Data-Driven Analysis of Realtime Vocal Tract MRI using Correlated Image Regions

Adam C. Lammert[1], *Michael I. Proctor*[2], *Shrikanth S. Narayanan*[1,2,3]

[1]Department of Computer Science, University of Southern California, USA
[2]Department of Linguistics, University of Southern California, USA
[3]Department of Electrical Engineering, University of Southern California, USA

lammert@usc.edu, mproctor@usc.edu, shri@sipi.usc.edu

Abstract

Realtime MRI provides useful data about the human vocal tract, but also introduces many of the challenges of processing high-dimensional image data. Intuitively, data reduction would proceed by finding the air-tissue boundaries in the images, and tracing an outline of the vocal tract. This approach is anatomically well-founded. We explore an alternative approach which is data-driven and has a complementary set of advantages. Our method directly examines pixel intensities. By analyzing how the pixels co-vary over time, we segment the image into spatially localized regions, in which the pixels are highly correlated with each other. Intensity variations in these correlated regions correspond to vocal tract constrictions, which are meaningful units of speech production. We show how these regions can be extracted entirely automatically, or with manual guidance. We present two examples and discuss its merits, including the opportunity to do direct data-driven time series modeling.

Index Terms: human speech production, phonetics, realtime mri, vocal tract, data reduction

1. Introduction

Realtime Magnetic Resonance Imaging (rtMRI) has been demonstrated as a useful and promising technique for collecting speech production data. It affords vocal tract imaging with good spatial and temporal resolution, as well as providing a full midsagittal view of the vocal tract. However, rich data like these bring all the challenges associated with high-dimensional data (i.e., the curse of dimensionality). For instance, a commonly published rtMRI protocol involves reconstructing images with a frame size of 68×68 pixels. Considering each pixel as a time-varying feature of the data, this means that even these small images constitute 4624-dimensional data.

The inherent dimensionality of the data is obviously much lower. The images represent a small number of objects moving in constrained ways. Linguistically-motivated descriptions of speech production (e.g., articulatory phonology) posit between 8 and 16 articulatory variables [1]. Even biomechanical models of the vocal tract contain only about 40 to 50 variables [6]. Given this, one would expect a dimensionality reduction of at least two orders of magnitude.

Performing this reduction is a complex task, of course, and various methods of doing it can be devised. The method one chooses can depend on (a) theoretical concerns about the linguistic variables of interest, and how naturally they correspond to the dimensions of the reduction (i.e., interpretability of the reduction), and (b) practical requirements of processing the data, such as robustness, precision, efficiency, automaticity.

The most intuitive way to process data of this type is by finding the air-tissue boundaries represented in the image. Such boundary-tracing can be done in the spatial domain [7] or in the frequency domain [4]. This kind of data reduction has a straight-forward interpretation, since it traces an outline of physically extant structures. Highly useful measurements, such as the vocal tract area function, can be determined from these boundaries. However, in seeking linguistically-relevant variables, boundary-finding is simply a first step. Variables like constrictions must be defined and similarly extracted from the boundaries. This poses a number challenges, both theoretically and practically. By that same token, these methods tend to lack robustness, since boundaries may appear poorly defined due to noise and smearing. Higher degrees of automaticity can bring high computational demand, as well.

This paper discusses an alternative kind of data reduction, which directly uses the pixel intensity values of the vocal tract images. It takes advantage of how those intensities vary and co-vary over time to segment the images into regions in which pixels are highly correlated with each other. Such methods hold promise for robustly capturing the details of articulatory movement directly from the images. They are simple, and they avoid the need for computing intermediate segmentations, such as air-tissue boundaries.

Over time, the intensity values of an individual pixel reflect changes in tissue density at a particular location in the image plane. If we assume a lack of head motion, then pixel locations reflect tissue changes in the midsagittal plane, as well. At the same time, localized changes in tissue density along the vocal tract is the definition of a vocal tract constriction, except that constrictions are on a larger scale than a single pixel. Thus, vocal tract constrictions should manifest in the images as localized regions, in which the pixels are highly correlated over time.

We will show that intensity variations in these regions correspond to constriction degrees. To that end, we demonstrate how these correlated regions can be extracted with minimal manual guidance, and we will show an algorithm for extracting them automatically from the data. This kind of data analysis is a simple, efficient and robust method for extracting certain variables relevant to speech production. We hope to discuss and illustrate the kinds of applications for which it is appropriate, as well as to mention the kinds of applications for which it is inappropriate.

