Statistical multi-stream modeling of real-time MRI articulatory speech data

Erik Bresch[1], *Athanasios Katsamanis*[1], *Louis Goldstein*[2], *Shrikanth Narayanan*[1]

[1]Department of Electrical Engineering, University of Southern California, USA
[2]Department of Linguistics, University of Southern California, USA
bresch@usc.edu, nkatsam@sipi.usc.edu, louisgol@usc.edu, shri@sipi.usc.edu

Abstract

This paper investigates different statistical modeling frameworks for articulatory speech data obtained using real-time (RT) magnetic resonance imaging (MRI). To quantitatively capture the spatio-temporal shaping process of the human vocal tract during speech production a multi-dimensional stream of direct image features is extracted automatically from the MRI recordings. The features are closely related, though not identical, to the tract variables commonly defined in the articulatory phonology theory. The modeling of the shaping process aims at decomposing the articulatory data streams into primitives by segmentation. A variety of approaches are investigated for carrying out the segmentation task including vector quantizers, Gaussian Mixture Models, Hidden Markov Models, and a coupled Hidden Markov Model. We evaluate the performance of the different segmentation schemes qualitatively with the help of a well understood data set which was used in an earlier study of inter-articulatory timing phenomena of American English nasal sounds.

Index Terms: speech production, articulatory modeling, real-time magnetic resonance imaging

1. Introduction

The recent technological advances in real-time (RT) magnetic resonance imaging (MRI) allow the speech researcher access to large quantities of rich articulatory data of running speech [1]. As opposed to previously available speech production data from electro-magnetometry (EMA), which provides spatially sparse point tracking, and ultrasound, which is confined to capturing the tongue shape, RT-MRI captures the air-tissue boundaries along the entire vocal tract from the glottis to the lips. RT-MRI data hence appear to be a good basis for studying the vocal tract shaping process in a holistic way, i.e., they allow the investigation of individual articulators while simultaneously taking into account the effects of inter-articulatory coupling. However, the identification of shaping primitives from RT-MRI data (or from any other articulatory data) is not trivial, due to the data's high dimensionality, the complexity of the deformation space of the vocal tract, and the inter and intra subject variability in articulation.

In this article we will address the problem of identifying articulatory gestures from streams of RT-MRI image sequences. According to the theory of articulatory phonology [2], a gesture is a goal directed action of constriction forming by a vocal tract articulator. This process is modeled using the response of a second order linear system to a constriction target input step function. An articulator may be used to execute a sequence of consecutive gestures which leads to *temporal* gestural overlap. The gestures are quantified using tract variables, and it is important to realize that the mechanical coupling, due to the anatomical constraints, may produce *spatially* correlated measurement noise across different tract variables. So, the recognition of gestures from articulatory data must undo or at least take into account this spatio-temporal mixing.

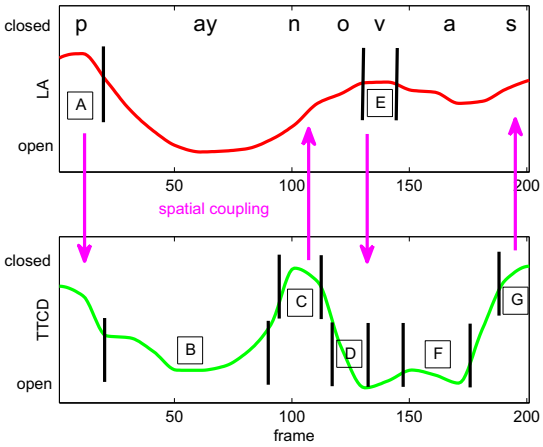
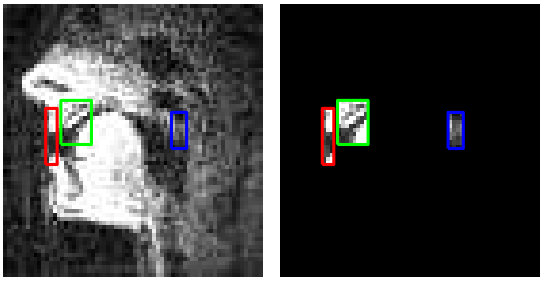
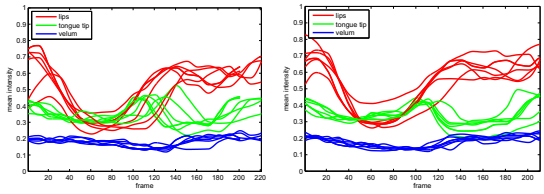


