AN APPROACH TOWARD UNDERSTANDING THE INVARIANT AND VARIANT ASPECTS OF SPEECH PRODUCTION USING LOW-RANK–SPARSE MATRIX DECOMPOSITIONS

Vikram Ramanarayanan and Shrikanth S. Narayanan

Signal Analysis and Interpretation Lab, Ming Hsieh Department of Electrical Engineering,
University of Southern California, Los Angeles, CA - 90089
vramanar@usc.edu, shri@sipi.usc.edu

ABSTRACT

An understanding of the invariant aspects of speech articulation is vital for the development of more accurate models of speech production. It is also equally important to find and describe the sources of variability in production to completely understand the cognitive processes behind the speech code. In this paper, we propose a novel application of low-rank–sparse matrix decomposition and signal warping techniques on real-time magnetic resonance image sequences obtained during human speech production to decompose those articulatory data into "invariant" and "variant" sequences. These sequences can then be used in linguistic analysis or statistical modeling.

Index Terms— Speech production, invariance, real-time MRI, sparsity, low-rank matrices, signal warping

1. INTRODUCTION

Separating the invariant aspects of speech production from those that arise due to variabilities such as articulation differences, speaker-dependent vocat tract morphology, head orientation and noise is vital from both a speech science and a speech technology perspective. For example, in phonetics research, it is of interest to see how speakers produce different sounds differently maintaining linguistically-relevant invariant aspects of their articulation. A potential application of such a decomposition in the long term would be towards understanding the intricacies of speech planning and execution. Understanding the invariant aspects of the executed speech plan while simultaneously extracting variabilities in this execution might give us better insights into the planning process. On the technological front, such a decomposition could find use as a pre-processing step in the statistical modeling of speech production. It is often desirable to work with just the invariant aspects of speakers' production so as to control for the variability that poses problems for pattern learning algorithms. In addition, such a technique might help us in developing a more optimal representation of the speech signal in quantitatively linking speech generation and processing from the articulatory as well as acoustic domains. In this paper, as a first step, we propose a method for extracting the invariant and variable aspects of speech articulation using data drawn from real-time magnetic resonance imaging technology.

Real-time magnetic resonance imaging (rt-MRI) [1] provides a powerful tool to examine the choreography of all articulators in the vocal tract as speech is produced. 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., and is thus well-suited for our purposes of examining invariant aspects of articulation.

A novel result that has come out of the recent literature on sparsity theory is that if we have a data matrix which can be modeled as the superposition of a low-rank component and a sparse component, it is possible (under some conditions) to *exactly* recover these components by solving a convex optimization program that minimizes a weighted sum of the nuclear norm and the l_1 norm [2, 3, 4]. This can be thought of as an approach to robust principal component analysis, where we can recover the principal components of a data matrix even though a positive fraction of its entries are arbitrarily corrupted. For our purposes, this technique can be applied to a matrix composed of several images of the same vocal tract posture(re-formed into column vectors) obtained through magnetic resonance imaging (MRI), since we would like to capture what is invariant about the vocal-tract postures in each image while separating out the effects due to slight postural variations, noise and imaging artifacts.

However, since speech production is an inherently dynamic process, the analysis of static postures (captured by single MRI images) is not enough. Indeed we would like to compare several *sequences of images* which could be of varying lengths and compute what is invariant and variant about those dynamic production sequences. Hence, as a first step, we resort to signal time-warping techniques like Dynamic Time Warping (DTW), Derivative Dynamic Time Warping (DDTW) or Canonical Time Warping (CTW) (see [5] for a

Work described in this paper was supported by NIH grant DC007124, the USC Imaging Sciences Center, and the USC Center for High Performance Computing and Communications (HPCC).

review) in order to transform these image sequences into sequences of equal length, following which sparse–low-rank matrix decomposition techniques are applied individually to the corresponding images of every sequence (at each time epoch) to obtain a final "invariant" sequence.

The rest of the paper is organized as follows: Section 2 details the mathematical formulation of the low-rank—sparse decomposition and how it is equivalent to extracting the invariant and variant aspects of articulation from MR images. Section 3 lays out the data used and experiments performed, followed by a discussion of applications and future directions in Section 4.