Section 2 provides a brief description of how the rtMRI data were acquired. In section 3, we describe the methods for manual and automatic selection of correlated regions. Section 4 provides an illustration of the proposed methods by application to specific data sets. This is followed by a brief discussion of the

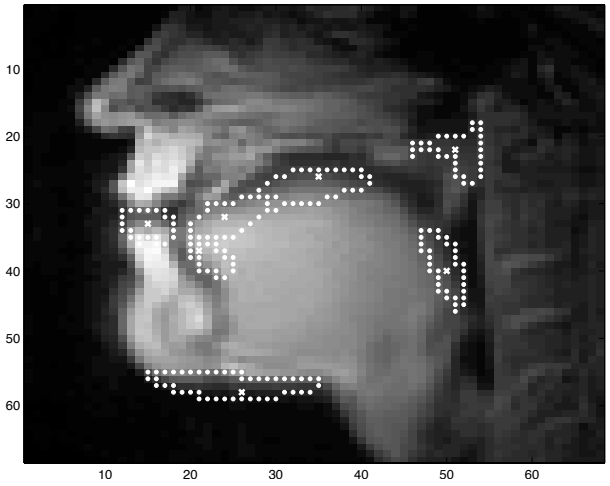

Figure 1: *Mean image with correlated regions overlaid (manually selected). White crosses mark the selected pixels, which are used to build the regions. Regions are outlined by white dots.*

relative merits of our method (section 5), and finally some concluding remarks in section 6.

2. Data Acquisition

The subjects spoke while laying in the scanner. Meanwhile, we collected real-time MR image sequences of the vocal tract from a midsagittal view. Audio recordings of the subjects vocalizations were also acquired, but were not utilized in this study. Our acquisition setup utilizes a GE Signa 1.5T scanner with a 13 interleaf spiral gradient echo pulse sequence. The MR pulse repetition time was TR = 6.5 ms. The slice thickness was approximately 3mm. A sliding window reconstruction was employed, at a rate of 22.44 frames per second. The field of view was adjusted for the subject's head size, so that images covered an area of 18.4 cm by 18.4 cm at a resolution of 68 $\times$ 68 pixels. Further details regarding the recording/imaging setup can be found in [2] and [3], and a sample video can be found at http://sail.usc.edu/span.

Subjects were asked to speak a variety of read stimuli, as well as to speak spontaneously about various topics. Indeed, we have collected a fairly large corpus of rtMRI data for various studies. The analyses presented here were designed to help tackle those data, and so no data were collected specifically for this study, per se. The relevant stimuli for the examples presented in this paper will be described below.

3. Extracting Correlated Regions

3.1. Correlation Images

Given a rtMR image M, which is $r \times c$ pixels in size, we can reshape that image into a column vector X. The vector X will be of length $p = rc$, and the pixels which were located at (s, d) in M are at location $i = c(d-1) + s$ in X.

For a sequence of n images, we can then compile them into a single $n \times p$ matrix

$$\mathbf{X} = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix} \tag{1}$$

Thus, we consider that each image is a single data point in a p-dimensional space, and we have n data points. From this, a $p \times p$ correlation matrix $\mathbf{C}$ can be easily calculated. Each element of the correlation matrix, $\mathbf{C}_{ij}$, shall represent the Pearson product-moment correlation coefficient between the ith and jth column of $\mathbf{X}$.

When a pixel of interest is selected, we can look at the column vector $\mathbf{C}_{*,\mathbf{i}}$ to find how it is correlated with every other pixel across time. Moreover, that column vector can be reshaped into an image of size $r \times c$. This is a "correlation image," where the pixel intensities are the correlation coefficients. Correlation images can, in general, be rich with information about how different parts of the vocal tract are coordinated. Even disparate areas of the vocal tract can sometimes show moderate correlation, due to particular coordinative relationships. For instance, labial constrictions are often correlated with jaw movement since the jaw is sometimes recruited to bring the lips together. Here, we focused on extracting localized regions of high correlation, because of their correspondence with vocal tract constrictions.