Figure 1: *Lip aperture (LA) and tongue tip constriction degree (TTCD) time series for the utterance /pay nova s/ as derived from RT-MRI data (details given below).*

For example, we can consider the lip aperture (LA) and tongue tip constriction degree (TTCD) time series for the token /pay nova s/ as segmented from the carrier *"Type pay nova slowly."* (Fig. 1). Here, we have manually marked the critical constriction forming processes. The segment labeled "A" of the LA trace corresponds to the bilabial closure for the formation of the /p/. Note that TTCD is also relatively constricted during this interval, due to articulatory coupling: the jaw contributes to lip closure, and brings the tongue tip towards the palate as a side-consequence. The purple arrow pointing down is meant to represent the direction of this coupling effect – from a phonologically controlled gesture to a passive coupling consequence. This is followed by the TTCD segment "B" for the formation of the diphthong /ay/. The diphthong is made using tongue body gestures which couple into the TTCD measurements. The subsequent tongue tip closure at the alveolar ridge in segment "C" is critical for the formation of the nasal, and we can identify a subtle effect on the LA trace due to the spatial coupling of the lips and the tongue via the jaw. As the tongue body is then used to produce the /o/ in segment "D", the lips move closer for the labiodental /v/ in segment "E," which again has an effect on the TTCD through spatial coupling. Finally, the production of the vowel /a/ ("F") with the tongue body is followed by a period of



(a) Sample image. (b) Masked image.



(c) Feature time series for 7 realizations of /pay nova s/. (d) Feature time series for 7 realizations of /pain over s/.

Figure 2: *Sample image and direct image feature time series.*

narrow TTCD for the sibilant /s/ in segment "G."

Previously, a variety of heuristical approaches have been pursued to model the shaping kinematics, such as the decomposition of individual EMA-traces into strokes [3], though with mixed results. In this paper we explore the use of a dynamical Bayesian network [4] to model the articulatory multi-stream data in a machine learning framework. Hereby, the joint modeling of different regions of the vocal tract is critical for coping with inter-articulator coupling, and the statistical processing will ensure a degree of robustness against intra subject variability.

This article is organized as follows. In Section 2 we will propose a simple yet robust way to obtain shaping information from the midsagittal MR images which aims at providing measurements closely related to the tract variables. Given a low-order parametric representation of the vocal tract shape we will, in Section 3, attempt a segmentation of image feature time series with vector quantizers (VQ), Gaussian Mixture Models (GMM), uncoupled Hidden Markov Models (HMM), and a coupled HMM (CHMM). The CHMM network is versatile, and it is particularly attractive since it is capable of handling asynchrony between data streams [5]. Finally, in Sections 4 and 5 we will discuss the results and draw conclusions.

2. Data preparation and parameterization

The data corpus for this case study consisted of two types of utterances produced by a female native American English speaker, namely *"Type pay nova slowly."* and *"Type pain over slowly."* The recordings were made using the scan protocol described in [6]. Seven realizations of each type, extracted from the carrier phrase, yielded the tokens /pay nova s/ and /pain over s/ used for our analysis. The starting frame was identified by the bilabial closure for /p/, and the end frame was chosen based on the narrow tongue tip constriction at the alveolar ridge for /s/. The token duration was on the order of 1 second, and our MRI frame rate is approximately 22 frames per second. No timing normalization was carried out. A sample midsagittal MR image is shown in Fig. 2(a).

The robust automatic extraction of the vocal tract shape in terms of its air tissue boundaries from the midsagittal MRI is not straightforward and is still considered to be an active domain of research [7, 8]. A versatile yet compact shape representation and parameterization, which would be beneficial for speech modeling purposes such as recognition, inversion, or synthesis, is not easy to obtain. Previous work in this domain includes the principal components based shape model used in [9, 10] or the constriction based vocal tract model implied by articulatory phonology [2]. Deriving such constriction measurements from image sequences can increase uncertainty of the data used for modeling. Given the complex geometry of the vocal tract using a region based description of constriction events, rather than pinpointing a specific constriction location or its degree, appears to be a more robust choice. We focus on such a parameterization of the image sequences directly so as to capture the constriction events implicitly but robustly.

