INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Convex Hull Convolutive Non-negative Matrix Factorization for Uncovering Temporal Patterns in Multivariate Time-Series Data

Colin Vaz, Asterios Toutios, and Shrikanth Narayanan

Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA 90089
cvaz@usc.edu, <toutios,shri>@sipi.usc.edu

Abstract

We propose the Convex Hull Convolutive Non-negative Matrix Factorization (CH-CNMF) algorithm to learn temporal patterns in multivariate time-series data. The algorithm factors a data matrix into a basis tensor that contains temporal patterns and an activation matrix that indicates the time instants when the temporal patterns occurred in the data. Importantly, the temporal patterns correspond closely to the observed data and represent a wide range of dynamics. Experiments with synthetic data show that the temporal patterns found by CH-CNMF match the data better and provide more meaningful information than the temporal patterns found by Convolutive Non-negative Matrix Factorization with sparsity constraints (CNMF-SC). Additionally, CH-CNMF applied on vocal tract constriction data yields a wider range of articulatory gestures compared to CNMF-SC. Moreover, we find that the gestures comprising the CH-CNMF basis generalize better to unseen data and capture vocal tract structure and dynamics significantly better than those comprising the CNMF-SC basis.

Index Terms: dictionary learning, articulatory gestures, speech production

1. Introduction

Observation of latent structure in data provides researchers with a tool for data analysis and interpretation. Dictionary learning methods are commonly used to uncover the latent structure. Non-negative matrix factorization (NMF) is a dictionary learning algorithm used in a wide range of fields, from speech enhancement [1, 2] and analysis [3] to computational biology [4] and molecular analysis [5, 6]. First proposed by Paatero and Tapper [7, 8] and developed further by Lee and Seung [9], NMF decomposes a data matrix into a basis matrix that contains basic units of the data and an activation matrix that encodes the data in terms of the basis matrix. Convolutive NMF (CNMF) [10] was proposed to consider temporal context in time-series data and extract temporal patterns observed in the data. CNMF was shown to find speech phones when operating on spectrograms of speech. Sparsity constraints on either the basis or activation matrix can be employed in order to get more interpretable outputs, depending on the application [11, 12]. In order to provide interpretability of the sparsity parameter, Hoyer proposed NMF with sparsity constraints (NMF-SC) [11], where the sparsity parameter ranges between 0 and 1, with 0 requiring no sparsity and 1 enforcing maximum sparsity. A convolutive extension to this algorithm, CNMF-SC, was recently proposed in [3] to find articulatory primitives in Electromagnetic Articulography (EMA) data.

Work supported by NIH grant R01DC007124 and NSF grant 1514544.

One drawback of NMF and the variants mentioned above is the requirement of a non-negative data matrix. This can prevent the use of NMF in cases where the data contain negative values. To overcome this, Ding et al. proposed the Convex NMF algorithm [13], where the basis matrix is formed as a convex combination of the data points. They showed that Convex NMF tends to find sparse solutions and the basis vectors correspond closely to observed data points, making the basis more interpretable over an NMF basis. Thurau et al. introduced Convex Hull NMF (CH-NMF) to improve computation speed on large datasets [14]. They proposed to form the basis matrix from convex combinations of the convex hull vertices rather than the data itself. They showed that the basis vectors from this approach tend to lie at the extremities of the data. Thus, the CH-NMF basis contains a wide range of basic units present in the data.

We propose the Convex Hull Convolutive NMF (CH-CNMF) algorithm that extends CH-NMF to incorporate temporal context in time-series data. Like CNMF, the basis will contain a set of temporal patterns found in the data. However, the basis will inherit the desirable properties of the CH-NMF basis: temporal patterns that correspond closely to temporal units in the data and represent a wide range of dynamics.

This paper is organized as follows. Section 2 describes the CH-CNMF algorithm. Section 3 discusses the experiments on synthetic time-series and real articulatory data and compares the performance quantitatively and qualitatively to CNMF-SC. Finally, Section 4 offers our conclusions and directions for future work.

