FACTOR ANALYSIS OF VOCAL-TRACT OUTLINES DERIVED FROM REAL-TIME MAGNETIC RESONANCE IMAGING DATA

Asterios Toutios, Shrikanth S. Narayanan

University of Southern California, Los Angeles, USA
{toutios, shri}@sipi.usc.edu

ABSTRACT

A factor analysis of vocal-tract outlines derived automatically from real-time magnetic resonance image (rtMRI) sequences has been performed. The analysis results in a compact representation of vocal-tract shapes, where every utterance is represented by a small set of trajectories corresponding to weights in linear combinations of linguistically interpretable vocal-tract deformations. Vocal-tract shapes can be reconstructed with good accuracy from these trajectories. The work uses information from a significantly larger number of speech frames compared to previous attempts in articulatory modeling. The proposed method is illustrated through a case study of rtMRI data corresponding to 250 sentences spoken by a single speaker and underscores the promise of the methodology for phonological analysis and articulatory synthesis.

Keywords: real-time MRI, articulatory modeling, guided PCA

1. INTRODUCTION

Real-time magnetic resonance imaging (rtMRI) has recently allowed the acquisition of dynamic data on the entire midsagittal slice of the vocal tract at a high enough spatio-temporal resolution, in ways and volumes that were hitherto not possible, especially given the limitations of X-ray imaging due to safety and ethical concerns [12, 1]. The recent release of the USC-TIMIT database has made a large corpus of such data freely available to the research community [13].

Information from rtMRI comes in the form of videos that may not be readily amenable to large-scale analysis before the application of some feature extraction method. Often, such methods target specific regions of the vocal tract, depending on the phenomenon under study [7, 16, 5, 15]. We propose here a method to extract a set of features covering the entire midsagittal slice that comprises two steps: first, the automatic derivation of the outlines of articulators in each rtMRI video frame, based on a previously published segmentation algorithm [2]; second, the conversion of the dynamics of these outlines into a series of phonetically meaningful trajectories, from which the entire midsagittal slice can be readily reconstructed with sufficient accuracy. Such a compact representation of the entire tract may prove beneficial for the analysis of vowels, where overall shaping is more relevant than particular constrictions, and in the context of articulatory synthesis [10, 17], since it may better capture the natural deformations of an exemplar speaker's vocal tract than a general articulatory model [11, 4, 19, 9].

We drew inspiration from Maeda's work on articulatory modeling [8, 9], in developing the analysis of this paper. That said, there are a few key differences between our modeling and Maeda's. First, while Maeda used measurements on an articulatory grid, we target directly the coordinates of points on the articulatory outlines. This is enabled by the segmentation method used, which displaces a fixed number of points on the outlines of 15 vocal-tract structures. Second, we introduce an analysis of the velum deformation, as well as a factor for the shaping of the arytenoid cartilage which accounts for voicing. The shaping of the arytenoid is visible in rtMRI which images a thin (5mm) slice cutting through the speaker's head, as opposed to X-ray which images a projection from the side (see Fig. 1). Third, while Maeda used data from 10 short sentences (about 1000 frames – at 50 fps), we use data from 250 larger sentences (18,130 frames – at 23.18 fps).

2. REAL-TIME MRI DATA AND SEGMENTATION

We used midsagittal rtMRI data from the F1 speaker of the USC-TIMIT database [13], a 23-year old female speaker born in New York. The speech material recorded with rtMRI corresponds to the 460 sentences of the MOCHA-TIMIT dataset [18]. These data were subjected to an updated version of an automatic segmentation algorithm previously published [2]. The segmentation method considers the outlines of 15 anatomical features comprising three connected regions of tissue (see Fig. 2). For every

Figure 1: Two examples of rtMRI segmentation. Left: /z/ from the utterance "This was easy for us". Right: /s/ from"Is this seesaw safe". Note that the orientation of the head is different in the two images. The left image belongs to the subset that is used in the analysis. Note also the different shape of the visible structure (arrow) at the arytenoid cartilage area that can distinguish voiced vs. unvoiced.

