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 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 MOCHA-TIMIT 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. We finally attempt an interpretation of the extracted basis set of image sequences in an Articulatory Phonology gesture-based framework [1].

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

1. Introduction

Extracting interpretable representations from raw articulatory data is critical for better understanding, modeling and artificial 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 [2]. 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 [2], and quantitive evidence has been put forth for sparse representations of sounds in the auditory cortex [3]. However, not many computational studies have been conducted into uncovering the primitives of speech production.

There are two broad approaches to attack this problem of formulating representations of speech production - knowledge-driven and data-driven. There have been many attempts at knowledge-driven formulations in the linguistics literature. An example is the framework of Articulatory Phonology [1] 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. 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.

Modeling data vectors as sparse linear combinations of basis elements is a general 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. These methods have been successfully applied to other problems in signal processing, machine learning, and neuroscience. 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* basis or dictionary elements, we focus on matrix factorization techniques such as Nonnegative Matrix Factorization (NMF)[1] and its variants [10, 11, 7, 8] with sparsity constraints imposed on the *activation* 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. We further validate the learned representations using a recognition task (as well as the regular

[1]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.

approximation error metric).

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 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].

2.1. Recording setup

The data was obtained using a GE Signa Excite HD 1.5 T scanner with designed spiral gradient waveforms capable of 22mT/m amplitudes and 77mT/m/sec slew rates. The pulse sequence used was a low flip angle 13-interleaf spiral gradient echo sequence with the following parameters: a repetition time of TR = 6.164ms, a 20 x 20 cm^2 field of view (FOV) resulting in an image of 68 pixels by 68 pixels, a 3 x 3 mm^2 in-plane spatial resolution, and an 80.1 ms temporal resolution corresponding to 12.5 frames per second. The slice thickness used was 5mm. A 4-channel upper airway receive coil array was used for RF signal reception. In the 4-channel receive coil array, two coil elements are anterior and the other two coil elements are posterior to the head and neck. Synchronized audio was recorded simultaneously using a fiber-optic microphone. In order to guarantee sample-exact synchronicity the audio sample clock is derived from the MRI scanner's 10MHz master clock and the recording is triggered on and off using the RF master-exciter unblank signal from the MRI scanner. For further details, please see [13].

2.2. Data postprocessing

The MRI scanner emits high intensity gradient noise in the audible range during scans, which makes acoustic analysis of any audio record very difficult. We use a model-based noise cancellation technique [13] which takes into account the periodicity of the gradient noise to solve this problem.

The audio data is phonetically aligned using the SailAlign tool [14]. In order to allow the aligner tool to adapt better phone models to our data, instead of aligning each audio file individually all audio files are concatenated into a master audio file which is passed as input to the aligner tool (the parameters of the aligner are optimized by trial and error to obtain a good working configuration). The final database consists of audio, MRI video and phone- and word-level transcriptions of 460 sentences (corresponding to those used in the MOCHA-TIMIT corpus [12]), split into 92 files contaning 5 sentences each.

3. Problem formulation

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 fails to account for potential dependencies across successive columns of $\mathbf{V}$; 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. This motivated the development of convolutive NMF [7], where instead we model $\mathbf{V}$ as:

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

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 and the $\vec{(\cdot)}^{i}$ 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.* [15] 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. However, 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 [16].

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

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, 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 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 $\mathbf{W}$ jointly for all files and $\mathbf{H}$ 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 till convergence is reached. The stepsize parameter $\mu_{\mathbf{H}}$ of the gradient descent procedure described in Step 2 is also set manually based on empirical observations.

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 approx. 170ms and 216ms respectively. 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[2] 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 $S_h = 0.85$ and $K = 15$.

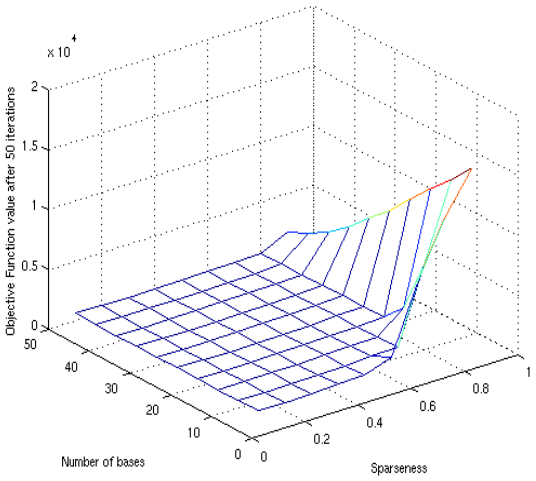

Figure 1: Performance of the algorithm as measured by the objective function value (as defined by equation 1) on a dataset file for $T = 1$ different values of sparseness S_h and number of bases K.

4. Validation and interpretation

5. Discussion 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. Other tongue shapes, such as the bunched shape formed during the production of an /i/ vowel, are also seen. We further notice a redundancy in some of the sequences extracted, such as those of the neutral vocal tract shape, but this is to be expected since this posture is adopted most frequently adopted during running speech. Note that some of the vocal tract shapes *not* represented well include extreme shapes, such as that assumed during an /a:/ vowel.

In future work, we would like to develop and extend the proposed method to find a link between knowledge-driven representations of articulatory movement and data-driven representations (such as the proposed method) to obtain truly interpretable bases of articulatory actions. In addition, we would like to explore other approaches, probabilistic and otherwise, from the sparse coding literature to improve the performance of the algorithm.

[2]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.

6. References

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

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

[3] 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.

[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] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans," *The Journal of the Acoustical Society of America*, vol. 120, p. 1791, 2006.

[14] A. Katsamanis, M. Black, P. Georgiou, L. Goldstein, and S. Narayanan, "SailAlign: Robust long speech-text alignment," *Proc. of Workshop on New Tools and Methods for Very Large Scale Research in Phonetic Sciences*, 2011.

[15] 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.

[16] 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.

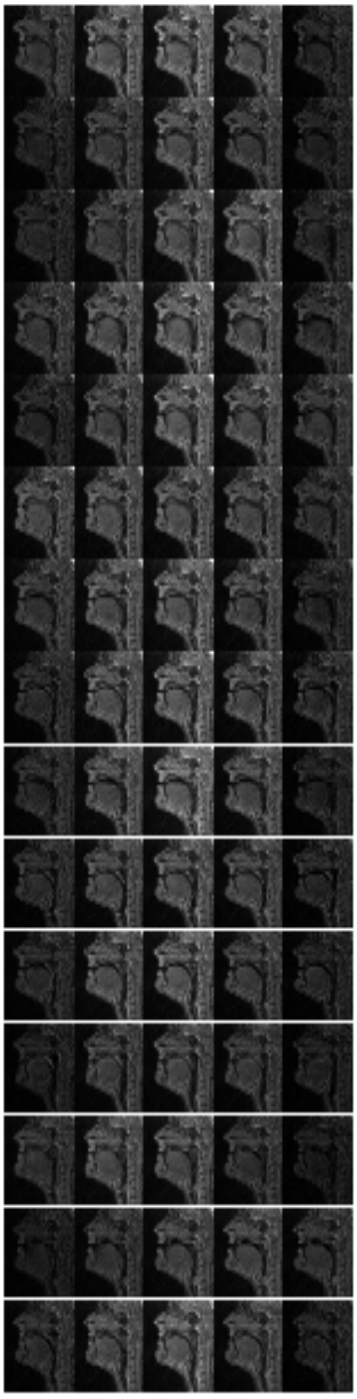

Figure 2: A set of 15 basis sequences of temporal extent 5 frames.