2. CH-CNMF Algorithm

Assume a multivariate time-series $V = [\mathbf{v}_1\, \mathbf{v}_2 \cdots \mathbf{v}_\mathbf{n}] \in \mathbb{R}^{m \times n}$ of m variables over n time frames. CH-CNMF tries to find K temporal patterns of duration T in V. To achieve this, we propose minimizing the cost function

$$(\hat{G}, \hat{H}) = \arg\min_{G,H} \left\| V - S \sum_{t=0}^{T-1} G(t) \overset{t\rightarrow}{H} \right\|_F^2 + \lambda \|H\|_1$$
$$\text{subject to } \|\mathbf{g}_k(t)\|_1 = 1,\ \forall k \in \{1, \ldots, K\},$$
$$\forall t \in \{0, \ldots, T-1\}. \quad (1)$$

$S \in \mathbb{R}^{m \times p}$ are p vertices of the convex hull of V. $G \in \mathbb{R}_+^{p \times K \times T}$ forms convex combinations of the columns of S to represent the time series in V. Because we want convex combinations of the columns of S, we require G to have non-negative entries and for each column to sum to 1. In Equation 1, $G(t) \in \mathbb{R}_+^{p \times K}$ indexes the third dimension of G. $W(t) = SG(t) \in \mathbb{R}^{m \times K}$ is a basis for V. By combining $W(t)$ over all t into a three-dimensional tensor $W \in \mathbb{R}^{m \times K \times T}$, we

http://dx.doi.org/10.21437/Interspeech.2016-571

have a time-varying basis for V that captures K temporal patterns of duration T. $H \in \mathbb{R}_+^{K \times n}$ represents the activations of V in terms of the basis W. The notation $\overset{t\rightarrow}{H}$ means that the columns of H are shifted t places to the right and t all-zero columns are filled in on the left. The λ parameter trades off reconstruction error for sparsity in the activation matrix. Sparsity in the activations forces each data point in V to be represented by a few basis vectors. This usually leads to more interpretable basis vectors.

To find the vertices of the convex hull, we follow the approach described in [14]. First, compute the D eigenvectors $\mathbf{e}_d$ corresponding to the D largest eigenvalues of the covariance matrix of V. D is chosen such that the eigenvectors account for 95% of the data variance. Then, project V onto 2D subspaces:

$$\tilde{V}_{p,q} = [\mathbf{e}_p \; \mathbf{e}_q]^T V \in \mathbb{R}^{2 \times n}, \; \forall p, q \in \{1, \ldots, D\}, p \neq q. \quad (2)$$

Next, find the vertices of the convex hull of $\tilde{V}_{p,q}$ using a convex hull-finding method (e.g. [16, 17]) and store the frame indices of the vertices in $\text{ch}(\tilde{V}_{p,q})$. Finally, form S by concatenating all the points in V marked as a convex hull vertex:

$$S = [V_{\text{ch}(\tilde{V}_{1,2})} \, V_{\text{ch}(\tilde{V}_{1,3})} \cdots V_{\text{ch}(\tilde{V}_{D-1,D})}], \quad (3)$$

where $V_{\text{ch}(\tilde{V}_{p,q})}$ are the columns of V corresponding to the indices in $\text{ch}(\tilde{V}_{p,q})$. There may be duplicate columns in S, so the repeated columns should be removed.

To find G and H that minimizes the cost function (Equation 1), we iteratively alternate updating G and H until the cost function converges or a given number of iterations have occurred. Let $F = \sum_{t=0}^{T-1} G(t) \overset{t\rightarrow}{H}$ and I_n be the $n \times n$ identity matrix. The update for G is

$$G(t) \leftarrow G(t) \otimes \frac{\left([S^T V]^+ + [S^T S]^- F\right) \overset{t\rightarrow}{H^T}}{\left([S^T V]^- + [S^T S]^+ F\right) \overset{t\rightarrow}{H^T}}, \; \forall t \in \{0, \ldots, T-1\}, \quad (4)$$

where $[A]^+ = 0.5(|A| + A)$ and $[A]^- = 0.5(|A| - A)$ represent the positive and negative elements of matrix A respectively. The update for H is

$$H \leftarrow H \otimes \frac{\sum_{t=0}^{T-1} G^T(t) \left([S^T V]^+ \overset{\leftarrow t}{I_n} + [S^T S]^- \overset{\leftarrow t}{F} \right)}{\sum_{t=0}^{T-1} G^T(t) \left([S^T V]^- \overset{\leftarrow t}{I_n} + [S^T S]^+ \overset{\leftarrow t}{F} \right) + \lambda}. \quad (5)$$

The operator $\otimes$ means element-wise multiplication, and the division is element-wise.