2. THEORY AND ANALYSIS

2.1. Mathematical formulation

We first consider the case where we would like to find the invariant and variant components associated with a static posture observed multiple times where we have, say, n input MR images corresponding to the canonical vocal posture[1] (an example of such a case might be in phonetic analysis, where we may want to analyze vocal tract shaping during a stop sound at the point of closure over multiple instances). If I_1, I_2, ..., I_n are the n MR images of the posture (dimension $n_1 \times n_2$) re-formed into $m \times 1$ column vectors (where $m = n_1 \times n_2$), then we can write:

$$D = [I_1 \,|I_2\,| \ldots |I_n\,] \in R^{m \times n} \tag{1}$$

Recall that we would like to obtain a decomposition:

$$D = L + S \tag{2}$$

where L is the low-rank "invariant" component and S is the sparse "variable" component. Notice that in the case of the former, we would like the resulting matrix L to be of the least possible rank (since that would imply that the data matrix L can be described by the least number of principal components (images)). Also note that S models these variabilities that corrupt the images, with the assumption that only a small fraction of all pixels in an image are corrupted thus, allowing us to model them as sparse errors whose non-zero entries can have arbitrarily large magnitudes [2, 3].

As the authors in the above-cited papers mention, our problem amounts to solving the following optimization problem:

$$\min_{L,S} \; rank(L) + \gamma \|S\|_0 \; s.t. \; D = L + S \tag{3}$$

where γ is a parameter which trades-off the rank of the solution versus the sparsity of the error and $\|S\|_0$ is the l_0-norm of the sparse matrix component S. The non-convexity of the above optimization problem is alleviated by means of a convex relaxation step which replaces the $rank(\cdot)$ term by the nuclear norm (or sum of singular values) of L and the l_0-norm of S by it's l_1-norm instead as follows:

$$\min_{L,S} \; \|L\|_* + \lambda \|S\|_1 \; s.t. \; D = L + S \tag{4}$$

where the weight parameter λ is generally selected as $\frac{1}{\sqrt{max(m,n)}}$. Note that this choice is weighting parameter is *optimal*; hence a major advantage of this algorithm is the *lack* of any tuning parameters [2, 3]. The convex program given by equation 4 can be solved by a variety of techniques using software made available online by the authors.

2.2. Conditions for separation

An important condition for achieving this decomposition is that the data matrix D should not be *both* low-rank and sparse [2]. In other words, an incoherence condition must be satisfied that prevents the singular vectors of the low-rank component from being sparse, i.e., they are reasonably spread out. If we write the singular value decomposition of L as :

$$L = U \Sigma V^* = \sum_{i=1}^{r} \sigma_i u_i v_i^* \tag{5}$$

where r is the rank of the matrix, $\sigma_1, \ldots, \sigma_r$ are the positive singular values and U and V are the left and right singular-vector matrices respectively. Then the incoherence condition with parameter μ states that:

$$\max_i \|U^* e_i\|_2^2 \leq \frac{\mu r}{m}, \max_i \|V^* e_i\|_2^2 \leq \frac{\mu r}{n}, \|UV^*\|_\infty \leq \sqrt{\frac{\mu r}{mn}}. \tag{6}$$

where $\|M\|_\infty = \max_{i,j} |M_{ij}|$ is the l_∞-norm of M seen as a long vector and e_i is a standard basis vector. In the cases we analyze in this paper, this condition is satisfied for a minimum incoherence parameter value of $\mu = 1$ ($n_1 = n_2 = 68$).

Also note that there exists not one, but a set of low-rank and sparse matrices will satisfy equation (2); however we are interested in finding the ones that allow L to be of least rank and at the same time, S to be as sparse as possible in the minimum l_1-norm sense because (i) we would like L not to be corrupted by any gross pixel corruptions or image intensity fluctuations (which the above-mentioned formulation is robust to [2, 3]), and (ii) we may not know apriori as to how many "invariant" components a set of images may contain.

2.3. Extension to sequences

Notice that so far we have only considered extracting the invariant and variant aspects of a single posture. However, we want to capture this information about the speech signal as it

[1] It makes sense to see what's varying and what's invariant about any signal in a data-driven manner only if we have more than 1 realization of the same signal.

unfolds in time. In this case also, we need many repetitions of the same utterance of interest in order to see what is invariant and variant about it's production. An important problem that arises here is that different realizations of the same utterance can be of different lengths. One solution to this problem is to align or warp these sequences to a single canonical length using signal time-warping techniques like Dynamic Time Warping (DTW)(see [5]), and performing the above-explained low-rank–sparse decomposition on images of every sequence at each time-epoch, such that we get a "low-rank" sequence and a "sparse" sequence.