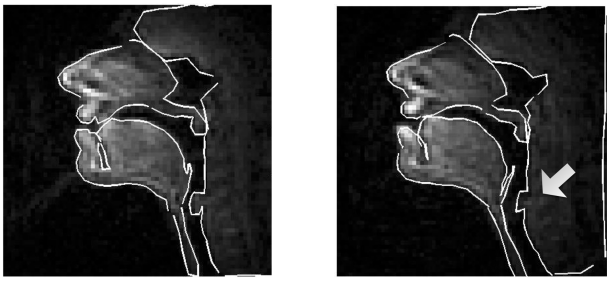
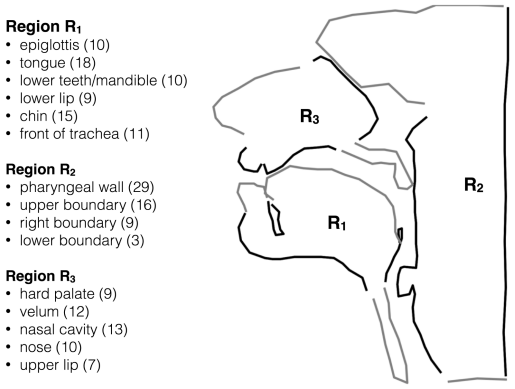


image in a video sequence, the method incrementally deforms an initial set of anatomical feature outlines (a template), by displacing a fixed number of points on each outline, until a fit to the observed image data is achieved. We did not perform a formal

Figure 2: Anatomical features used in the segmentation algorithm. Each feature is sampled at the number of points shown in parentheses.



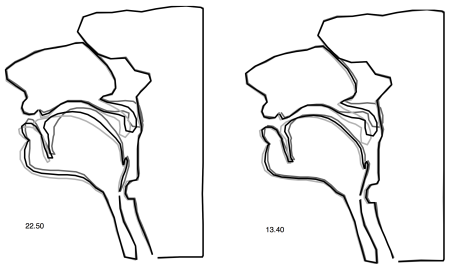
evaluation of the goodness-of-fit of these articulatory contours. That would require the manual segmentation of at least a subset of rtMRI video frames, which is in itself a process amenable to errors. We relied on visual inspection of the contours, which was satisfactory.

A technical problem during the recording of these data resulted in the first 250 utterances having a different orientation of the speaker's head compared to the rest of the dataset (see Fig. 1). To avoid any problems that might arise from co-registering these two subsets, the analysis presented in this paper is based only on the first 250 utterances.

3. FACTOR ANALYSIS METHODOLOGY

After segmentation, there are 184 points describing the articulatory contours corresponding to each rtMRI frame. Direct application of Principal Component Analysis (PCA [6]) on the *xy* coordinates of these points leads to a set of factors that optimally explain the variance of these coordinates across the dataset. Fig. 3 visualizes the two most significant principal components which appear to correspond to combined deformations of several articulators (e.g. jaw, tongue, lips, *and* velum). Our aim however

Figure 3: Visualization of the two most significant factors derived by PCA on the entire contours (± 2 standard deviations). Numbers at lower left corner indicate the percentage of variance explained by each factor.



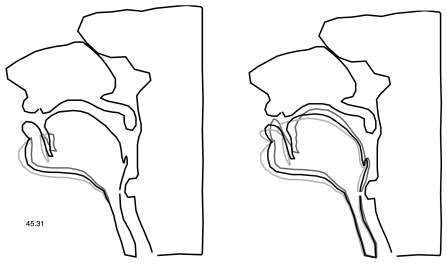
is that each factor has a specific articulatory correlate, i.e. is linguistically interpretable. In what follows, we are describing a methodology that aims to achieve this goal, starting from an analysis of the jaw movement.

3.1. Jaw

We apply PCA focused on the jaw and mandible outlines shown in Fig. 2. That is, we apply PCA on the data after setting the values of the *xy* coordinates of the points on all the other structures at their mean values across the dataset. The left panel of Fig. 4 visualizes the most significant principal component (a vector of length 368), which explains about 45% of the jaw and mandible variance, and corresponds well to the linguistically important *jaw opening gesture*. Further principal components (not shown here) appear unrelated to the opening gesture.