3. Experiments

We evaluated the CH-CNMF algorithm on two datasets. The first dataset was created synthetically to aid evaluation of the basis vectors and verify that the algorithm finds a meaningful basis. For the second dataset, we used vocal tract constrictions derived from real-time MRI images of a subject speaking TIMIT sentences. We chose this dataset to assess the performance of CH-CNMF on realistic time-series data and to uncover articulatory gestures in a data-driven manner. For both datasets, we compared the performance of CH-CNMF to CNMF-SC.

3.1. Synthetic data

To create synthetic time-series data, we created three Markov chains, each with four states. Each state generates a sample from a two-dimensional Gaussian distribution with a given mean vector and covariance matrix. The means were chosen such that each Markov chain produces distinct 4-sample sequences. Within each chain, the states transitioned from left to right with probability 1 to ensure that the chain outputs exactly four samples. After transitioning out of the last state, another chain is chosen, with each chain having equal probability of being selected. We used this procedure to generate a 1000-sample time series. Figure 1a shows an example output plotted in two-dimensional space, with circles indicating the states of the Markov chains and arrows indicating transitions between the states within each chain. Note that the output of one chain is separated spatially from the other two, while the other two chains share the same second state.

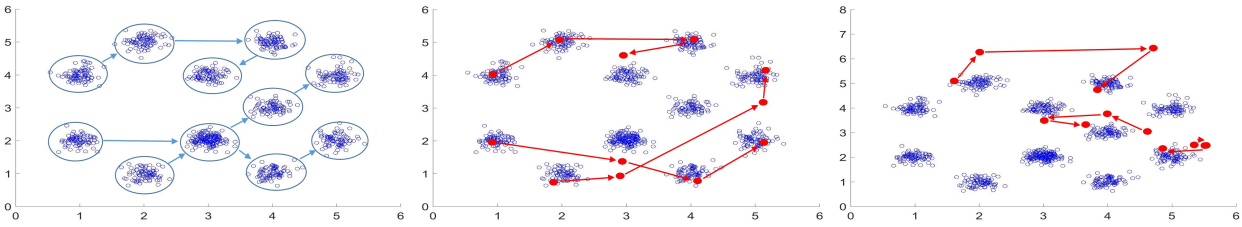
Since we know that the data has three distinct patterns with a length of four samples, we set $K = 3$ and $T = 4$. Additionally, we experimentally determined $\lambda = 1$ to be a good choice. In this data, we know that only one of the three chains are active at a particular time. This suggests that the activation matrix H should be about 67% sparse. Thus, we set the sparsity parameter in CNMF-SC to 0.67. We used 100 update iterations for each algorithm, and we ran the experiment 100 times to account for effects of random initialization.

Figures 1b and 1c show the basis for CH-CNMF and CNMF-SC respectively. From these figures, one can see that CH-CNMF recovers the three temporal patterns more clearly than CNMF-SC. Specifically, we see that CH-CNMF accurately chooses clusters that reside near the convex hull of the data. Meanwhile, inner clusters are represented less accurately; the algorithm tends to shift the basis vectors for the inner clusters closer to the convex hull. On the other hand, the CNMF-SC basis is less interpretable because the basis patterns don't correspond closely to the observed data points. CNMF-SC scales the rows of the activation matrix H to have unit ℓ_2 norm, so the basis patterns are scaled accordingly to minimize the CNMF-SC cost function. These results agree with those found by Thurau et al. [14], where the CH-NMF basis lies near the convex hull of the data while the NMF basis does not correspond well to the data points. While it is possible to scale the CNMF-SC basis vectors to lie closer to the data points, this procedure may not be feasible on larger and real-world datasets. Thus, we see that CH-CNMF captures temporal structure in the data more reliably than CNMF-SC.