3.2. Manual Selection

A single pixel can be used in defining a correlated region. Selecting a pixel (i.e., a location in the image plane) determines a correlation image. If the selected pixel is located in a region of the image plane where constrictions are consistently observed, then it will be highly correlated with its neighboring pixels. The correlation image will show high intensities throughout this region. In practice, the specific pixel chosen in this region changes the correlation image very little, since all the pixels in it are highly correlated with each other. Choosing a good pixel to define a region can be done manually, by inspecting the vocal tract anatomy. To that end, it is usually helpful to view the mean image (see figure 1), and choose a location near one known constriction location (e.g., the lips, palate, alveolar ridge, etc.).

Given a pixel i, we define its correlated region to be all the pixels in its correlation image which are (a) correlated above a threshold τ, and (b) connected to i via other pixels which are correlated above τ. These regions may be found by standard region growing algorithms [7]. An example set of regions can be found in figure 1. The regions correspond to constrictions near the lips, alveolar ridge, sublingual cavity, palate, velum, pharynx and jaw.

3.3. Automatic Selection

Extracting the correlated regions can also be done automatically. Here, we suggest one possible way. Areas of interest will show a high degree of localized correlation. Thus, we can obtain an estimate of the correlation "density" at each pixel location. To that end, we proceed to build a correlation density map, D, which has the same dimensions ($p \times 1$) as a correlation image. Each element of this density map can be calculated by averaging the correlations between the corresponding pixel, i, and its 4-neighbors from the image space. More specifically,

$$D_i = mean(\mathbf{C}_{i,N_4(i)}) \tag{2}$$

where the function $N_4(b)$ denotes the pixel 4-neighbors,

$$N_4(x) = \{x+1, x-1, x+c, x-c\} \tag{3}$$

We note that this calculation is similar to kernel density estimation. As such, using the 4-neighbors in this instance represents a box-cart kernel with a width of 2. While any other kernel

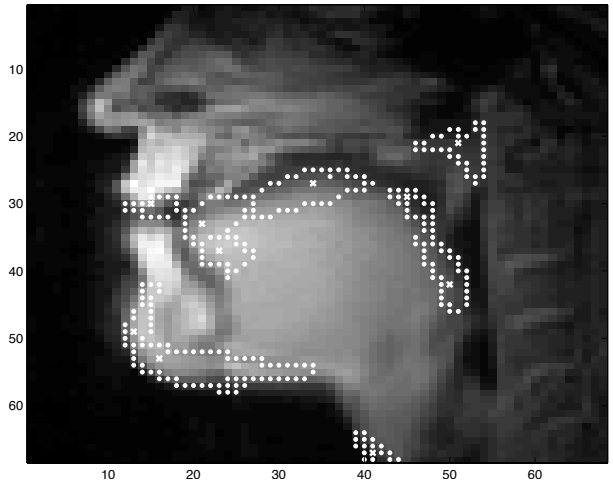

Figure 2: *Mean image with correlated regions overlaid (automatically selected). White crosses mark the selected pixels, which are used to build the regions. Regions are outlined by white dots.*

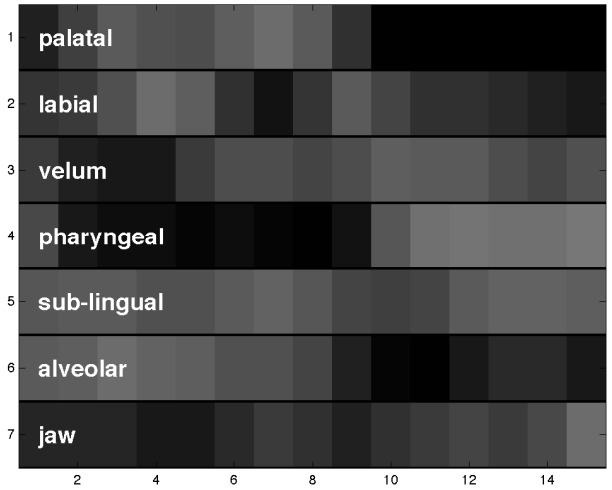

Figure 3: *Extracted articulatory time series for the utterance /pipəl/. Time progresses from left to right, and lighter shades represent more tissue (i.e., more constriction) in the region indicated.*

could be chosen, we used this kernel in order to get a fine-grain density map.