In this study we confine ourselves to investigating the articulatory processes involving the lips, the tongue tip, and the velum, and we select correspondingly in each image rectangular regions of interest as shown in Fig. 2(a) (shown as red, green, and blue box, respectively). The location of the regions is considered fixed, although this choice can also be dictated in a data driven way based on the region statistics such as the local image intensity correlation properties [11]. We can assume negligible head motion occurred during the experiment since the subjects head was well immobilized.

We then mask out the rest of MR image as shown in Fig. 2(b) and compute for each frame the average image intensity in each of the regions. The time series of these image intensity features are shown in Fig. 2(c) and 2(d) for all 7 realizations of /pay nova/ and /pain over/, respectively, and they have been ten-fold interpolated. The time series have a straightforward intuitive interpretation, since constriction forming events correspond to increasing the average image intensity because tissue moves into the particular region of interest. Conversely, a constriction release leads to a drop of average intensity over time since tissue moves out of the affected region. Hence the features closely resemble the constriction degree tract variables defined in articulatory phonology. Further, this representation can inherently capture the linguistically meaningful events in the presence of production variability, including due to inter-speaker morphological differences.

The two utterances were chosen because they differ minimally in the syllable position of the nasal, which is in coda position for /pain over/ and in onset position for /pay nova/. Previous studies [12, 13] have shown that systematic relative timing differences exist for the tongue tip closure gesture and the velum opening gesture during the nasal production depending on its position in the utterance, and we will hence use this data set as a test case for our modeling framework.

3. Data modeling

Due to the limited number of training realizations in the data set considered, we will confine ourselves to detecting the gross shaping phenomena, i.e., the closure events "A," "C," "E," and "G" in Fig. 1. A simplified gestural transcription is shown in Fig. 3, where "OP" means open, "CL" means closed, and "X" means irrelevant state. The challenge for the segmentation algorithm will be to not give a false "CL" detection result at the very end of the LA trace (solid red), since that maximum is due to coupling from the TTCD (solid green). Equivalently, we would like no false "CL" alarm in the beginning of the TTCD trace,

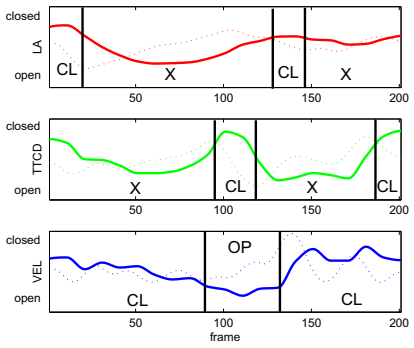


Figure 3: *Lip aperture (LA), tongue tip constriction degree (TTCD), and velum aperture (VEL) for the utterance /pay nova s/ with gestural transcription. Solid line - feature time series, dashed line - first derivative.*

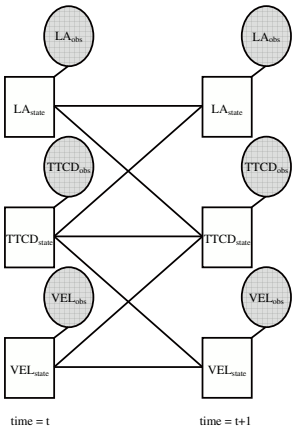


Figure 4: *3-chain CHMM layout (squares - hidden discrete nodes, shaded circles - continuous observations).*

since that maximum is due to spatial coupling with the LA trace. Both requirements are difficult to achieve robustly by a simple quantization of the time series. As noted earlier, the image sequences of vocal tract contours reflect a fairly complex dynamic geometry, and simple rule-driven ways of robustly identifying minimum constriction location/degree are difficult to implement, even with region based parameterization.

Generally, the time series data are quite noisy, and their first derivatives even more so (dotted lines in Fig. 3), especially for the velum (blue curves) due to the low image contrast in the pharyngeal region. So, rules such as through simple thresholding to find inflection points often do not yield reliable results. Hence, statistically capturing the time series behavior directly appears as a reasonable approach to pursue.

