ON THE NATURE OF DATA-DRIVEN PRIMITIVE REPRESENTATIONS OF SPEECH ARTICULATION

Vikram Ramanarayanan, Maarten Van Segbroeck and Shrikanth S. Narayanan

Signal Analysis and Interpretation Lab,
University of Southern California, Los Angeles, CA – 90089

ABSTRACT

A long standing view in speech production research posits that articulatory representations are low dimensional. Conceptual and computational models have been built based on this view. In this work we explore the nature of low dimensional representations derived directly from articulatory signals based on sparsity constraints. Specifically, we present a method to examine how well derived representations of "primitive movements" of speech articulation can be used to classify broad phone categories. We first extract these spatio-temporal primitives from a data matrix of human speech articulation data using a weakly-supervised learning method that attempts to find a part-based representation of the data in terms of basis units (or primitives) and their corresponding activations over time. For each phone interval, we then derive a feature representation that captures the co-occurrences between the activations of the various bases over different time-lags. We show that this feature, derived entirely from activations of these primitive movements, is able to achieve an accuracy of about 80% on an interval-based phone classification task. We discuss the implications of these findings in furthering our understanding of speech signal representations.

Index Terms— speech communication, movement primitives, phone classification, motor theory, information transfer.

1. INTRODUCTION

The motor theory of speech perception [1] states that the objects of speech perception are the intended gestures of the speaker, represented as invariant motor commands for linguistically significant movements. One of the implications of the theory is that the human speech production system must produce just the right maneuvers to fit the demands of the categories imposed by the auditory system. Recently researchers have presented evidence in favor of this [2, 3], showing that processing speech signals using an auditory cochlea-like filterbank preserves maximal mutual information between articulatory gestures and the processed speech signals. This suggests that speech gestures and the auditory system are well matched to one another and that the filtering properties of the human auditory system maximally preserve information about speech gestures, which is in accordance with the predictions of the motor theory. In this work, we would like to test a complementary hypothesis with respect to the speech *production* system. From a linguistics perspective, Articulatory Phonology [4] theorizes that the act of speaking is decomposable into units of vocal tract action called "gestures," and suggests that lexical items are assembled from these dynamic primitive units, i.e., constriction actions of the vocal organs. Note that these representations are essentially low dimensional in nature. The above ideas suggest that speech gestures are produced so that they can distinguish between broad phonetic categories imposed by auditory systems. Here we investigate whether articulatory representations derived by imposing constraints on signal properties can be explained in the light of conceptual proposals such as articulatory gestures.

There is also a strong case for using speech production knowledge to inform and improve speech technology applications such as automatic speech recognition; finding efficient representations is a key building block for such an effort [5]. Some reasons for this include: (i) improved noise robustness [6], (ii) better performance on spontaneous speech which exhibits a greater degree of coarticulation due to factored representations [7, 8, 9], (iii) better modeling of different sources of variability, e.g., morphology [10], (iv) provision of a complementary view of the information captured by acoustic features [11], and (v) the significantly lower-dimensional space of articulatory-based feature representations [4, 12]. To motivate the final argument in particular, observe that the speech signal at the acoustic level has a much higher bit rate (e.g., 64 kbits/sec assuming 8 kHz sampling rate and 8 bits/sample encoding) as compared to that of the underlying sound patterns that have an information rate of less than a 100 bits/sec [13]. The presence of this large redundancy in the speech signal means that we first need to extract a lower-dimensional representation of the signal that best captures the discriminative information required for a given task at hand. For example, in the case of a phone discrimination task, we would want to extract a representation that is able to capture the differences between various sounds in a language.

We gratefully acknowledge the support of NIH Grant R01 DC007124-01. We would also like to thank Prasanta Ghosh for help with the EMA data processing.