Adopting a jaw-based approach to articulatory modeling, it is crucial to extract the component of the deformation of the tongue and lower lip that is a direct consequence of the jaw opening gesture. To achieve this, we first set the *xy* coordinates of all outlines not it Region 1 of Fig. 2 to their mean values, and calculate the covariance matrix, let R, of this modified dataset. If t_1 is the first jaw principal component previously derived, define $v = t_1' R t_1$,

Figure 4: Left: First jaw principal component. Right: The component of deformation of the entire lower vocal-tract region, which is a direct consequence of the first jaw principal component.

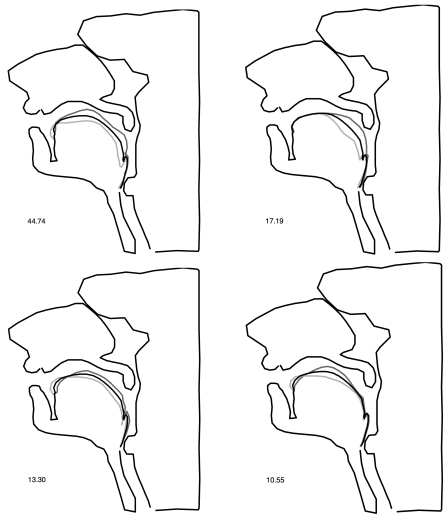


$h_1 = t_1/\sqrt{v}$, and $f_1' = Rh_1'$ [14, 3]. A visualization of f_1, which can be regarded as a factor of the data replacing the first jaw principal component, is shown in the left panel of Fig. 4.

3.2. Tongue

We first subtract the contribution of factor f_1 from the data, by $a_{new} = a(1 - f_1 f_1^{\dagger})$ where $\dagger$ denotes the pseudo-inverse. PCA focusing on the tongue outline is applied on the modified dataset. We also find the contribution of the these components to the epiglottis, in the same way we found the contribution of the jaw opening to the tongue. The first four resulting factors are shown in Fig. 5. It is interesting to note

Figure 5: First four tongue components (after removal of the contribution of the jaw)

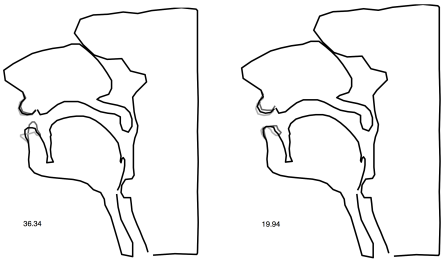


that the first three tongue factors associate well with the tongue factors of Maeda's model: *tongue dorsum position, tongue dorsum shape, tongue tip* (the latter encoding also some deformation at the tongue root, as in Maeda's model).

3.3. Lips

We subtract the contribution of jaw factor f_1 from the data, and apply PCA focused on the lips. The first two factors thus derived are shown in Fig. 6. The first of these factors can be associated to *lip opening* and the second to *lip protrusion*.

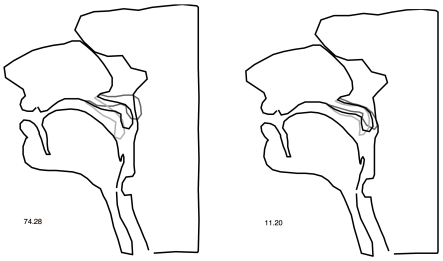
Figure 6: First two lip components (after removal of the contribution of the jaw)



3.4. Velum

We apply PCA focused on the velum. The first component captures more the 75% of the variance and is associated to *velum lowering*. Subsequent components appear to capture linguistically non-important deformations (Fig. 7).

Figure 7: First two velum components



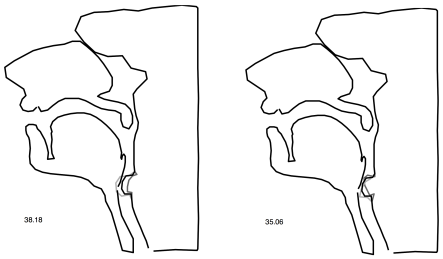
3.5. Larynx

We apply PCA focused on a small number of points that outline the arytenoid cartillages (Fig. 8). The first component captures can be associated to *voicing*, given that when that the more the cartillages approach, the more prominent they are on the midsagittal MRI slice. The second component associates well to *larynx height*.