In the following we will augment the feature streams by their first derivatives, and attempt the modeling using VQ, GMM, HMM, and CHMM systems. These methods were chosen for a variety of reasons. The VQ is the most straightforward way to implement a simple instantaneous, i.e., time independent, thresholding mechanism for the individual 2-dimensional augmented feature data streams. The quantization levels can be found robustly using the well known k-means procedure, which, given the number of quantization levels, is otherwise parameter free. A manual transcription of all 14 data tokens as shown in Fig. 3 was produced, and it was used for the training of all of the methods. For the VQ, two centers were allocated corresponding to the two class labels. It should be noted that a VQ could also be implemented on the joint feature streams of all measurements, though we chose to keep the streams separate to allow "fair" comparisons of the VQ, GMM, and HMM methods.

The GMM can be considered a more sophisticated statistical way to achieve an instantaneous quantization, and it affords soft output values. However, in our case we implemented subsequent hard clipping and thereby lose this advantage, but we included the GMM approach since it is often used in practice, and it can provide initialization parameters for the subsequent HMM systems. Just as the HMM and CHMM, the GMM is trained using the expectation maximization (EM) algorithm, which for all applications in this study was employed with a convergence threshold of 10^{-5}. The GMMs were implemented using the MATLAB Netlab toolbox which is a component of the BNT toolbox [4]. The models were initialized using k-means, and they had a full covariance matrix.

The HMM is a step up from the GMM in terms of modeling power and system complexity. It can be thought of as a time-dependent quantizer, and this method was chosen to address the temporal gestural overlap within a tract variable feature time series. Three individual HMMs were used for the LA, TTCD, and VEL data. The HMMs were implemented using the MATLAB HMM toolbox which is also included in the BNT package. The hidden nodes had two states corresponding to the two segmentation labels used for each tract variable. Using the transcriptions, we initialized the observation models as bi-variate Gaussians with full covariance matrices, as well as the state priors and the ergodic state transition model.

The CHMM is the most complex system that we tested for this study, and it allows spatio-temporal modeling of the combined time series data. The model had three chains corresponding to the LA, TTCD, and VEL features (see Fig. 4), and it was implemented using the MATLAB BNT toolbox. The three hidden nodes had two states each, and the observations were bi-variate Gaussians. The CHMM parameters were intialized using the previously trained uncoupled HMMs.

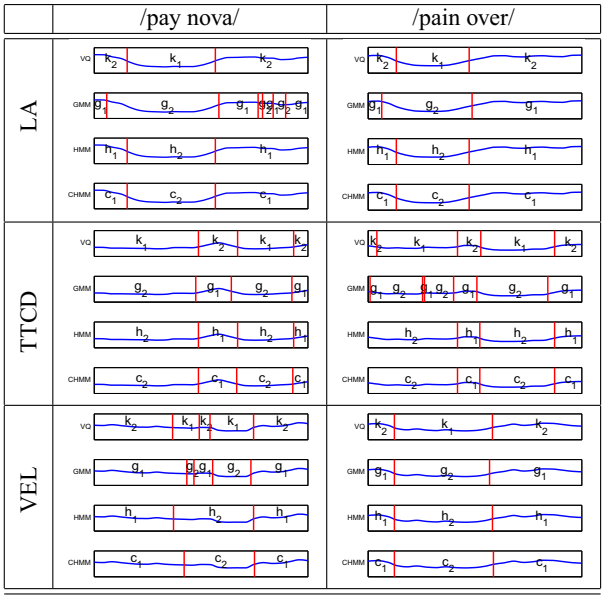
We carried out the segmentation of our 14 observed articulatory traces using leave-one-out cross-validation, and we present in Table 1 some typical results. The graphs in the left column correspond to a realization of /pay nova/ while the right column come from a realization of /pain over/. The top row shows the segmentation results for the LA trace, the middle row for TTCD, and the bottom row for VEL. The 4 plots in each row show the tract variable trace versus time (blue) and the segment boundaries (red vertical bars) as found by the VQ, GMM, HMM, and CHMM methods (top to bottom). The segments are labeled $k_{1,2}$ for VQ, $g_{1,2}$ for GMM, $h_{1,2}$ for HMM, and $c_{1,2}$ for CHMM.

4. Discussion

In general we observed that the VQ and the GMM methods produced more spurious transitions, as shown for LA and VEL segmentation for /pay nova/, and TTCD segmentation for /pain over/ in Table 1. Generally, the HMM and the CHMM produce comparable and more consistent results.