3.2. Vocal tract data

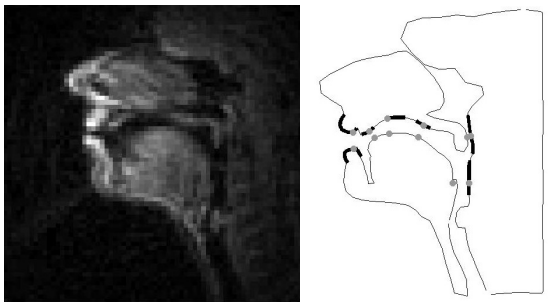
To supplement the synthetic data experiment, we tested our algorithm on measurements of constriction degrees (Euclidean distances) between active and passive articulators during a speech task. Articulatory Phonology [18] theorizes that the movements of the vocal tract can be decomposed into units of vocal tract actions called gestures. Gestures implement constrictions of variable degrees along the vocal tract, and the timings of the gestures are arranged into a gestural score. The goal in this experiment is to derive such gestures along with a gestural score in a data-driven manner.

We used mid-sagittal real-time MRI data of the vocal tract from the F1 speaker of the USC-TIMIT database [19]. The frame rate of the MRI data in this corpus is 23.18 frames per second. We used an automatic segmentation algorithm [20] to find the contours of the air-tissue boundaries of the vocal tract in each frame. Figure 2 shows an example MRI frame and the contours found from this frame. Based on the contours, the constriction degrees were measured at five places of articulation (bilabial, alveolar, palatal, velar, and pharyngeal), plus the



(a) Input data (b) CH-CNMF basis patterns (c) CNMF-SC basis patterns

Figure 1: Input synthetic data and the recovered temporal patterns from CH-CNMF and CNMF-SC. The circles in (a) indicate the states of the Markov chains. The arrows represent the temporal progression within the Markov chains and the recovered basis patterns.



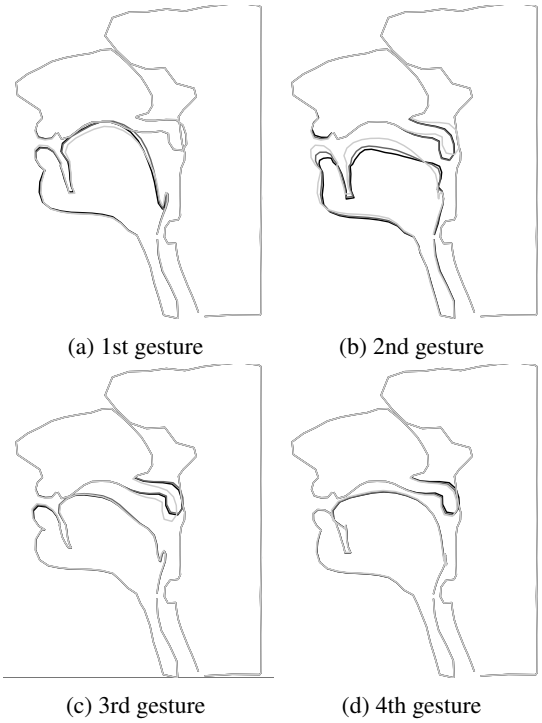
(a) MRI frame (b) Contours

Figure 2: Air-tissue boundary contours (thin black lines) from a single real-time MRI frame. The thick black lines mark the places of articulation where constrictions are measured (from lips to glottis: bilabial; alveolar; palatal; velar; velopharyngeal port; pharyngeal). For a given frame, constrictions are measured as the Euclidean distances between opposing gray dots.

velopharyngeal port opening. The locations of the constriction measurements are indicated in Figure 2b with thick black lines. There are 460 TIMIT sentences, but a technical error during data collection caused the head orientation of the first 250 sentences to be different from the rest. To avoid any errors due to incorrect vocal tract segmentation, we performed our analysis using only the first 250 sentences. We assigned the first 150 sentences as the training set, which we used to learn a basis of gestures. The remaining 100 sentences were assigned to the testing set to evaluate how well the gesture basis generalizes to unseen data.

