Validating rt-MRI based articulatory representations via articulatory recognition

Athanasios Katsamanis, Erik Bresch, Vikram Ramanarayanan, Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory
University of Southern California, Los Angeles

http://sail.usc.edu

Abstract

The large corpus of real time magnetic resonance image sequences of the vocal tract during speech production that was recently acquired and can be referred to as MRI-TIMIT, provides us with a unique platform for systematically studying articulatory dynamics. Compared to previously collected articulatory datasets, e.g., using articulography or X-rays, MRI-TIMIT is a rich source of information for the entire vocal tract and not only for certain articulatory landmarks and further has the potential to continue increasing in size covering a large variety of speakers and speaking styles. In this work, we investigate an articulatory representation based on full vocal tract shapes. We employ an articulatory recognition framework in MRI-TIMIT to analyze its merits and drawbacks. We argue that articulatory recognition can serve as a general validation tool for real-time MRI based articulatory representations.

Index Terms: vocal tract shape, articulation, real-time MRI, articulatory recognition

1. Introduction

The selection of an appropriate real data-driven vocal tract representation is dependent both on the needs and specificities of a particular speech production study but also, to a significant degree, on the flexibility allowed by the amount and variability of the available data. For example, Mermelstein in [1] presented a parametric, geometric vocal tract shape model that was based on roughly 300 mid-saggittal x-ray tracings. Not exactly data-driven, the model was manually fitted to the real data and then used for articulatory speech synthesis. Maeda in [2] approached the problem in a statistical manner, i.e., using guided principal component analysis, and developed a full vocal tract shape model based on the manually annotated, X-ray based vocal tract tracings for 10 French utterances by a single speaker. This model has been used for articulatory synthesis, speech inversion and other articulatory studies. Despite their success, admittedly both these models are constrained by the fact that they are based on only very few articulations.

Such constraints are partially removed when we consider representations based on Electromagnetic Articulography (EMA) speech production data, which have been made available in significant amounts more recently. However, these representations can only account for specific critical points on the articulatory system and do not describe the entire vocal tract shape. So, they may be appropriate for studying articulatory kinematics or speech inversion but they can not easily be used for speech synthesis, for example. In our work, we present yet another vocal tract representation, this time based on a significant number of real-time Magnetic Resonance Image sequences of mid-sagittal vocal tract profiles. The proposed representation accounts for the entire vocal tract shape, is derived in a fully automatic manner and is validated via articulatory recognition.

More specifically, we develop a statistical deformable vocal tract shape model based on automatically extracted vocal tract outlines from the real-time upper airway MRI recordings of 460 phonetically-balanced TIMIT utterances by a single male speaker. A discriminative shape representation is also derived using linear discriminant analysis. Compared to other statistically derived vocal tract shape models the presented model is uniquely combining the two following properties:

1. It is not based on any vocal tract specific reference grid, as, e.g., the semi-polar grid used in [2] .
2. It is estimated in a fully automatic manner and is not based on manually annotated vocal tract contours as is the case in [3, 2].

Further, it is the first time a vocal tract shape model is built on so many and so variant articulations. A similar model based on MRI images in [3] was only built on 25 MRI frames, i.e., corresponding to a dataset of portuguese phonemes acquired via sustained articulations, while in our case we have roughly 30000 vocal tract shapes corresponding to continuous articulations.

Typically, validation of shape models is achieved by comparison of reconstructed shapes with manually annotated outlines of the original vocal tract shapes. In our case, we only have automatically extracted vocal tract contours available, since, given the nature and amount of the original data, manual annotation is not possible. Comparison of the reconstructed shapes with possibly not fully accurate vocal tract contours may be misleading however. So, we validate the resulting shape representations by articulatory recognition experiments. The hypothesis is that the proposed vocal tract representation can successfully capture the critical shape properties of the vocal tract during speech production and discriminate between different types of articulation. We apply hidden Markov models for the articulatory recognition and are able to recognize sequences of different articulation types in continuous speech with almost 50% accuracy.

2. Articulatory representations

2.1. Articulatory corpus

For our experiments we use the real-time MRI recordings of 460 TIMIT utterances by one male, native American English speaker. These data were collected as part of a larger parallel audio-articulatory database, referred to as MRI-TIMIT [4]. Vocal tract images have been reconstructed at a framerate of

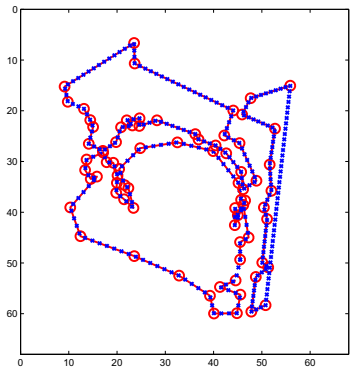

Figure 1: Original vocal tract polyline desribed by the red circles and equidistantly resampled vocal tract contour described by blue 'x' markers. Resampling was performed to cope with the "point correspondence" problem. The points are in pixel coordinates. The size of the original image is 68×68 pixels (200mm×200mm field of view)

23.18Hz. In our current work we focus on only one of the four speakers in the database.