With respect to the HMM and CHMM method, we found that both of them consistently labeled the intial bilablial closure segment in the LA trace. They also found the onset of the labio-

Table 1: *Sample segmentation results.*



dental segment, but they repeatedly failed to identify its correct ending. However, both methods managed to avoid giving a false closure segment in the beginning of the TTCD trace. We found one realization of /pay nova/ for which the HMM, as opposed to all other methods, did not identify the VEL gesture at all.

Using the CHMM segmentation, we can now investigate the lag time difference between TTCD and VEL events for the formation of the nasal for the two types of tokens, i.e., we measure the time difference between the onset of the VEL opening (labeled c_2 in the bottom row, bottom graph of Table 1) and the onset of the TTCD closure (labeled c_1 in the center row, bottom graph). For the /pay nova/ tokens we obtain an average lag time of 96.8ms (σ=68ms), whereas for /pain over/ we obtain a lag of 279ms (σ=39.5ms). These results are encouraging since they are in accordance with previous findings [12, 13], and they seem to suggest that the proposed feature extraction procedure and the CHMM segmentation method appear to be robust and provide results that are consistent with our expectations.

In general we can suggest a number of ways to continue this study in order to improve the segmentation performance. On the one hand, one can certainly choose more complex models, e.g., higher-order mixtures for modeling the observations. And of course one can also scale up the entire procedure to include other image regions, leading to more chains in the CHMM. In any case, as more model parameters will have to be estimated a larger data corpus will be necessary. The possibility of collecting significant amounts of imaging data with RT-MRI holds promise in this regard.

5. Conclusions

We conclude from our study that the proposed method of image feature extraction has merit, and that the CHMM framework is a promising candidate for the discovery of articulatory primitives from RT-MRI data.

On a wider scope, this study indicates that if we combine (a) an explicit multistream transcription (gestures) with (b) appropriate techniques for extraction of articulatory time functions from RT-MRI data and with (c) the appropriate statistical models, we are well positioned to derive phonological information automatically from a rich set of articulatory data.

6. Acknowledgements

This work was supported by NIH Grant DC007124.

7. References

[1] E. Bresch, Y.-C. Kim, K. Nayak, D. Byrd, and S. Narayanan, "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging," *IEEE Signal Processing Magazine*, May 2008.

[2] C. Browman and L. Goldstein, "Towards an articulatory phonology," *Phonology Yearbook*, vol. 3, pp. 219–252, 1986.

[3] T. Kato, S. Lee, and S. Narayanan, "An analysis of articulatory-acoustic data based on articulatory strokes," *Proc. IEEE Int'l Conf. Acous., Speech, and Signal Processing*, pp. 4493–4496, 2009.

[4] K. Murphy, "The Bayes Net Toolbox for Matlab," *Computing Science and Statistics*, vol. 33, 2001.

[5] A. Nefian, L. L., X. Pi, X. Liu, and K. Murphy, "Dynamic bayesian networks for audio-visual speech recognition," *J. Applied Signal Processing*, 2002.

[6] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *Journal of the Acoustical Society of America*, vol. 115, pp. 1771–1776, 2004.

[7] E. Bresch and S. Narayanan, "Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images," vol. 28, pp. 323–338, 2009.

[8] J. Fontecave and F. Berthommier, "A semi-automatic method for extracting vocal tract movements from x-ray films," *Speech Communication*, vol. 51, pp. 97–115, 2008.

[9] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux, "Three-dimensional linear articulatory modeling of tongue, lips and face based on MR and video images," *Journal of Phonetics*, vol. 30, pp. 533–553, 2002.

[10] S. Maeda, "Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model," in *Speech Production and Speech Modeling*, W. Hardacastle and A. Marchal, Eds. Kluwer Academic Publishers, 1990.

[11] A. Lammert, M. Proctor, and S. Narayanan, "Data-driven analysis of realtime vocal tract MRI using correlated image regions," *Interspeech*, 2010, accepted.

[12] D. Byrd, S. Tobin, E. Bresch, and S. Narayanan, "Timing effects of syllable structure and stress on nasals: a real-time MRI examination," *Journal of Phonetics*, vol. 37, pp. 97–110, 2009.

[13] E. Bresch, L. Goldstein, and S. Narayanan, "An analysis-by-synthesis approach to modeling real-time mri articulatory data using the task dynamic application framework," *157th Meeting of the Acoustical Society of America*, May 2009.