The parameters K and T are chosen based on the data and application. We chose $T = 3$ to capture gestures with a duration of 130 ms ($3 \times \frac{1 \text{ second}}{23.18 \text{ frames}} \approx 130$ ms), which is roughly the average duration of a phoneme. Since we measured constrictions at 6 locations, we chose $K = 6$. Choice of K is highly data dependent and can be chosen in a more principled manner for a specific application. We used the same K and T values for CNMF-SC, and we set CNMF-SC's sparsity parameter to 0.7, as suggested in [3]. We ran both algorithms with 200 update iterations.

After running the algorithms, the bases contain constriction degrees at the six locations in the vocal tract. In order to visualize the bases, we used a forward map [21] to convert the constriction degrees to articulatory weights [22] that describe the relative contributions of ten vocal tract-shaping components towards the formation of a given vocal tract shape. Figure 3 shows vocal tract movements (gestures) due to the constrictions found in the CH-CNMF basis. Figure 4 shows the same for the CNMF-SC basis. In the interest of space, we only show four



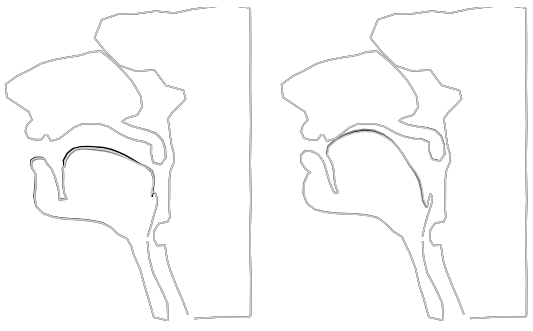
(a) 1st gesture (b) 2nd gesture

(c) 3rd gesture (d) 4th gesture

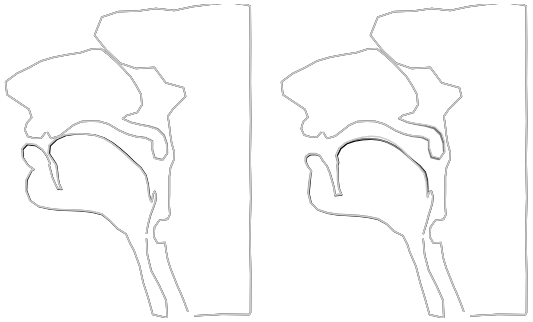
Figure 3: Visualization of the CH-CNMF gesture basis. The vocal tract at time step 1 is shown in light grey, time step 2 in dark grey, and time step 3 in black.

gestures from each algorithm. The CH-CNMF basis shows interpretable articulatory gestures; for example, Figure 3a shows the tongue body rising, while Figure 3c shows the tongue forming a dental/alveolar constriction while the velum simultaneously closes. In general, the gestures found by CH-CNMF are more overt and display a wider range of vocal tract movement than those found by CNMF-SC. This agrees with the results of the synthetic data experiment, where CH-CNMF tends to find temporal patterns at the extremities of the data, while the temporal patterns from CNMF-SC generally don't correspond to the data.

To evaluate how well the learned gestures generalize to unseen data, we fix the basis for each algorithm and find the activation matrix H_{test} for each sentence in the test set using Equation 5. We then reconstruct the constrictions and find the root mean square error (RMSE) and correlation between the input and reconstructed constrictions. Table 1 shows the average RMSE and correlations for both algorithms on the test set. We used a one-sided Wilcoxon rank-sum test to find that the RMSE was significantly lower ($p \approx 0$) and the correlation was sig-



(a) 1st gesture (b) 2nd gesture



(c) 3rd gesture (d) 4th gesture

Figure 4: Visualization of the CNMF-SC gesture basis. The vocal tract at time step 1 is shown in light grey, time step 2 in dark grey, and time step 3 in black.

nificantly higher ($p \approx 0$) for CH-CNMF than for CNMF-SC. This indicates that the gestures found by CH-CNMF generalize better than the CNMF-SC basis.