2.2. Vocal tract outlines

We extracted the vocal tract outlines from the images using the segmentation algorithm described in [5]. Each outline is formed by three separate contours delineating the lower, upper and back parts of the vocal tract respectively. The lower part includes the tongue, lower lip and epiglottis, the upper part includes the upper lip, hard palate and velum while the back part corresponds to the pharyngeal wall and the larynx. Each contour is essentially represented by a dense ordered set of M points $C = \langle x_1, x_2, \ldots x_M \rangle$ where consecutive points are connected by a line segment. By concatenating the Cartesian coordinates of these points we can describe each contour by a $2M$-dimensional shape vector $\mathbf{s}$.

Our goal is to build a point distribution model [6] to describe the vocal tract deformations during speech production. This model will indirectly capture how the shape changes based on the joint statistical properties of the vocal tract points. To reliably estimate these statistics from a representative set of contours, i.e., a training set of shape vectors, it is important to establish a correspondence between the points describing different contours. This is the so-called correspondence problem [7, Chap. 4]. This is often solved by meticulous manual annotation of specific landmark points on the training shapes. However, identifying landmarks on certain parts of the vocal tract, e.g., the tongue, that is homogeneous and its length may vary in a non-uniform manner, is not straightforward. We tried to minimize the related implications by densely and equidistantly sampling each contour with a fixed number of points, always starting from a point which can be relatively robustly identified on the images, e.g., a point on the tip of the chin. At a final preprocessing stage, we compensate for possible rigid head movement of the speaker during the recordings by properly aligning the shape vectors via an iterative application of the iterative closest point method [8] implementing a modified version of the alignment algorithm in [6].

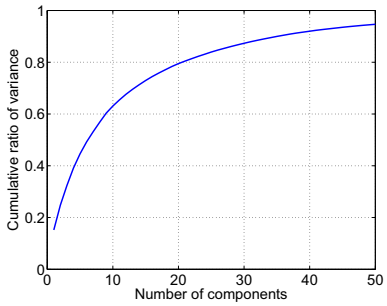

Figure 3: Cumulative ratio of the total variance explained by the principal components of the shape model.

2.3. Statistical deformable vocal tract model

Given the aligned set of N vocal tract shape vectors we can build a point distribution model using principal component analysis [6]. We estimate the mean vocal tract shape:

$$\bar{\mathbf{s}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{s_i} \tag{1}$$

and the corresponding covariance matrix S:

$$S = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{s} - \bar{\mathbf{s}})(\mathbf{s} - \bar{\mathbf{s}})^T. \tag{2}$$

The major modes of shape variation are described by the eigenvectors of S [6]. The first four vocal tract eigen-shapes are shown in Fig.2 . The mean shape is in solid line. The dashed lines describe the effect of the eigenmodes when they are weighted by three different values. So, each shape can be approximated as:

$$\mathbf{s} \approx \bar{\mathbf{s}} + P\mathbf{b} \tag{3}$$

where P is a matrix formed by a reduced number of eigenvectors/principal components and $\mathbf{b}$ is a weight vector. Our analysis showed that 90% of the variance can be explained by 35 principal components. The cumulative ratio of the total variance explained with the inclusion of increasing number of components is given in Fig 3. So, each shape can be quite accurately represented by the corresponding 35-dimensional weight vector. This vector will be:

$$\mathbf{b_{PCA}} = P^T(\mathbf{s} - \bar{\mathbf{s}}). \tag{4}$$

An alternative representation can be obtained using linear discriminant analysis. We aim at a vocal tract representation of reduced dimensionality that would preserve as much of the discriminative information among classes of shapes as possible. In our current exploratory study, our classification is based on phonetic classes. Essentially, we assume that we have a separate class of vocal tract shapes for each phoneme. For a specific vocal tract shape, the resulting description will be:

$$\mathbf{b_{LDA}} = W^T\mathbf{s} \tag{5}$$

The optimal projection matrix W, assuming unimodal Gaussian distributions for each class can be found via linear discriminant analysis [9, Sec. 3.8.3], by maximizing the ratio of between-class to within-class scatter.

The derived statistical models and the corresponding vocal tract contours are available online at `http://sipi.usc.edu/~nkatsam`. and will also become available as part of the MRI-TIMIT database.