3.6. Hard palate

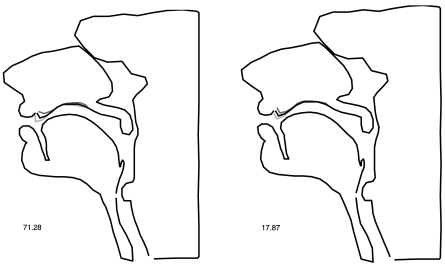
An inherent problem with rtMRI and the automatic segmentation method is that it fails to properly differentiate between the hard palate and the tongue tip.

Figure 8: First two larynx components



This results to observing a deformation of the hard palate outline. Applying PCA on the hard palate outlines gives the factors shown in Fig. 9.

Figure 9: Hard palate factors (as a result of inherent problems of rtMRI and segmentation).



3.7. Compact representation and reconstruction

Articulatory outlines of any frame in the dataset can be approximated by a weighted sum of the previously derived factors: $a = \sum_i w_i f_i$. While the factors (vectors) f_i are constant across the dataset, the set of weights (scalars) w_i is dynamic and can be used as a compact representation of the vocal-tract outlines. Fig. 10 shows trajectories of the weights associated with four of the presented factors along an utterance from the dataset. The weight of the jaw opening presents a rhythmic pattern; the tongue dorsum position weight is largely correlating with fronting; the velum weight has peaks at nasals; and the voicing weight presents peaks for unvoiced segments.

Given the factors, vocal-tract outlines can be reconstructed from the weights. Fig. 11 shows a reconstruction using the factors of Figs. 4-9 with the exception of the left panel of Fig. 4 and the right panel of Fig. 7. The approximation is in all satisfactory, retaining salient features of the vocal-tract shaping. We note that we had to include the hard palate factors (which of course are not linguistically or physiologically relevant) otherwise we would not get right the constriction for alveolar stops and fricatives. Correcting this problem is something we intent to look further into.

Figure 10: Trajectories of weights associated with factors: (from top to bottom) jaw opening; tongue dorsum position; velum; and voicing

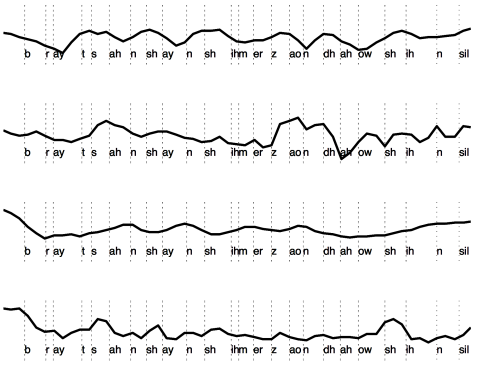
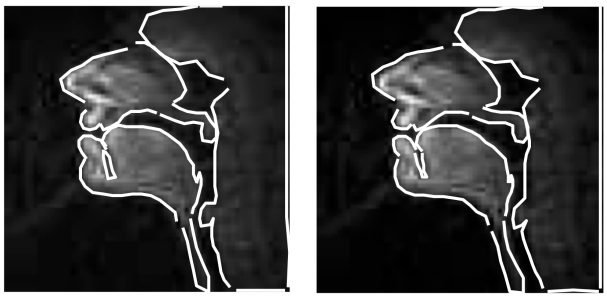


Figure 11: Left: Result of segmentation (same as in Fig. 1, left. Right: reconstruction from 12 factors.



4. CONCLUSION

We have presented a method to derive compact representations of vocal-tract outlines automatically tracked on rtMRI sequences, as a linear combination of (mostly) linguistically interpretable vocal-tract deformations. The analysis was illustrated using data drawn from 250 sentences spoken by an American English speaker. The results underscore the promise of the method. There are however limitations that need to be further addressed. The study has been speaker- and recording session- specific, and thus does not lead to a general-purpose articulatory model. Further generalizations will be required, which is among our future work plans, alongside exploring the utility of our current analysis for phonological analysis and articulatory synthesis.