3. DATA AND EXPERIMENTS

Midsagittal real-time MR images of the vocal tract were acquired with a repetition time of TR=6.5ms on a GE Signa 1.5T scanner with a 13 interleaf spiral gradient echo pulse sequence. The slice thickness was approximately 3mm. A sliding window reconstruction at a rate of 22.4 frames per second was employed. Field-of-view (FOV), which can be thought of as a zoom factor, was set depending on the subjects head size. Details regarding the recording and imaging setup can be found in [1, 6], and sample MRI movies can be found at http://sail.usc.edu/span.

We tested this approach on 7 repetitions of the utterance "pa-sop" produced by a male American English speaker. In this and many other cases in linguistic analysis, it is of interest to examine the invariant and variable aspects of the shaping of specific speech sounds (an alveolar sibilant fricative, in this example). In this study, the D matrix at *each* time epoch of the sequence was of dimension 4624×7 (one image for each rep). The value of the λ parameter in the optimization program formulated in Section 2 was chosen optimally as $\sqrt{\frac{1}{4624}} = 0.0147$, i.e., *no* parameters of this algorithm needed to be tuned explicitly. The dimension of the sparse matrix S was always 7 while that of the low-rank matrix L was 2 or 3. The maximum per-pixel reconstruction error, $D-(L+S)$ obtained was of the order of 10^{-6}, which gives an indication of the high accuracy of the method. The results are shown in Figure 1. The original set of 7 sequences of varying lengths were aligned using DTW (the other techniques produced similar results to those obtained with DTW for this particular dataset) to the same length of 16 image frames. For better visualization purposes, the low-rank and sparse frames computed at each epoch were each summed and normalized such that all pixel values lay in the interval $[0, 1]$ to obtain one normalized-sum low-rank image and one normalized-sum sparse image at each time epoch. This process was done for each of the 16 time epochs to obtain a sequence of normalized-sum low-rank and sparse images, which are plotted in Figure 1. We observe that the low-rank sequence (penultimate row) characterizes the canonical articulation of the /pa sop/ sequence by this speaker which is relatively invariant across the different realizations, with the slight variabilities due to vocal tract posture (and noise to a certain extent) being captured by the corresponding sparse component sequence (bottom row). The sparse sequence also captures other errors, such as those obtained during transcription or alignment (for example, in this case, we see slight disagreements with the segmentation of the start and end frame of the lip closure which show up in the sparse component along with other postural differences, such as the slighly more raised lip and tongue front position observed in some realizations of the sequence). We can extend this technique other utterances in continuous speech to understand how the speaker *generally* shapes his vocal tract to produce a particular utterance, allowing us to build more accurate models of that speakers' speech production mechanisms. In addition, this decomposition allows us to look at the "sparse" sequences and model the variabilities in production separately, if needed.

In another study, we tested out this approach in obtaining qualitative differences between images of postures assumed by one female American English speaker during pauses in read speech (27 images) and at absolute rest (6 images). See [7] for experimental details and motivation. Here it is of interest to understand the differences in vocal tract shape assumed and it's variability in the two cases, which may provide insights into the speech planning mechanism. The decomposition was performed for each set of images separately and the low-rank and sparse images obtained as a result of the decomposition were summed and normalized as before to finally obtain a normalized-sum low-rank image and a normalized-sum sparse image for each case. The results are plotted in Figure 2. Notice that while the normalized-sum low-rank images give a clear picture of the canonical shape assumed in each case (the vocal tract is relatively closed in the case of an absolute rest posture, which is significantly different from postures assumed during read inter-speech pauses), their sparse counterparts give us an idea of the amount of variability observed in different regions, with the pixel intensity proportional to the amount of variability observed. For example, in the case of an absolute rest position, we see (from the sparse normalized-sum component) a high degree of variability associated with the tongue dorsum position, which might suggest that this articulator may not be under a high degree of active control by the speech planning mechanism. These results are consistent and complimentary to the work described in [7].