Choosing regions can be done iteratively by selecting the pixel corresponding to the highest value in D, defining the region as per section 3.2, and then downweighting the elements of D based on the correlation image of that pixel. This downweighting ensures that we avoid pixels from the same region on subsequent iterations. Specifically, the correlation image will be,

$$\mathbf{c} = \mathbf{C}_{*,arg_i max(D)} \tag{4}$$

We proceed to downweight D using the square of $\mathbf{c}$ (i.e., the coefficient of determination). Prior to this, we zero all the negative elements in $\mathbf{c}$, because our regions are defined in terms of high positive correlation. Each downweighted value of D will become an element of $\hat{D}$:

$$\hat{D}_i = D_i(1 - f(\mathbf{c_i})^{\mathbf{2}}) \tag{5}$$

$$f(x) = \begin{cases} x : x > 0 \\ 0 : x < 0 \end{cases} \tag{6}$$

We then we set $D = \hat{D}$, and repeat this procedure until we obtain the desired number of regions. Examples of the automatically extracted regions can be seen in figure 2. As with the regions in figure 1, the regions correspond to constrictions near the lips, alveolar ridge, sublingual cavity, palate, velum, pharynx and jaw. In addition, the automatic selection method found regions relevant to the chin (jaw protrusion), velar constrictions of the tongue, and movement of the laryngeal prominence.

3.4. From Regions to Time Functions

Once a region has been defined, a constriction time function can be obtained by simply averaging the intensity values of pixels within the region for each frame. This amounts to estimating the average tissue density in the segmented region. This kind of averaging also results in substantial noise reduction, as compared to the signal from an individual pixel. Larger regions will, in general, show even less noise.

It should be noted that, since the relationship of intensity values and tissue density is unknown, these measurements are inherently relative, not absolute. Also note that averaging represents the simplest way to combine pixel intensities from a region. Various other methods could be implemented, from weighted means to complex transform representations.

4. Examples

For the purposes of illustration, we applied our methods to two sequences of rtMRI images. The regions were selected manually for both examples.

4.1. Example 1: Spontaneous Speech Data

We applied our methods to a sequence of rtMRI images taken from 14.45 seconds of spontaneous speech. The subject was a male American English speaker. He was asked to speak about his experiences as a graduate student. Figure 3 displays a set of 7 time functions, corresponding to constrictions in the palatal, labial, pharyngeal, alveolar regions, as well as near the velum, jaw and sub-lingual cavity. The time functions were normalized between 0 and 1. The utterance was "people," spoken by the subject in the context "I've met a lot of interesting people, I've learned ... ,".

By inspection, one can see that the expected linguistic events are preserved by this method. The timing of these events is also clearly visible. For instance, the pair of bilabial closures are clearly visible. A palatal constriction can be seen between the two bilabial closures, corresponding to the high front vowel in the first syllable. The dynamic development of these events can be observed, as well. Note that the palatal constriction begins during the first labial constriction and ends abruptly, prior to the second. Distinct characteristics of this speaker are also visible, such as the strong pharyngeal constriction during the [l], and the relative lack of jaw movement over the entire utterance.

4.2. Example 2: Speech Errors Data

These methods are exceptionally well-suited to studies focused on the timing of articulatory events. Consider studies that scru-

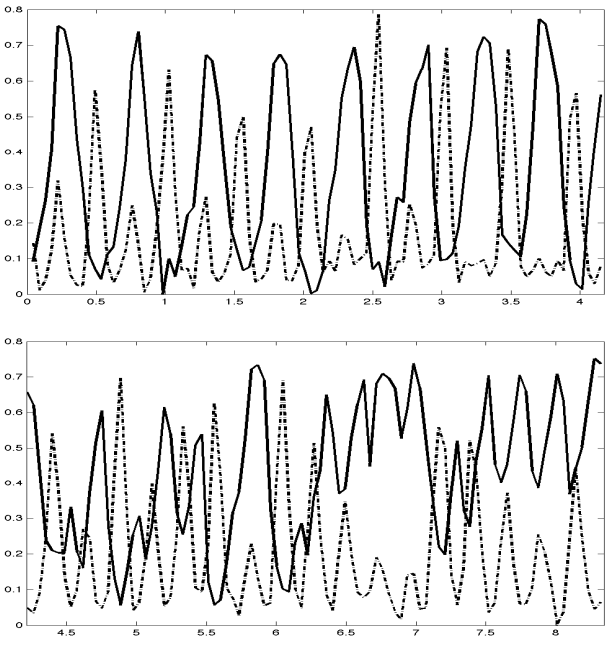

Figure 4: Time series from a single recording session. The plots represent constrictions in the velar region (solid line) and the alveolar (dotted line). Time continues from the top graph to the bottom.

tinize speech errors. To elicit intrusion errors, subjects will often repeat similar syllables in alternating fashion [8]. These alternations produce rhythmic constrictions, which show up readily with the methods described here.