5. ACKNOWLEDGMENT

Research presented in this paper was supported by NIH grant R01DC007124.

6. REFERENCES

[1] Bresch, E., Kim, Y.-C., Nayak, K., Byrd, D., Narayanan, S. May 2008. Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging [Exploratory DSP]. *IEEE Signal Processing Magazine* 25(3), 123–132.

[2] Bresch, E., Narayanan, S. Mar 2009. Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. *IEEE Transactions on Medical Imaging* 28(3), 323–338.

[3] Cai, J., Laprie, Y., Busset, J., Hirsch, F. 2009. Articulatory modeling based on semi-polar coordinates and guided PCA technique. *Interspeech* Brighton,UK. 56–59.

[4] Harshman, R., Ladefoged, P., Goldstein, L. 1977. Factor analysis of tongue shapes. *The Journal of the Acoustical Society of America* 62(3), 693–707.

[5] Hsieh, F.-Y., Goldstein, L., Byrd, D., Narayanan, S. S. Aug. 2013. Truncation of pharyngeal gesture in english diphthong [ai]. *Interspeech* Lyon, France.

[6] Jolliffe, I. 1986. *Principal component analysis*. New York: Springer-Verlag.

[7] Lammert, A., Proctor, M., Narayanan, S. Sept. 2010. Data-driven analysis of realtime vocal tract MRI using correlated image regions. *Interspeech* Makuhari, Japan. 1572–1575.

[8] Maeda, S. 1979. Un modèle articulatoire de la langue avec des composantes lineaires. *10ème Journées d'Etude sur la Parole* Grenoble, France. 152 – 162.

[9] Maeda, S. 1990. Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In: Hardcastle, W. J., Marchal, A., (eds), *Speech production and speech modelling*. Amsterdam: Kluwer Academic Publisher 131–149.

[10] Maeda, S. 1996. Phonemes as concatenable units: VCV synthesis using a vocal-tract synthesizer. In: Simpson, A., Pätzold, M., (eds), *Sound Patterns of Connected Speech: Description, Models and Explanation*. 145–164.

[11] Mermelstein, P. 1973. Articulatory model for the study of speech production. *The Journal of the Acoustical Society of America* 53(4), 1070–1082.

[12] Narayanan, S., Nayak, K., Lee, S., Sethy, A., Byrd, D. 2004. An approach to real-time magnetic resonance imaging for speech production. *The Journal of the Acoustical Society of America* 115, 1771.

[13] Narayanan, S., Toutios, A., Ramanarayanan, V., Lammert, A., Kim, J., Lee, S., Nayak, K., Kim, Y.-C., Zhu, Y., Goldstein, L., others, 2014. Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). *The Journal of the Acoustical Society of America* 136(3), 1307–1311.

[14] Overall, J. E. 1962. Orthogonal factors and uncorrelated factor scores. *Psychological Reports* 10(3), 651–662.

[15] Parrell, B., Narayanan, S. S. 2014. Interaction between general prosodic factors and language-specific articulatory patterns underlies divergent outcomes of coronal stop reduction. *International Seminar on Speech Production (ISSP)* Cologne, Germany.

[16] Proctor, M. I., Katsamanis, N., Goldstein, L., Hagedorn, C., Lammert, A., Narayanan, S. 27-31 Aug. 2011. Direct estimation of articulatory dynamics from real-time Magnetic Resonance Image sequences. *Interspeech* Florence, Italy. 281–284.

[17] Toutios, A., Narayanan, S. Aug. 2013. Articulatory synthesis of French connected speech from EMA data. *Interspeech* Lyon, France. 2738–2742.

[18] Wrench, A., Hardcastle, W. May 2000. A multichannel articulatory speech database and its application for automatic speech recognition. *Proc. 5th Seminar on Speech Production, Kloster Seeon, Bavaria* 305–308.

[19] Zheng, Y., Hasegawa-Johnson, M., Pizza, S. 2003. Analysis of the three-dimensional tongue shape using a three-index factor analysis model. *The Journal of the Acoustical Society of America* 113(1), 478–486.