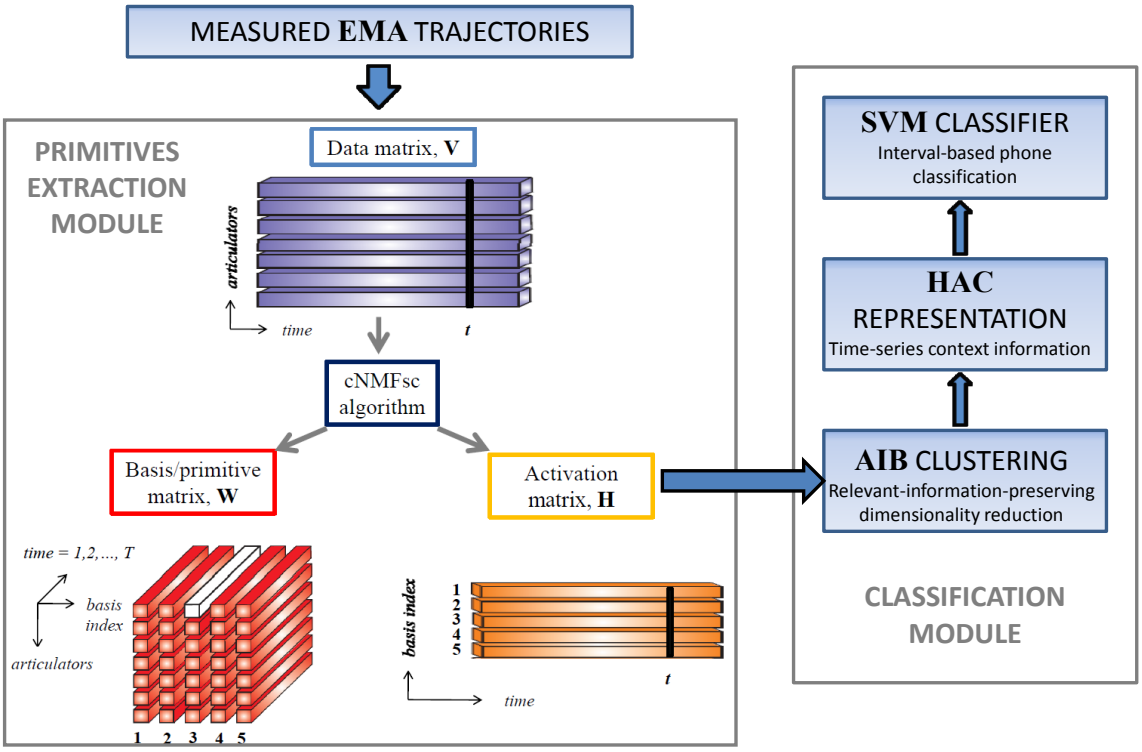

Fig. 1: Schematic of the experimental setup. The input matrix $\mathbf{V}$ is constructed either from real (EMA) articulatory data. In this example, we assume that there are $M = 7$ articulator fleshpoint trajectories. We would like to find $K = 5$ basis functions or articulatory primitives, collectively depicted as the big red cuboid (representing a three-dimensional matrix $\mathbf{W}$). Each vertical slab of the cuboid is one primitive (numbered 1 to 5). For instance, the white tube represents a single component of the 3^{rd} primitive that corresponds to the first articulator (T samples long). The activation of each of these 5 time-varying primitives/basis functions is given by the rows of the activation matrix $\mathbf{H}$ in the bottom right hand corner. The activation matrix is used as input to the classification module, which consists of 3 steps – (i) dimensionality reduction using agglomerative information bottleneck (AIB) clustering, (ii) conversion to a histogram of cooccurrence (HAC) representation to capture dependence information across timeseries, and (iii) a final support vector (SVM) classifier.

Extracting such a representation from acoustic data is not straightforward. However, if we are able to extract from speech discriminative information about articulatory gestures (see [4]), which we know are useful in distinguishing different sounds in a language, we might be better positioned to solve this problem. In this paper, we explore the question of how well *low-dimensional* "articulatory movement primitives" derived from data by imposing sparsity constraints can discriminate between broad phone categories. Articulatory movement primitives (or, exemplars) may be defined as a dictionary or template set of articulatory movement patterns in space and time, weighted combinations of the elements of which can be used to represent the complete set of coordinated spatio-temporal movements of vocal tract articulators required for speech production [14, 15]. Although we do not claim that this is a completely validated model for human speech production, such a representation captures information regarding movement synergies, i.e., combinations that simplify the production of movements by reducing the degrees of freedom that need to be specified by the motor control system [16].

Figure 1 presents a schematical overview of the paper. We describe the articulatory data used for experiments in Section 2. Sections 3 and 4 present the mathematical formalism used for primitive extraction and a brief quantitative evaluation of the extraction procedure respectively. Next, in Section 5, we describe the classification setup including appropriate feature preprocessing steps. Finally, we present our experimental observations along with a brief discussion of possible implications in Sections 6 and 7.

