of the system such as the goals and constraints of the plan and the architecture of the system among others. For this we would need an understanding of how these characteristics are specified or represented in inputs and outputs of the system, i.e., so-called primitive representations. Recently there have been studies in the literature that have attempted to further our understanding of primitive representations in biological systems using ideas from linear algebra and sparsity theory. For example, studies have suggested that neurons encode sensory information using only a few active neurons at any point of time, allowing an efficient way of representing data, forming associations and storing memories [1]. It has been also been argued that for human vision the spatial visual receptive fields in the brain might be employing a sparse and overcomplete basis for representation [1], and quantitive evidence has been put forth for sparse representations of sounds in the auditory cortex [2]. However, not many computational studies have been conducted into uncovering the primitives of speech production.

There are two broad (but not mutually exclusive) approaches to address this problem of formulating representations of speech production - knowledge-driven and data-driven. There have been many compelling attempts at, and accounts of, knowledge-driven formulations in the linguistics literature. An example is the framework of Articulatory Phonology [3] which theorizes that the act of speaking is decomposable into units of vocal tract actions termed “gestures.” So in this framework, a simple set of linguistically-meaningful primitives are so-called ‘tract variables’ (or a set of constriction degrees and locations); this is one possible basis set that can be used to characterize the gestural lexicon of a language used in speech planning. In this paper, however, we choose to adopt the less-explored data-driven approach to extract sparse primitive representations from real-time magnetic resonance imaging (rt-MRI) data. We view this as a first step towards our ultimate goal of bridging knowledge-driven and data-driven approaches. rt-MRI is a recently-developed medical imaging technique that has been successfully used to obtain simultaneous observations of dynamic vocal tract shape deformations in the midsagittal plane along with synchronized audio speech data [4]. It can provide a complete view of all vocal tract articulators as compared to other imaging technologies such as ultrasound, electromagnetic midsagittal articulography (EMMA), etc., thus affording useful data for articulatory modeling and large-scale phonetics research. The rt-MRI data of vocal tract movements hence offer a rich source of information for deriving articulatory primitives that underlie speech production.

Modeling data vectors as sparse linear combinations of basis elements is a general computational approach (termed variously as dictionary learning or sparse coding or sparse matrix factorization depending on the problem formulation) which we will use to solve our problem of seeking articulatory representations. These methods have been successfully applied to a number of problems in signal processing, machine learning, and neuroscience, and are in general applicable and useful for our task since we would like to obtain primitive articulatory representations, weighted combinations of which can be used to synthesize any temporal sequence of articulatory movements. More specifically, we say that a signal $\mathbf{x}$ in $\mathbb{R}^m$ admits a sparse approximation over a basis set of vectors or ’dictionary’ $\mathbf{D}$ in $\mathbb{R}^{m \times k}$ with k columns referred to as ’atoms’ when one can find a linear combination of a small number of atoms from $\mathbf{D}$ that is as “close” to $\mathbf{x}$ as possible (as defined by a suitable error metric) [5]. Note that sparsity constraints can be imposed over either the dictionary or the coefficients of the linear combination (or ’activations’) or both. In this paper, since one of our main goals is to extract *interpretable*[1] basis or dictionary elements from observed articulatory data, we focus on matrix fac-

This work was supported by NIH Grant R01 DC007124-01.

[1] By interpretable we mean a basis that a trained speech researcher can assign linguistic meaning to; for example, a basis of sequences of rt-MRI images of the vocal tract.

Copyright © 2011 ISCA

torization techniques such as Nonnegative Matrix Factorization (NMF)[2] and its variants [10, 11, 7, 8] with sparsity constraints imposed on the *activation* matrix (but none on the basis matrix since not constraining the basis image sequences would allow them a greater degree of interpretability). In addition, we would like to find a factorization such that only a few basis functions are "activated" at any given point of time, i.e., a sparse activation matrix.

The rest of this paper is organized as follows: we give a brief description of the data used in Section 2 followed by a detailed layout of the problem formulation in Section 3. We next present a validation and interpretation of the representations extracted by our approach in Section 4 followed by a discussion of future work.

2. Data

For this study we used the MRI-TIMIT database recently collected by our lab which currently consists of read speech data (MRI image sequences and synchronous noise-cancelled audio) collected from 4 native (2 male and 2 female) American English speakers while lying supine in an MRI scanner. The stimuli consisted of 460 sentences corresponding to those used in the MOCHA-TIMIT corpus [12]. A more detailed description of this rt-MRI corpus is provided in a companion submission [13].