Figure 4 shows an 8.5 second segment of speech from a male, native Tamil speaker. The subject was asked to alternately repeat the utterances /kɑp/ and /tɑp/, producing alternating constrictions in the alveolar and velar regions. Using our methods, these are seen as highly periodic time signals with equal frequency, but which are 180 degrees out of phase (see top of figure 4). The signals remain periodic in this way until errors begin to occur (see bottom of figure 4). In the errors are particularly catastrophic, and take a variety of forms. At first, non-periodic behavior can be observed, followed by synchronization of phase, and then dominance of only velar constrictions, with alveolar constrictions being minimized.

5. Discussion

The methods presented here represent a highly data-driven way of processing image sequence data. User input is not required, but it can help to guide the region finding, if desired. The applicability of these methods to a particular data set rests on several assumptions. We assume no – or, possibly slight – head motion. This assures that the image plane coincides with the reference frame of interest (e.g., the midsagittal plane). The data should reflect repeated actions in the reference frame, at a scale larger than a single pixel. Obviously, the number of frames in the sequence should be large enough to mitigate correlations that arise by chance alone.

Correlated regions take temporal variations into account from the outset, which makes them very appropriate for rtMRI data. By averaging pixel intensities across all the pixels in a region, the resulting time functions become reasonably robust to noise. The methods are extremely easy to implement, and the measures are easily and quickly extracted. There is no need to optimize complex cost functions, so most of the computation comes from calculating the correlations. If an image sequence contains images with n pixels, then n^2 correlations must be calculated. After that, the regions are defined once, for the entire sequence of images.

One possible concern is interpretation of the specific values obtained with methods of this kind. The regions correspond to constriction degrees, but they do not necessarily imply the scale of that constriction. If the relative degree of constriction or the dynamic information is what is needed, there is still a strong motivation for using methods like those presented here. In the end, we will not argue that this kind of data analysis is a panacea. It is, in some ways, a less intuitive way to approach data of this kind. However, it is a simple, efficient and robust method for extracting certain variables that are relevant to speech production.

6. Conclusion

Realtime MR image sequences contain rich information about movement and coordination within the vocal tract. The most intuitive way to deal with high-dimensional data of this kind is by reducing it to air-tissue boundaries. We have described and illustrated an alternative way to perform the data reduction. We segment the images into highly correlated regions in the image plane. Intensity variations in the regions reflect the time-course of consistent, repeated changes in tissue density along the midsagittal plane, corresponding to vocal tract constrictions.

We plan to apply these methods to a variety of rtMRI data that we have collected. In the near future, our plan is to utilize the correlated regions for direct modeling of feature time series, similar to the example presented here (see section 4.2).

7. Acknowledgements

Work described in this paper was supported by NIH Grant DC007124, as well as a graduate fellowship from the Annenberg Foundation.

8. References

[1] Browman, C. and Goldstein, L., "Towards an articulatory phonology", Phonology Yearbook, 3:219, 1986.

[2] Narayanan, S., Nayak, K., Lee, S., Sethy, A. and Byrd, D., "An approach to real-time magnetic resonance imaging for speech production", JASA, 109:2446, 2004.

[3] Bresch, E., Nielsen, J., Nayak, K. and Narayanan, S., "Synchronized and noise-robust audio recordings during realtime MRI scans", JASA, 120:1791, 2006.

[4] Bresch, E. and Narayanan, S., "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images", IEEE Trans. Med. Imaging, 28(3):323, 2009.

[5] Ladefoged, P., "A Course in Phonetics: Fifth Edition", Thomson Wadsworth, 2006.

[6] Vogt, F., Lloyd, J., Buchaillard, S., Perrier, P., Chabanas, M., Payan, Y. and Fels, S., "An efficient biomechanical tongue model for speech research", in Proceedings of ISSP, 2006.

[7] Forsyth, D. and Ponce, J., "Computer Vision: A Modern Approach", Prentice Hall, 2002.

[8] Goldstein, L., Pouplier, M., Chen, L., Saltzman, E. and Byrd, D., "Dynamic action units slip in speech production errors", Cognition, 103(3):386, 2007.