2. DATA

We analyze ElectroMagnetic Articulography (EMA) data from the Multichannel Articulatory (MOCHA) database [17], which consists of data from two (British English) speakers - one male and one female. Acoustic and articulatory data were collected while each speaker read a set of 460 phonetically-diverse TIMIT sentences. The articulatory channels include EMA sensors directly attached to the upper and lower lips, lower incisor (jaw), tongue tip (5-10mm from the tip), tongue blade (approximately 2-3cm posterior to the tongue tip sensor), tongue dorsum (approximately 2-3cm posterior to the tongue blade sensor) and soft palate. Each articulatory channel was sampled at 500 Hz with 16-bit precision and zero-phase low-pass filtered with a cut-off frequency of 35 Hz [18]. Next, for every utterance, we subtracted the mean value from each articulatory channel [19, 18]. Then we added the mean value of each channel averaged over all utterances to that corresponding channel. Finally, we downsampled each channel by a factor of 5 to 100 Hz and further normalized data in each channel (by its range) such that all data values lie with the range [0,1].

3. EXTRACTION OF PRIMITIVE MOVEMENTS

Modeling data vectors as sparse linear combinations of basis elements is a general computational approach[1] which we will use to solve our problem [20, 21, 22, 23, 24]. If x_1, x_2, ..., x_M are the M time-traces (represented as column vectors of dimension $N \times 1$) of EMA articulator trajectory variables, then we can design our data matrix $\mathbf{V}$ to be:

$$\mathbf{V} = [x_1 | x_2 | \ldots | x_M]^\dagger \in R^{M \times N} \quad (1)$$

where $\dagger$ is the matrix transpose operator. We will use convolutive nonnegative matrix factorization or cNMF [22] to solve our problem. cNMF aims to find an approximation of the data matrix $\mathbf{V}$ using a basis tensor $\mathbf{W}$ and an activation matrix $\mathbf{H}$:

$$\mathbf{V} = \sum_{t=0}^{T-1} \mathbf{W}(t) \cdot \vec{\mathbf{H}}^t \quad (2)$$

where each column of $\mathbf{W}(t) \in R^{\geq 0, M \times K}$ is a time-varying basis vector sequence, each row of $\mathbf{H} \in R^{\geq 0, K \times N}$ is its corresponding activation vector (h_i is the i^{th} row of $\mathbf{H}$), T is the temporal length of each basis (e.g., no. of data samples or frames), and the $\vec{(\cdot)}^k$ operator is a shift operator that moves the columns of its argument by k spots to the right, as detailed in [22]. In order to derive primitives that are maximally discriminative of different phone classes, we augmented the data matrix $\mathbf{V}$ with phone label information (after [25]):

$$\begin{bmatrix} \mathbf{V} \\ \mathbf{V_{lab}} \end{bmatrix} = \sum_{t=0}^{T-1} \begin{bmatrix} \mathbf{W}(t) \\ \mathbf{W_{lab}}(t) \end{bmatrix} \cdot \vec{\mathbf{H}}^t \quad (3)$$

where each column of $\mathbf{V_{lab}}$ is a 40×1 vector whose entries are all 0 save for one – we set the entry corresponding to the phone label of the current frame[2] to 1 (there are 40 phone labels in all annotated for this dataset). To force the training algorithm to extract one unique primitive for each phone, we (i) added a (weak) supervision step to the multiplicative update rules of the cNMF training algorithm by forcing the $\mathbf{W_{lab}}$ matrix to be a 40×40 identity matrix, and (ii) set the number of primitives K equal to the number of unique phone classes (40). We further add a sparsity constraint on the rows of the activation matrix to obtain the final formulation of our optimization problem, termed cNMF with sparseness constraints (or cNMFsc) [14, 15]:

$$\min_{\mathbf{W},\mathbf{H}} \| \begin{bmatrix} \mathbf{V} \\ \mathbf{V_{lab}} \end{bmatrix} = \sum_{t=0}^{T-1} \begin{bmatrix} \mathbf{W}(t) \\ \mathbf{W_{lab}}(t) \end{bmatrix} \cdot \vec{\mathbf{H}}^t \|^2 \text{s.t. } sparseness(h_i) = S_h, \forall i. \quad (4)$$

Note that the level of sparseness ($0 \leq S_h \leq 1$) is user-defined. See Ramanarayanan et al. [14, 15] for the details of an algorithm that can be used to solve this problem.