3. Problem formulation

Recall that the primary aim of this research is to extract dynamic articulatory primitives, weighted combinations of which can be used to resynthesize the various dynamic articulatory movements in the vocal tract. Techniques from linear algebra such as non-negative matrix factorization (NMF) which factor a given non-negative matrix into a linear combination of (non-negative) basis vectors are thus an excellent starting point to solve our problem.

3.1. Nonnegative Matrix Factorization and its extensions

The aim of NMF (as presented in [10]) is to approximate a non-negative input data matrix $\mathbf{V} \in \mathbb{R}^{\geq 0, M\times N}$ as the product of two non-negative matrices, a basis matrix $\mathbf{W} \in \mathbb{R}^{\geq 0, M\times K}$ and an activation matrix $\mathbf{H} \in \mathbb{R}^{\geq 0, K\times N}$ (where $K \leq M$) by minimizing the reconstruction error as measured by either a Euclidean distance metric or a Kullback-Liebler (KL) divergence metric. Although NMF provides a useful tool for analyzing data, it suffers from 2 problems of particular relevance in our case. First, it fails to account for potential dependencies across successive columns of $\mathbf{V}$ (in other words, capture the (temporal) dynamics of the data); thus a regularly repeating dynamic pattern would be represented by NMF using multiple bases, instead of a single basis function that spans the pattern length. Second, there is no guarantee that a given column will be represented by as few bases as possible, which is important to identify primitives. While one approach to solving this second problem is to impose sparsity conditions on the activation matrix, the first problem motivated the development of convolutive NMF [7], where instead we model $\mathbf{V}$ as:

$$\mathbf{V} \approx \sum_{t=0}^{T-1} \mathbf{W}(t) \cdot \overset{t\rightarrow}{\mathbf{H}} = \mathcal{V} \tag{1}$$

where each column of $\mathbf{W}(t) \in \mathbb{R}^{\geq 0, M\times K}$ is a time-varying basis vector sequence, each row of $\mathbf{H} \in \mathbb{R}^{\geq 0, K\times N}$ is its corresponding activation vector, T is the temporal length of each basis (number of image frames) and the $\overset{i\rightarrow}{(\cdot)}$ operator is a shift operator that moves the columns of its argument by i spots to the right, as detailed in [7]. In this case the author uses a KL divergence-based error criterion and derives iterative update rules for $\mathbf{W}(t)$ and $\mathbf{H}$ based on this criterion. This formulation was extended by O'Grady and Pearlmutter [8] to impose sparsity conditions on the activation matrix. However the parameter which trades-off sparsity of the activation matrix against the error criterion in their case (λ) is not readily interpretable, i.e., it is not clear what value λ should be set to to yield optimal interpretable bases. We instead choose to use a sparseness metric based on a relationship between the l_1 and l_2 norms (as proposed by [11]) as follows:

$$sparseness(\mathbf{x}) = \frac{\sqrt{n} - \frac{(\sum_i |x_i|)}{\sqrt{\sum_i x_i^2}}}{\sqrt{n}-1} \tag{2}$$

where n is the dimensionality of $\mathbf{x}$. This function equals unity iff $\mathbf{x}$ contains only a single non-zero component and 0 iff all components are equal upto signs and smoothly interpolates between the extremes. More recently Wang *et al.* [14] showed that using a Euclidean distance-based error metric was more advantageous (in terms of computational load and accuracy on an audio object separation task) than the KL divergence-based metric and further derived the corresponding multiplicative update rules for the former case. It is this formulation along with the sparseness constraints on $\mathbf{H}$ (as defined by Equation 2) that we use to solve our problem. Note that incorporation of the sparseness constraint also means that we can no longer use multiplicative update rules for $\mathbf{H}$ – so we use gradient descent followed by a projection step to update $\mathbf{H}$ iteratively (as proposed by [11]). The added advantage of using this technique is that it has been shown to find a unique solution of the NMF problem with sparseness constraints [15].

$$\min_{\mathbf{W},\mathbf{H}} \|\mathbf{V} - \sum_{t=0}^{T-1} \mathbf{W}(t) \cdot \overset{t\rightarrow}{\mathbf{H}} \|^2 \text{ s.t. } sparseness(h_i) = S_h, \forall i. \tag{3}$$

where h_i is the i^{th} row of $\mathbf{H}$ and $0 \leq S_h \leq 1$ is user-defined.