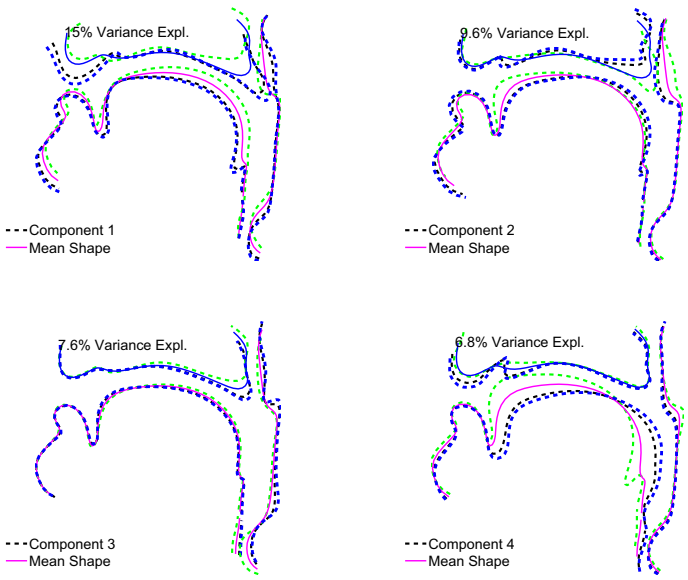

Figure 2: First four eigenshapes as estimated by principal component analysis on approximately 30000 aligned vocal tract shapes. The three dashed lines for each component correspond to the component weights being equal to plus one, minus one and plus two standard deviations of the respective eigenvalues.

3. Articulatory recognition using HMMs

By articulatory recognition we refer to the identification of a sequence of phonemes based on the corresponding sequence of vocal tract shapes. The term "phonemes" in this context refers to the corresponding articulatory configurations. For example, a vocal tract shape with a labial closure and a velopharyngeal opening would be recognized as the articulatory configuration for the phoneme /m/. Such experiments have been reported for articulatory data (EMA) in the past mainly as part of speech recognition studies that exploit articulatory information, e.g., in [10].

For the recognition experiments, the vocal tract shape parameters for the sequences of magnetic resonance images are based on the models described above and are resampled at 100Hz. We built 3-state ergodic HMM models, trying to have one for each type of articulation that we expect to be discriminable by the corresponding vocal tract shaping. The choice of the number of states was made empirically, mainly motivated by the conventional approach followed in speech recognition. Roughly, we train one articulation model for each phoneme, with the exception of pairs of phonemes like /p/ and /b/, /t/ and /d/ that only differ in terms of voicing and we wouldn't expect them to be discriminable in terms of vocal tract shape differences. Each of these pairs of phonemes, six in total, were grouped together. So, our recognition system is based on 30 models, including one for silence.

Our models are continuous density hidden Markov models with state observation probability distributions described by Gaussian mixture models (GMMs). Each Gaussian component is described by the corresponding mean vector and a diagonal covariance matrix. The number of Gaussian components was fixed to four. This increased model complexity per state is expected to allow capturing coarticulation effects. It would be too strict to just assume unimodal shape probability distributions per state for each phoneme. A diagonal matrix is chosen based on the fact that our shape features are expected to be largely decorrelated. The models are trained using HTK [11]. We followed a standard three-step training process.

1. Initialization of the models using the timed phonetic transcriptions of the training sequences. Timed transcriptions were originally generated by Viterbi-based forced alignment of the synchronously recorded acoustic data with the corresponding text. The parameters of each phoneme articulation model are initialized by the global statistics of the shape parameters over all the corresponding segments. Each model's parameters are updated separately in an iterative fashion. In each iteration, the Viterbi algorithm provides an alignment of the model with the corresponding segments. Each state's parameters are updated by pooling together the data assigned to that state and estimating their statistics.
2. Baum-Welch reestimation of each model's parameters separately. Again, the parameters are updated iteratively. This time, the Baum-Welch algorithm is applied in each iteration instead of the Viterbi.
3. Final composite model training. For each training sequence a composite model is created by concatenating all the models found in the corresponding transcription. The standard Baum-Welch algorithm is then applied to update all the parameters. The process is repeated till convergence, i.e., till the change of model likelihood given all the data between two consecutive iterations is negligible.

It should be noted that although the timed phonetic transcriptions are used for the initialization of the articulatory models, in the final training step this constraint is removed. So, the articulatory models are allowed to be asynchronously aligned with the corresponding acoustics, as it would be intuitively expected.

A model representing the articulation of /k/ for example may be activated earlier than the point when the corresponding sound becomes apparent in acoustics.

We randomly split our data in 95%-5% training and testing parts and we repeated our experiments 10 times. We report the average recognition accuracy over these repetitions. The shape models were also developed on the same training set each time. We experimented with both the Principal Component and the Linear Discriminant Analyses-based features and the corresponding results are ginen in Table 1. We also give results obtained with image intensity-based Discrete Cosine Transform features (100 features were used). DCT results serve as a more generic baseline that does not use any explicit shape information. The accuracy achieved with the LDA-based features is similar to that of the PCA-based features. Overall, the articulation recognition accuracy shows that our fully automatically derived shape models and the temporal modeling imposed on top of them via the hidden Markov models can capture significant part of the articulation variation during continuous speech production.