Additionally, we used an experimental procedure described in [3] to evaluate the extent to which the bases captured temporal structure in the data. If we suppose that the bases contain random temporal patterns, then we don't expect a significant change in the RMSE and correlation between the input and reconstructed constrictions when we substitute H_{test} with a random matrix H_{rand} with the same sparsity as H_{test}. To ensure H_{rand} has the same sparsity as H_{test}, we used the method proposed by Hoyer [11] to set the ℓ_1 and ℓ_2 norms of each row of H_{rand} to the ℓ_1 and ℓ_2 norms of the corresponding rows of H_{test}. The results of reconstructing with a random matrix are shown in Table 1. Using a one-sided Wilcoxon rank-sum test, we found that the RMSE was significantly lower when reconstructing with H_{test} than with H_{rand} for both CH-CNMF ($p \approx 0$) and CNMF-SC ($p \approx 0$). Also, the correlation was significantly higher when reconstructing with H_{test} than with H_{rand} for both CH-CNMF ($p \approx 0$) and CNMF-SC ($p \approx 0$). These results suggest that both algorithms learn meaningful temporal structure from the training set data.

Table 1: Root mean square errors (RMSE) and correlations when reconstructing vocal tract constriction using a calculated activation matrix H_{test} and a random activation matrix H_{rand}.

Algorithm	Activation matrix	RMSE (mm)	Correlation
CH-CNMF	H_{test}	0.824	0.964
	H_{rand}	3.419	-0.002
CNMF-SC	H_{test}	6.058	0.619
	H_{rand}	8.127	0.168

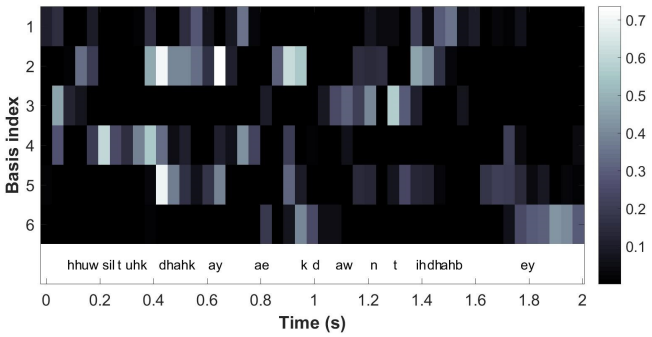


Figure 5: CH-CNMF activation matrix for the TIMIT sentence "Who took the kayak down to the bay?". The time-aligned Arpabet transcription is shown below the matrix.

For additional insight, we plot the CH-CNMF activation matrix of a TIMIT sentence from the testing set in Figure 5. The activation matrix indicates how much a gesture is activated at a particular time, with lighter colors indicating greater activation. A time-aligned Arpabet transcription of the sentence is shown below the activation matrix to help correlate acoustics with gesture activations. From inspection of Figure 5, it appears that the third gesture occurs during coronal stops while the fourth gesture is associated with high back vowels. These observations correspond closely to the gestures in Figure 3, where the third gesture shows a dental/alveolar constriction formed by the tongue tip and the fourth gesture shows the tongue positioned high in the mouth. Further observations about the remaining gestures can be made more clearly from activation matrices of other TIMIT sentences. Thus, the activation matrix can be interpreted as a "gestural score" because it indicates the occurrences of different gestures.

4. Conclusion

We introduced the Convex Hull Non-negative Matrix Factorization (CH-CNMF) algorithm to find temporal patterns in multivariate time-series data. It factors a data matrix into a basis tensor that contains temporal patterns and an activation matrix that indicates the times at which the temporal patterns occur in the data. Using synthetic data, we showed that CH-CNMF extracts better, more interpretable temporal patterns than Convolutive Non-negative Matrix Factorization with sparsity constraints (CNMF-SC). With vocal tract constriction data, we were able to find a wider range of articulatory gestures using CH-CNMF than using CNMF-SC. We also demonstrated that the gestures contained in the CH-CNMF basis generalized better to unseen data and extracted better vocal tract dynamics than the CNMF-SC basis by reconstructing the data with significantly lower RMSE and significantly higher correlation. Finally, we showed that the activation matrix can be interpreted as a "gestural score".