3.2. Extraction of primitive representations from rt-MRI data

If $I_1, I_2, \ldots, I_N$ are the N images (of dimension $n_1 \times n_2$) in an rt-MRI sequence re-formed into $M \times 1$ column vectors (where $M = n_1 \times n_2$), then we can design our data matrix $\mathbf{V}$ to be:

$$\mathbf{V} = [I_1 \,|I_2\,| \ldots \,|I_N\,] \in \mathbb{R}^{M\times N} \tag{4}$$

In our case, each image is of dimension 68 pixels by 68 pixels, i.e., $M = 68 \cdot 68 = 4624$. We now aim to find an approximation of this matrix V using a basis tensor $\mathbf{W}$ and an activation matrix $\mathbf{H}$. A complication which arises here is that for a given speaker in our dataset, there are 92 files (or image sequences), each of which results in a $4624 \times N$ data matrix V (where N is equal to the number of frames in that particular sequence). However we would like to obtain a *single* basis tensor $\mathbf{W}$ for

[2] We use NMF-based techniques since these have been shown to yield basis elements that can be assigned meaningful interpretation depending on the problem domain [6, 7, 8]. It is also worth noting that [9] gives specific conditions required for NMF algorithms to give a "correct" decomposition into parts, which affords us some mathematical insight into the decomposition.

all files so that we obtain a primitive articulatory representation for any sequence of articulatory movements made by that speaker. One possible way to do this is to concatenate all 92 image sequences into one huge matrix, but the dimensionality of this matrix makes computations intractably slow. In order to avert this problem we propose a second method that optimizes **W** jointly for all files and **H** individually per file. The algorithm is as follows:

```
1. Initialize $\mathbf{W}$ to a random tensor of appropriate dimension.
2. W Optimization.
   for Q of N files in the database do
     (a) Initialize $\mathbf{H}$ to a random matrix of requisite dimensions.
     (b) PROJECT. Project each row of $\mathbf{H}$ to be non-negative, have unit $l_2$ norm and $l_1$ norm set to achieve the desired sparseness [11].
     (c) ITERATE.
         i. H Update.
            for t = 1 to T do
              · Set $\hat{\mathbf{H}}(t) = \mathbf{H} - \mu_{\mathbf{H}} \mathbf{W}(t)(\overleftarrow{\mathcal{V}}^{t} - \overleftarrow{\mathbf{V}}^{t})$.
              · PROJECT $\mathbf{H}$.

            $\mathbf{H} \leftarrow \frac{1}{T} \sum \hat{\mathbf{H}}(t)$.
         ii. W Update.
            for t = 1 to T do
              · Set $\mathbf{W}(t) = \mathbf{W}(t) \otimes \mathbf{V}(\overrightarrow{\mathbf{H}}^{t})^{T} \oslash \mathcal{V}(\overrightarrow{\mathbf{H}}^{t})^{T}$.

3. for the rest of the files in the database do
     · H Update keeping $\mathbf{W}$ constant.
```

Step 2 is repeated for an empirically-specified number of iterations. The stepsize parameter $\mu_{\mathbf{H}}$ of the gradient descent procedure described in Step 2 is also set manually based on empirical observations. Note that the total number of rt-MRI image sequences in the database was $N = 92$ and **W** was optimized over $Q = 10$ files.

3.3. Selection of optimization parameters

In this section we briefly describe how we set the values of the various free parameters of the algorithm. The temporal extent of each basis sequence (T) was set to either 4 or 5 (since this corresponds to a reconstructed image sequence time period of approximately 170ms and 216ms respectively) to capture effects of the order of a syllable length on average. Since we want the activations of these basis vectors to be as sparse as possible (and as few basis vectors active at any given point of time) we choose the sparseness parameter (S_h) to be in the range $0.7 - 0.9$. This parameter as well as the optimal number of bases (K) was chosen by looking at the performance of the algorithm for different values of S_h and K (an example graph is plotted in Figure 1). Note that the figure shows the performance of the algorithm for $T = 1$. Since increasing the value of T just causes an increase in the number of NMF operations by a factor of T, we can use this to get a general idea of how the algorithm performs[3] with different values of S_h and K. One general trend which is seen is that the squared error (or value of the objective function) after 50 iterations decreases as K increases – this makes intuitive sense since we expect to get a better approximation of V as K approaches the rank of V. In addition, the objective function is lower for lower values of the sparseness parameter S_h. Based on such observations and the fact that we would like the dimension of the extracted basis to be as small (for better interpretability), we choose $S_h = 0.75$ and $K = 15$.