Table 1: *Recognition accuracy (%) of articulatory types using shape-model based parameters. The model has been built either using Principal Component Analysis of a Point Distribution Model or using Linear Discriminant Analysis. The result based on intensity-based DCT features is also given.*

	Articulatory recognition accuracy
intensity DCT	31.2%
shape PCA	47.3%
shape LDA	49.4%

4. Discussion and future work

We presented a statistical deformable vocal tract shape model developed in a fully automatic manner from the real-time magnetic resonance recordings of the vocal tract of a speaker while uttering 460 utterances. These data are part of larger audio-articulatory database that is currently been acquired and is known as MRI-TIMIT [4]. The described shape model is essentially trained on more than 30000 automatically derived vocal tract outlines. It is the first time that a vocal tract shape model is built on so many and so variant articulations.

Compared to previously presented vocal tract models that are based on similar approaches but were trained on different and much smaller datasets our current model appears to 1) include a much higher number of components and 2) to explain variations that are not necessarily relevant to speech production. For example, the first component appears to describe motion of the upper lips and at the same time shrinkage of the velum, which could be argued to be unrealistic. These observations most probably can be attributed to contour tracking errors. As mentioned earlier, the proposed approach is fully automatic and is based on the vocal tract shapes that are extracted by an automatic contour tracking algorithm. By a simplistic outlier detection mechanism we managed to identify a large proportion of erroneous vocal tract contours and we excluded them from our analysis. However, we haven't yet been able to verify and ensure the correctness of all the shapes that are finally included. We are currently working on a semi-automatic approach towards this direction. More specifically, our goal is to incrementally train the vocal tract shape model so that only acceptable training shapes are used and possibly noisy observations are cleaned up using the already trained model.

We validate the derived vocal tract representation via articulatory recognition experiments based on hidden Markov models. Using the proposed scheme we are able to achieve articulatory recognition accuracy close to 50%. This effort is part of our ongoing work to establish rich and flexible articulatory representations that can be automatically derived or adapted to real-time articulatory data. We would like to devise a representation that would allow us to robustly exploit these data in a computational manner and in a way that could potentially also inform and be informed by related speech production theories.

In this direction, we plan to extend our shape model to multiple speakers. For this purpose we are also investigating alternative representations that would probably be less speaker-dependent. The articulatory recognition framework we have described will be our validation platform. Further, we investigate ways to use the statistical shape model in the vocal tract contour tracking as well, in a properly adapted Bayesian framework, as for example the one presented in [12] for tongue tracking in ultrasound images.

5. Acknowledgements

Research supported by NIH Grant R01 DC007124-01.

6. References

[1] P. Mermelstein, "Articulatory model for the study of speech production," *J. Acoust. Soc. Am.*, vol. 53, no. 53, pp. 1070–1082, 1973.

[2] S. Maeda, "Un modèle articulatoire de la langue avec des composantes lineaires," *10ème Journées d'Etude sur la Parole*, pp. 1–9, 1979.

[3] M. J. M. Vasconcelos, S. M. R. Ventura, D. R. S. Freitas, and J. M. R. S. Tavares, "Using statistical deformable models to reconstruct vocal tract shape from magnetic resonance images," *Journal of Engineering in Medicine*, vol. 224, 2010.

[4] S. Narayanan, E. Bresch, P. K. Ghosh, L. Goldstein, A. Katsamanis, Y.-C. Kim, A. Lammert, M. I. Proctor, V. Ramanarayanan, and Y. Zhu, "A multimodal real-time MRI articulatory corpus for speech research," in *Proc. Int'l Conf. on Speech Communication and Technology*, Florence, Italy, Aug 2011.

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

[6] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models – their training and application," *Computer Vision and Image Understanding*, vol. 61, pp. 38–59, 1995.

[7] R. H. Davies, "Learning shape: Optimal models for analysing natural variability," Ph.D. dissertation, Division of Imaging Science and Biomedical Engineering, University of Manchester, 2002.

[8] P. Besl and N. McKay, "A method for registration of 3-d shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, pp. 239–256, 1992.

[9] R. O. Duda and P. E. Hart, *Pattern classification and scene analysis*. New York: Wiley, 1973.

[10] J. Frankel and S. King, "ASR-articulatory speech recognition," in *Proc. European Conf. on Speech Communication and Technology*, 2001.

[11] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. Entropic Cambridge Research Laboratory, Cambridge, England., 2002.

[12] A. Roussos, A. Katsamanis, and P. Maragos, "Tongue tracking in ultrasound images with active appearance models," in *Proc. IEEE Int'l Conf. on Image Processing*, 2009.