[1] This approach is termed variously as dictionary learning or sparse coding or sparse matrix factorization depending on the exact problem formulation.

[2] Phone labels of each frame were obtained through automatic phonetic alignment of the data.

4. ALGORITHM PERFORMANCE

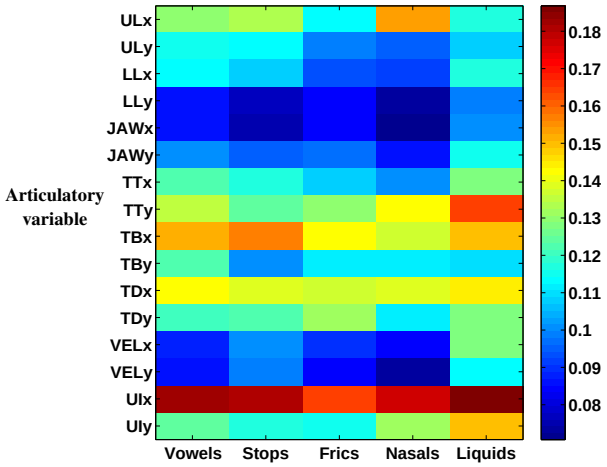

Fig. 2: Root mean squared error (RMSE) for each articulator and broad phone class obtained as a result of running the algorithm on all 460 sentences spoken by male speaker *msak0*.

In order to choose model parameters appropriately, we computed the Akaike Information Criterion (AIC, [26]), which overwhelmingly prefered parameter values that resulted in low model complexity. Based on this analysis we decided to set the temporal extent of each basis sequence (T) to 10 samples (since this corresponds to a time period of approximately 100ms, factoring in a sampling rate of 100 samples per second) to capture effects of the order of the length of a phone on average. As mentioned earlier, we chose the number of bases, K, to be equal to the number of phone classes, i.e., 40. The sparseness parameter S_h was set to 0.65 based on experiments with synthetic data.

In order to see how the algorithm performs for different phone classes, we first performed a phonetic alignment of the audio data corresponding to each set of articulator trajectories (using the Hidden Markov Model toolkit [27, HTK]) to enable association with different phone classes. Figure 2 shows the root mean squared error (RMSE) for each articulator and broad phone class for MOCHA-TIMIT speaker *msak0*. Recall that since we are normalizing each row of the original data matrix to the range $[0, 1]$ (and hence each articulator trajectory), the error values in Figure 2 can be read on a similar scale. We see that in general, error values are high. The errors were highest ($0.13 - 0.2$) for tongue-related articulator trajectories and the upper incisor variable. On the other hand, trajectories of the lip (LLx and LLy) and jaw ($JAWx$ and $JAWy$) sensors were reconstructed with lower error (≤ 0.1). We further computed the fraction of variance that was not explained (FVU) by the model for each sentence in the database. The histograms of these distribution are plotted in Figure 3. The mean and standard deviation of this distribution was 0.079 ± 0.028 for speaker *msak0* (i.e., approx. 7.9% of the original data variance was not accounted for on

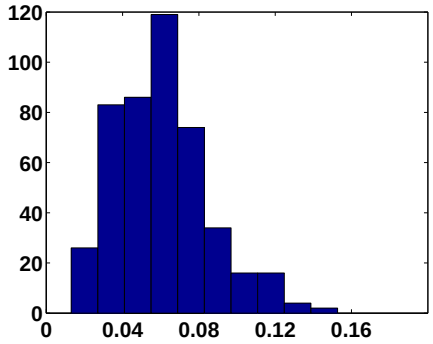

Fig. 3: Histograms of the fraction of variance unexplained (FVU) by the proposed cNMFsc model for MOCHA-TIMIT speaker *msak0*. The samples of the distribution were obtained by computing the FVU for each of the 460 sentences. (The algorithm parameters used in the model were $S_h = 0.65$, $K = 40$ and $T = 10$).