4. CONCLUSIONS AND FUTURE WORK

We have described a procedure to decompose rt-MRI image sequences of the human vocal tract into relatively "invariant" (or low-rank) and "variant" (or sparse) image sequences using a convex optimization formulation that minimizes a weighted sum of the nuclear norm of the low-rank component and the l_1-norm of the sparse component *without* requiring the explicit tuning of any optimization parameters. In addition, the

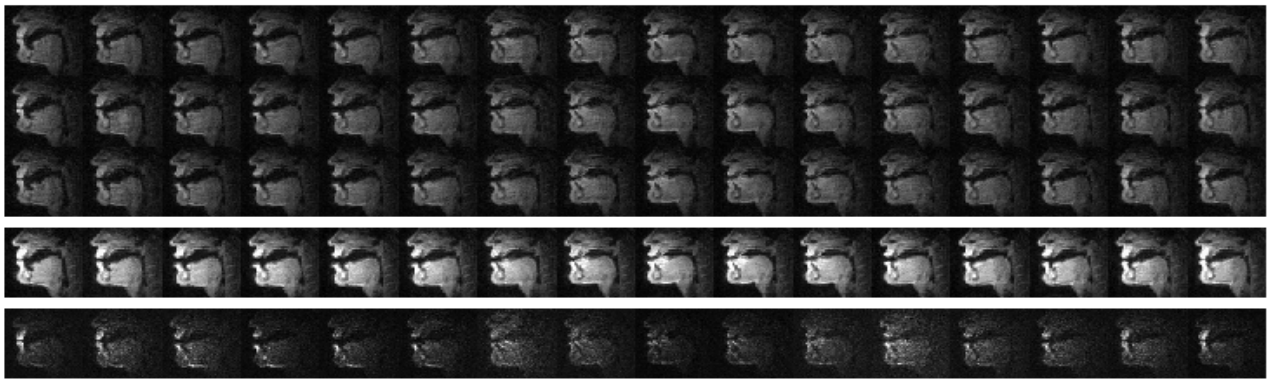

Fig. 1: (Top) 3 out of the original 7 time-aligned set of MR image sequences corresponding to a speaker's production of "pa sop" (after alignment with DTW). (Penultimate row) Sequence of normalized-sum images obtained by summing the low-rank frames of the decomposition at each time epoch. (Bottom row) Corresponding sparse normalized-sum image sequence.

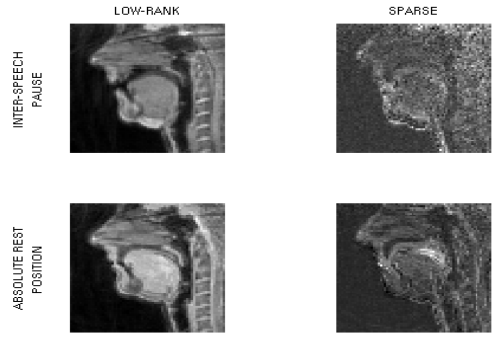

Fig. 2: (Top row) Normalized-sum images obtained by summing the low-rank components (left) and sparse components (right) of vocal postures obtained during an inter-speech pause in read speech. (Bottom row) Similar images for an absolute rest posture.

convex program formulation gives an exact solution (upto 6 significant digits) to the problem and is solvable in real-time. Figures 1 and 2 gives us an indication of the usefulness of this method, both in phonetic analysis as well as a preprocessing step for statistical modeling of articulatory data. However, care must be taken in interpreting the results of this method, since a mathematical formulation of invariance as described in this paper may or may not correspond to invariance in the cognitive speech planning domain. In addition, we implicitly assume that the data matrix of images D is decomposable exactly into a low-rank and a sparse component. In future extensions to this work, we would like to account for speaker-specific anatomy and head movement/rotation so that comparisons across speakers can be performed. Two principal long-term goals of this project are (i) to see what insights this invariance-variance decomposition can give us regarding the planning and execution process for continuous speech of different speakers and speaking styles, and (ii) to obtain better representations of the acoustic speech signal using information about invariant articulatory representations.

5. REFERENCES

[1] 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, pp. 1771, 2004.

[2] E.J. Candes, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis," *preprint*, 2009.

[3] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images," in *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*. Citeseer, 2010.

[4] V. Chandrasekaran, S. Sanghavi, P.A. Parrilo, and A.S. Willsky, "Sparse and low-rank matrix decompositions," in *IFAC Symposium on System Identification*, 2009.

[5] F. Zhou and F. de la Torre, "Canonical time warping for alignment of human behavior," *Advances in Neural Information Processing Systems (NIPS)*, 2009.

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

[7] V. Ramanarayanan, D. Byrd, L. Goldstein, and S. Narayanan, "Investigating articulatory setting - pauses, ready position and rest - using real-time MRI," *Interspeech, Makuhari, Japan*, 2010.