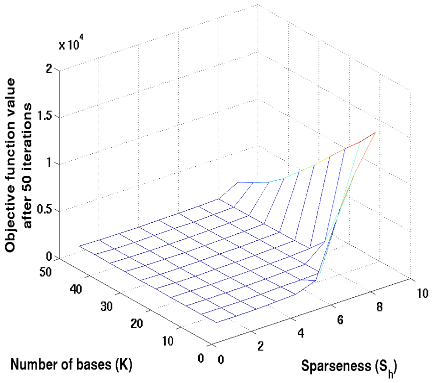

Figure 1: Performance of the algorithm as measured by the objective function value (as defined by equation 3) on a dataset file for $T = 1$ different values of sparseness S_h and number of bases K.

4. Linguistic interpretation and validation

Figure 2 depicts 4 out of 15 basis sequences extracted using the cNMFsc algorithm and how they can be combined (after weighing by corresponding activation functions) to approximate a given sequence (example shows the word "cut"). Notice that basis sequence A, B and C correspond to the formation of a dorsal constriction, release of a dorsal constriction and a coronal constriction respectively. Also notice that each of these sequences are activated one after the other in sequence as is required to produce the sequence "cut". These basis sequences A-C are interpretable and yield a good approximation of the input sequence. In addition the sequence D depicts a vocal tract posture or 'setting' at rest – basis sequences like these are important for linguistic studies of articulatory setting or basis of articulation to understand how articulatory postures are controlled by the speech planning mechanism [16]. This sequence of "average" vocal tract postures is typically combined with the other basis sequences (A-C) in a weighted manner to approximate a given sequence of interest. The illustration in Figure 2 also shows that this method doesn't always give clear-cut results, for instance, there is no image in any of the basis sequences that clearly depicts a complete dorsal closure; the basis sequences extracted by the algorithm depend on choice of parameters and the content of the input sequences that W is optimized on. In addition, we have not yet explored as to what extent the activation functions corresponding to the extracted bases are able to differentiate stops versus fricatives, where there is a fine distinction in the constriction degree.

Validating the performance of such algorithms is in general a difficult task. One way is estimating how well the features are able to distinguish between broad linguistic classes such as manners and places of articulation. One method to compare our algorithm against and which may be expected to do well on this task is principal component analysis (PCA), which extracts K basis vectors that account for as much variance in the observed data as possible, but is not readily interpretable. So one

[3] Given the large dimensionality of the videos in our problem, the algorithm takes a long time to run for a given set of parameters; hence we used a temporal dimension of $T = 1$ to optimize S_h and K.

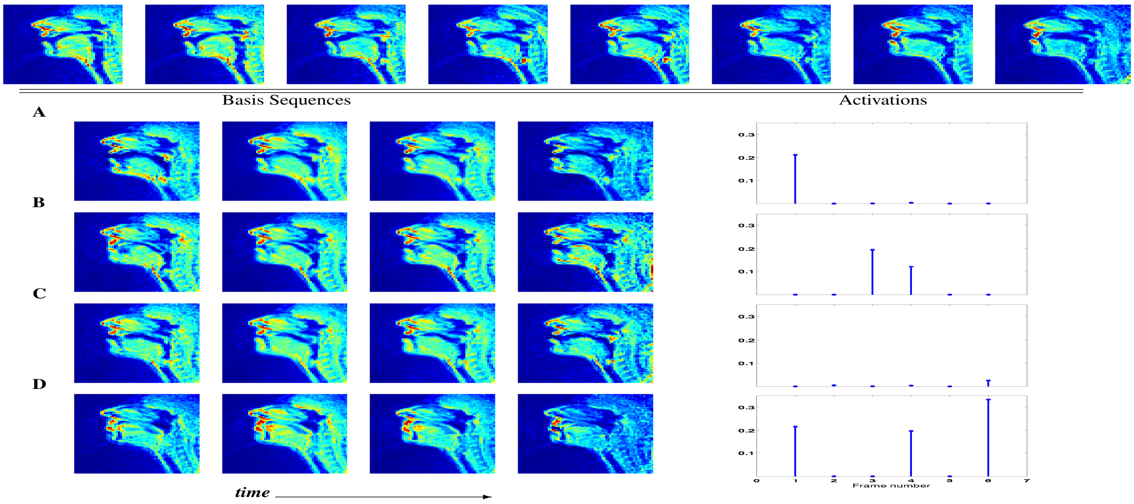

Figure 2: Example decomposition of the word "cut" spoken by a female subject (top row shows 8 frames corresponding to "cut"). Rows A-D show activations (rightmost panel) of 4 component bases (out of a total of $K = 15$) at times corresponding to the first 6 frames (out of the 8) shown in the top panel ($S_h = 0.75$ and $T = 4$). Activations of other bases were zero for the time frame displayed. See text for details.