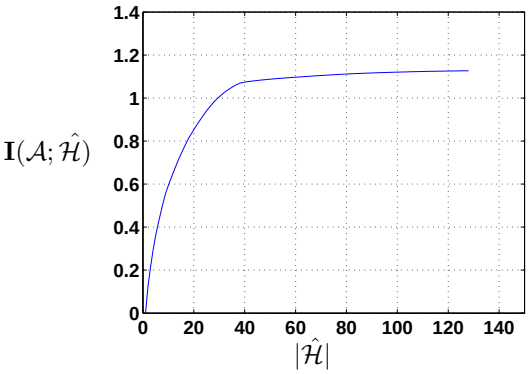

Fig. 4: Mutual information $\mathbf{I}(\mathcal{A};\hat{\mathcal{H}})$ between quantized activation space $\hat{\mathcal{H}}$ and the space of acoustic features $\mathcal{A}$ as a function of the cardinality of $\hat{\mathcal{H}}$ (in other words, the number of quantization levels).

average). These statistics suggest that the cNMFsc model accounts for more than 90% of the original data variance.

5. BROAD PHONE CLASSIFICATION SETUP

In this section, we describe how activation matrices obtained using the algorithm described above are transformed into features suitable for phone classification experiments. We can hypothesize the sequence of phones corresponding to a given utterance along with their corresponding time-boundaries by phonetically aligning the audio. Therefore in this work, the phone categories are entirely based on categorical information obtained from the audio signal.

Since the activation matrices are sparse by formulation, it does not make sense to use columns of the activation matrix (one per frame) as feature prototypes in a frame-based phone classification experiment (since there will be zeros corresponding to time-frames where no basis is activated). Instead, we choose to compute *one* feature per phone interval. This way, we are formulating the classification problem as an *interval*-based phone classification experiment. Therefore, given a segment of activation columns for a given phone interval (i.e., a block subset of columns of the activation matrix), we have to compute a single feature. First, we quantize the space of activation vectors (columns of the activation matrix) to generate a codebook representation of the time-series using an agglomerative information bottleneck-based clustering technique; second, we compute histograms of co-occurrences (denoted HAC [28]) of the codebook indices over the time-series[3]. HAC representations are useful since they explicitly model cooccurrences of articulatory feature instances over time. We describe the procedure in more detail below.

[3]Notice that the initial quantization step is needed because the column entries are not discrete-valued, making it impractical to compute meaningful co-occurrences directly.

5.1. Codebook generation

We perform vector quantization (VQ) of the columns of the activation matrices using the agglomerative information bottleneck (AIB) principle [29]. We formulate the problem as that of finding a quantization or a compressed representation $\hat{\mathcal{H}}$ of the activation space $\mathcal{H}$ that minimizes the mutual information $\mathbf{I}(\mathcal{H};\hat{\mathcal{H}})$ between them, while simultaneously maximizing the mutual information $\mathbf{I}(\mathcal{A};\hat{\mathcal{H}})$ between $\hat{\mathcal{H}}$ and the space of acoustic features $\mathcal{A}$. In other words, we would like to find that quantization of the duration space that achieves maximal compression while retaining as much discriminative information as possible about acoustic features[4]. We use the VLFeat software [30] to perform this clustering. Figure 4 plots the mutual information $\mathbf{I}(\mathcal{A};\hat{\mathcal{H}})$ as a function of the number of clusters/codebook entries. We observe a rapid drop in mutual information as the number of clusters drops below 20. Based on empirical observation of this graph, we choose a codebook size of 32 clusters for our experiments.

5.2. Computing histograms of co-occurrences

We first replace each frame of the activation matrix $\mathbf{H}$ with the best matching centroid of the codebook. This way, activation matrix is now represented by a single row vector of VQ-labels, $\mathbf{H}_{quant}$. A HAC-representation of *lag* τ is then defined as a vector where each entry corresponds to the number of times all pairs of VQ-label are observed τ frames apart. In other words, we construct a vector of lag-τ co-occurrences where each entry (m, n) signifies the number of times that the input sequence of activation frames is encoded into a VQ-label m at time t (in the row vector H_{quant}), while encoded into VQ-label n at time $t+\tau$ [25]. By stacking all (m, n) combinations, each phone interval can be represented by a single column vector where the elements express the sum of all K^2

[4]Note that we aren't *adding* any extra info from acoustics to the activation features obtained from articulatory data. We are just clustering it differently using acoustic information. Thus the argument that we are using only articulatory information to cluster phone categories still holds water.

possible lag-τ co-occurrences (where K is the number of VQ clusters). We can repeat the procedure for different values of τ, and stack the results into one "supervector". Note however, that the dimensionality of the HAC feature increases by a factor of K^2 for each lag value τ that we want to consider. In our case, we empirically found that choosing four lag values of 2, 3, 4 and 5 frames worked well.