Building upon this work, we will explore training the basis to be discriminative of the labels in a labeled dataset (e.g. phonemes in an utterance). To make this algorithm computationally tractable on large datasets, we will explore different formulations and optimization techniques to speed up calculations and scale down memory requirements. We will apply this algorithm for other speech-related tasks, such as phoneme classification and automatic speech recognition, and to domains beyond speech where extracting temporal patterns from data are useful.

5. References

[1] T. Virtanen, "Sound source separation using sparse coding with temporal continuity objective," in *Int. Computer Music Conf.*, 2003, pp. 231–234.

[2] C. Vaz, V. Ramanarayanan, and S. Narayanan, "A two-step technique for MRI audio enhancement using dictionary learning and wavelet packet analysis," in *Interspeech*, Lyon, France, 2013, pp. 1312–1315.

[3] V. Ramanarayanan, L. Goldstein, and S. Narayanan, "Spatio-temporal articulatory movement primitives during speech production – extraction, interpretation and validation," *J. Acoustic Society of America*, vol. 134, no. 2, pp. 1378–1394, 2013.

[4] K. Devarajan, "Nonnegative matrix factorization: an analytical and interpretive tool in computational biology," *PLoS Computational Biology*, vol. 4, no. 7, 2008.

[5] J.-P. Brunet, P. Tamayo, T. R. Golub, J. P. Mesirov, and E. S. Lander, "Metagenes and molecular pattern discovery using matrix factorization," *Proc. Nat. Academy Sciences*, vol. 101, no. 12, pp. 4164–4169, 2004.

[6] Y. Gao and G. Church, "Improving molecular cancer class discovery through sparse non-negative matrix factorization," *Bioinformatics*, vol. 21, no. 21, pp. 3970–3975, 2005.

[7] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

[8] P. Paatero, "Least squares formulation of robust non-negative factor analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 37, no. 1, pp. 23–35, 1997.

[9] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Adv. in Neu. Info. Proc. Sys. 13*, 2001, pp. 556–562.

[10] P. Smaragdis, "Convolutive Speech Bases and Their Application to Supervised Speech Separation," *IEEE Trans. Acoustics, Speech, and Lang. Process.*, vol. 15, no. 1, pp. 1–12, Jan. 2007.

[11] P. O. Hoyer, "Non-negative Matrix Factorization with Sparseness Constraints," *J. Machine Learning Research*, vol. 5, pp. 1457–1469, Dec. 2004.

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

[13] C. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 45–55, 2010.

[14] C. Thurau, K. Kersting, M. Wahabzada, and C. Bauckhage, "Convex non-negative matrix factorization for massive datasets," *Knowledge and Information Systems*, vol. 29, no. 2, pp. 457–478, 2011.

[15] G. M. Ziegler, "Lectures on polytopes," *Springer Science & Business Media*, 1995.

[16] R. L. Graham, "An Efficient Algorithm for Determining the Convex Hull of a Finite Planar Set," *Information Process. Letters*, vol. 1, no. 4, pp. 132–133, 1972.

[17] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls," *ACM Trans. Mathematical Software*, vol. 22, no. 4, pp. 469–483, 1996.

[18] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, no. 3-4, pp. 155–180, 1992.

[19] S. S. Narayanan, A. Toutios, V. Ramanarayanan, A. Lammert, J. Kim, S. Lee, K. S. Nayak, Y.-C. Kim, Y. Zhu, L. Goldstein, D. Byrd, E. Bresch, P. K. Ghosh, A. Katsamanis, and M. I. Proctor, "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)," *J. Acoustical Society of America*, vol. 136, no. 3, pp. 1307–1311, 2014.

[20] E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," *IEEE Trans. Medical Imaging*, vol. 28, no. 3, pp. 323–338, 2009.

[21] T. Sorensen, A. Toutios, L. Goldstein, and S. Narayanan, "Characterizing vocal tract dynamics with real-time MRI," in *LabPhon*, 2015.

[22] A. Toutios and S. Narayanan, "Factor analysis of vocal-tract outlines derived from real-time magentic resonance imaging data," in *Int. Congress Phonetic Sciences*, 2015, pp. 523–532.