test for our algorithm is whether it is able to bridge this gap – discriminate between broad linguistic classes atleast as well as PCA while giving us interpretable bases simultaneously. *Preliminary* HMM-based articulatory recognition results indicate that the (activation) features extracted using the cNMFsc algorithm ($S_h = 0.75, K = 15, T = 4$) correctly distinguished between broad linguistic classes 30.33% of the time for one speaker while activation (features) extracted using PCA performed slightly better at 40.34%. Thus there is scope for improvement. It must be however noted that the recognition setup has not yet been optimized to deal with sparse feature vectors (such as the activation features in our case); in addition, we have not preprocessed/denoised the data, and some of the extra variation captured by PCA might be due to noise in the image sequences.

5. Conclusions and future work

We have presented an algorithm to extract basis image sequences of articulatory movements from real-time MRI data. As one can see in Figure 2, the extracted basis is somewhat interpretable to the trained linguist; for example, one can see the formation of a tongue-tip and tongue dorsum closures captured by 2 of the basis functions. Note that some of the vocal tract shapes *not* represented well include extreme shapes, such as that assumed during an /a:/ vowel. In addition, noting that different articulatory actions might extend over different temporal durations, designing the T parameter to be variable for different bases might yield better results.

In future work, we hope to develop a more concrete validation and articulatory recognition framework for our algorithm. We would like to develop and extend the proposed algorithm/method to find a link between knowledge-driven representations of articulatory movement and data-driven representations (such as the proposed method) to obtain interpretable bases of articulatory actions that are provably commensurate with linguistic theories. In addition, we would like to explore other approaches, probabilistic and otherwise, from the sparse coding literature to improve the performance of the algorithm.

References

[1] B. Olshausen and D. Field, "Sparse coding of sensory inputs," *Current opinion in neurobiology*, vol. 14, no. 4, pp. 481–487, 2004.

[2] T. Hromádka, M. DeWeese, and A. Zador, "Sparse representation of sounds in the unanesthetized auditory cortex," *PLoS Biol*, vol. 6, no. 1, p. e16, 2008.

[3] C. Browman and L. Goldstein, "Dynamics and articulatory phonology," *Mind as motion: Explorations in the dynamics of cognition*, 1995.

[4] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *The Journal of the Acoustical Society of America*, vol. 115, p. 1771, 2004.

[5] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *The Journal of Machine Learning Research*, vol. 11, pp. 19–60, 2010.

[6] B. Mel, "Computational neuroscience: Think positive to find parts," *Nature*, vol. 401, no. 6755, pp. 759–760, 1999.

[7] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 1–12, 2007.

[8] P. O'Grady and B. Pearlmutter, "Discovering speech phones using convolutive non-negative matrix factorisation with a sparseness constraint," *Neurocomputing*, vol. 72, no. 1-3, pp. 88–101, 2008.

[9] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts," *Advances in neural information processing systems*, vol. 16, 2004.

[10] D. Lee and H. Seung, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, 2001.

[11] P. Hoyer, "Non-negative matrix factorization with sparseness constraints," *The Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[12] A. Wrench, "A multi-channel/multi-speaker articulatory database for continuous speech recognition research," in *Workshop on Phonetics and Phonology in ASR*, 2000.

[13] S. Narayanan, E. Bresch, P. Ghosh, L. Goldstein, A. Katsamanis, Y. Kim, A. Lammert, M. Proctor, V. Ramanarayanan, and Y. Zhu, "A multimodal real-time mri articulatory corpus for speech research," in *Proc. 12th Conf. Intl. Speech Communication Assoc. (Interspeech 2011)*, Florence, Italy, Submitted.

[14] W. Wang, A. Cichocki, and J. Chambers, "A multiplicative algorithm for convolutive non-negative matrix factorization based on squared Euclidean distance," *Signal Processing, IEEE Transactions on*, vol. 57, no. 7, pp. 2858–2864, 2009.

[15] F. Theis, K. Stadlthanner, and T. Tanaka, "First results on uniqueness of sparse non-negative matrix factorization," in *Proceedings of the 13th European Signal Processing Conference (EUSIPCO05)*. Citeseer, 2005.

[16] V. Ramanarayanan, D. Byrd, L. Goldstein, and S. Narayanan, "Investigating Articulatory Setting-Pauses, Ready Position, and Rest-Using Real-Time MRI," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.