5.3. Classification experiments

We used support vector machine (SVM) classifiers to perform classfication experiments [31]. We experimented with both linear as well as radial basis function (RBF) kernels and empirically found that the former gave better classification accuracy. This could be due to the large dimensionality of the HAC feature space. Hyperparameters were tuned using a grid-search method.

6. OBSERVATIONS AND RESULTS

Table 1 shows the performance of the activation features (after appropriate HAC-feature transformation) on an interval-based phone classification task. Also shown for comparison purposes are the performances of the raw EMA pellets themselves, as well as mel-frequency cepstral coefficient features (13-dimensional) on the same task. Initial experiments suggest that the activation features learnt by the cNMFsc algorithm significantly outperform both raw MFCC and raw EMA features in terms of classification accuracy.

For a deeper understanding of what the classification accuracy numbers in Table 1 actually mean, we also computed the entropy of each feature set and mutual information[5] (MI) between each feature set and the phone labels. We observe that although the entropy (and consequently bit rate assuming a fixed encoding scheme) of primitive activation features is lower than that of the raw MFCC or EMA features, the mutual information between the phone labels and the different features considered is still comparable. This, along with the weak supervision during the learning process, suggests that primitive activations are a useful, low-dimensional representation capable of disciminating phone classes. In addition, we can see that although the MFCC and EMA features have a similar entropy value, the former has a higher MI. This is in agreement with the observation of a higher classification accuracy. The challenge for future work will be finding representations that push the classification accuracy envelope while minimizing the required bitrate.

[5]To estimate the probability of a given feature value: (i) We clustered the data using k-means clustering ($K = 128$) and assigned each feature to a cluster. (ii) We set the probability of occurrence of a feature to be equal to the (maximum likelihood estimate of the) probability of occurrence of its corresponding cluster.

Table 1: Performance of various features on a interval-based phone classification experiment (after appropriate transformation to HAC-representations). For clarity of understanding we also show the entropy of the feature set along with the mutual information between the feature set and classification labels $\mathcal{L}$ in each case.

Feature set $\mathcal{X}$	Class. Acc. (%)	$\mathbf{H}(\mathcal{X})$	$\mathbf{I}(\mathcal{X};\mathcal{L})$
MFCC	71%	6.9	1.68
Raw EMA pellets	61.78%	6.9	1.59
Primitive activations	80.59%	6.5	1.63
Phone labels $\mathcal{L}$	100%	4.9	4.9

7. DISCUSSION AND OUTLOOK

Our results suggest that articulatory movement primitives offer information for discriminating between broad phone classes. It is important to note that the performance of these features is contingent upon the way they are extracted, and therefore, algorithmic choices, such as the sparseness value S_h, number of primitives, and the temporal extent of each primitive, will greatly influence the outcome of subsequent classification experiments[6]. Having said that, it is encouraging to observe that the computationally estimated lower-dimensional primitive representations of speech articulation contain useful information to distinguish between broad phone categories. The question that remains is in explaining the gap between the abstract (conceptual) low dimensional production representations and those derived by imposing sparseness constraints on observed movement data.

A partial answer to this may arise from the fact that the (EMA) articulatory data used in the experiments offers only a limited view of the complex articulatory mechanisms. Although they do encode information about phonetic categories, these movements represent only a part of the picture with respect to the phonetic categories. But more generally, the Motor Theory of speech perception [1] suggests that the human speech production system must produce just the right maneuvers to fit the demands of the categories imposed by the auditory system. Assuming that articulatory movement primitives can be considered as a surrogate for atleast a subset of these maneuvers, our results are in agreement with the theory. An exciting future research direction that this sets up is understanding whether information transfer during speech production is performed so as to effect efficient perception of auditory categories. The work described in this paper, along with other efforts (such as [2, 3]), are an initial step at answering these questions and could open up new avenues of research into speech production.

[6]The primitives we extract will depend on the cost function that we formulate and optimize. Development of better problem formulations and algorithms to extract primitives is an exciting area of ongoing and future research.

8. REFERENCES

[1] A.M. Liberman and I.G. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, no. 1, pp. 1–36, 1985.

[2] P.K. Ghosh, L.M. Goldstein, and S.S. Narayanan, "Processing speech signal using auditory-like filterbank provides least uncertainty about articulatory gestures," *The Journal of the Acoustical Society of America*, vol. 129, pp. 4014, 2011.

[3] A. Bertrand, K. Demuynck, V. Stouten, and H. Van hamme, "Unsupervised learning of auditory filter banks using non-negative matrix factorisation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4713–4716.

[4] C.P. Browman and L. Goldstein, "Dynamics and articulatory phonology," *In T. van Gelder and B. Port (Eds.), Mind as motion: Explorations in the dynamics of cognition*, pp. 175–193, 1995.

[5] V. Ramanarayanan, P. Ghosh, A. Lammert, and S. Narayanan, "Exploiting speech production information for automatic speech and speaker modeling and recognition – possibilities and new opportunities," in *Fourth Annual Conference of the Asia-Pacific Signal and Information Processing Association*, 2012.

[6] R.C. Rose, J. Schroeter, and MM Sondhi, "The potential role of speech production models in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 99, pp. 1699–1709, 1996.

[7] L. Deng, G. Ramsay, and D. Sun, "Production models as a structural basis for automatic speech recognition," *Speech Communication*, vol. 22, no. 2, pp. 93–111, 1997.

[8] E. Farnetani, "Coarticulation and connected speech processes," *The handbook of phonetic sciences*, pp. 371–404, 1997.

[9] E. Mcdermott and A. Nakamura, "Production-oriented models for speech recognition," *IEICE transactions on information and systems*, vol. 89, no. 3, pp. 1006–1014, 2006.

[10] A. Lammert, M. Proctor, and S. Narayanan, "Morphological variation in the adult vocal tract: A study using rtmri," *Proc. 9th ISSP*, 2011.

[11] R. Arora and K. Livescu, "Multi-view cca-based acoustic features for phonetic recognition across speakers and domains," in *Int. Conf. on Acoustics, Speech, and Signal Processing*, 2013.

[12] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, pp. 723–742, 2007.

[13] B.S. Atal, "Automatic speech recognition: A communication perspective," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing- Proceedings*, 1999, vol. 1, pp. 457–460.

[14] V. Ramanarayanan, A. Katsamanis, and S. Narayanan, "Automatic data-driven learning of articulatory primitives from real-time mri data using convolutive nmf with sparseness constraints," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[15] V. Ramanarayanan, L. Goldstein, and S. Narayanan, "Articulatory movement primitives – extraction, interpretation and validation," *The Journal of the Acoustical Society of America*, 2013 (accepted).

[16] J.A.S. Kelso, "Synergies: atoms of brain and behavior," *Progress in motor control*, pp. 83–91, 2009.

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

[18] P.K. Ghosh and S. Narayanan, "A generalized smoothness criterion for acoustic-to-articulatory inversion," *The Journal of the Acoustical Society of America*, vol. 128, pp. 2162–2172, 2010.

[19] K. Richmond, *Estimating articulatory parameters from the acoustic speech signal*, Ph.D. thesis, University of Edinburgh, 2002.

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

[21] A. d'Avella and E. Bizzi, "Shared and specific muscle synergies in natural motor behaviors," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 8, pp. 3076, 2005.

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

[23] 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, 2008.

[24] T. Kim, G. Shakhnarovich, and R. Urtasun, "Sparse coding for learning interpretable spatio-temporal primitives," *Advances in neural information processing systems*, vol. 22, 2010.

[25] M. Van Segbroeck and H. Van hamme, "Unsupervised learning of time–frequency patches as a noise-robust representation of speech," *Speech Communication*, vol. 51, no. 11, pp. 1124–1138, 2009.

[26] H. Akaike, "Likelihood of a model and information criteria," *Journal of Econometrics*, vol. 16, no. 1, pp. 3–14, 1981.

[27] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, XA Liu, G. Moore, J. Odell, D. Ollason, D. Povey, et al., "The htk book (for htk version 3.4)," 2006.

[28] H. Van hamme, "HAC-models: a novel approach to continuous speech recognition," in *Interspeech*, 2008.

[29] N. Slonim and N. Tishby, "Agglomerative information bottleneck," *Advances in Neural Information Processing Systems*, vol. 12, pp. 617–623, 1999.

[30] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[31] C.C. Chang and C.J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.