INTERSPEECH 2011

Direct Estimation of Articulatory Kinematics from Real-time Magnetic Resonance Image Sequences

Michael Proctor[1,2], *Adam Lammert*[3], *Athanasios Katsamanis*[1],
Louis Goldstein[2], *Christina Hagedorn*[2], *Shrikanth Narayanan*[1,2]

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

mproctor@usc.edu

Abstract

A method of rapid, automatic extraction of consonantal articulatory trajectories from real-time magnetic resonance image sequences is described. Constriction location targets are estimated by identifying regions of maximally-dynamic correlated pixel activity along the palate, the alveolar ridge, and at the lips. Tissue movement into and out of the constriction location is estimated by calculating the change in mean pixel intensity in a circle located at the center of the region of interest. Closure and release gesture timings are estimated from landmarks in the velocity profile derived from the smoothed intensity function. We demonstrate the utility of the technique in the analysis of Italian intervocalic consonant production.

Index Terms: speech production, real-time MRI, consonant articulation, tongue shaping, articulatory phonology

1. Introduction

Real time magnetic resonance imaging (rtMRI) is an important emerging method for studying human speech production [1, 2]. However, analysis of speech data acquired using rtMRI presents several challenges. Unlike articulometry, the sensing modality is not designed to track the location of flesh points across time; as a result, analysis of midsagittal imaging data has typically relied on segmentation of tissue boundaries in the upper airway (e.g. [3, 4, 5]) – a time consuming approach which demands expert anatomical knowledge, and is prone to inconsistencies and experimenter biases.

Although *automatic* segmentation techniques have been developed for the analysis of real-time MRI [6, 7], additional processing is still required in order to identify articulatory events in the resulting sequences of tissue boundaries. No method currently exists for characterizing gestural dynamics from such data in a manner which would allow comparison with articulatory data obtained from other sensing modalities, including X-ray microbeam [8] and EMA [9, 10, 11, 12].

The goal of this study is to develop a robust method of identifying and automatically locating constriction events – typically corresponding to intervocalic consonant productions – in sequences of midsagittal MR images, with minimal supervision. Additionally, we propose a method of dynamically characterizing constriction formation and release which allows for quantification of hypothesized underlying gestural events, consistent with a task dynamic analysis of speech production [13]. The validity of this approach is assessed by comparing the derived constriction kinematics with lingual and labial trajectories tracked over the same image sequences.

2. Method

2.1. Image Acquisition

All data analyzed in this study were acquired using a rtMRI protocol developed specifically for the dynamic study of speech production [2]. Subjects' upper airways were imaged in the midsagittal plane with spatial resolution 68 x 68 pixels, field of view 200 x 200 mm, and a temporal reconstruction rate of 33.18 f.p.s. Subjects' heads were fixed throughout the scan, to allow for inter-frame comparison of image data. Pixel intensity was quantized into 8-bit values. Video sequences were reintegrated with de-noised audio simultaneously recorded inside the MRI scanner at 20 kHz [14]. The resulting video allows for dynamic visualization of midsagittal articulation, and acoustic analysis of the companion speech signal.

2.2. Automatic Constriction Location

We have developed a method for automatically identifying constriction location targets, by combining pixel-wise temporal dynamics with anatomical prior knowledge. For each speaker, the midsagittal trace of the passive articulators – extending from the velum to the upper lips – was defined (Fig. 1, solid line). A set of pixels below this line was automatically identified as a reasonable search space for each potential constriction region (e.g., Fig. 1: broken line defines dorsal search space). For each speech interval of interest, a cohort of maximally-active neighboring pixels was located within the specified search space, over the corresponding sequence of image frames.

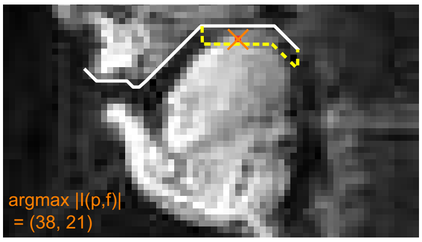

Figure 1: ***Automatic location of dorsal constriction target**: intervocalic stop [akːo] produced by adult male Italian speaker. Cross indicates center of maximum change in locally-correlated pixel intensity over surrounding 20-frame sequence.*

At each point p with coordinates (x, y) and intensity I_p in the search space P, the mean intensity $\bar{I}_p(\mathrm{f})$ of all pixels lying within a circle centered at p with radius r was calculated for

28 – 31 August 2011, Florence, Italy

each frame f (Eq. 1). The dynamic range of local intensity DR_p was calculated over the frame sequence f $\in [a..b]$ (Eq. 2). Pearson correlation coefficients $\rho(\bar{I}_p(\mathrm{f}), \bar{I}_q(\mathrm{f}))$ were computed between unit-radius intensity functions centered at p and each nearest neighbor q, and the degree of locally-correlated activity CoI_p was calculated as the mean correlation of changes in intensity at all pixels lying in a $2d \times 2d$ grid centered at p (Eq. 3).

The locus of articulatory activity for frame sequence $[a..b]$ was chosen to be the pixel c at which the weighted sum of DR_p and CoI_p was maximized over the search space P (Eq. 4), where α is a weighting factor used to preferentially select greater fluctuations in intensity ($\alpha \to 1$) or more highly correlated activity in a region ($\alpha \to 0$).

$$\bar{I}_p(\mathrm{f}) = \mathrm{mean}(I_q), \;\; \forall q : ||p-q|| < r \quad (1)$$

$$DR_p = \max_{\mathrm{f}\in[a..b]}(I_p(\mathrm{f})) - \min_{\mathrm{f}\in[a..b]}(I_p(f)) \quad (2)$$

$$CoI_p = \mathrm{mean}(\rho(\bar{I}_p, \bar{I}_q)), \;\; \forall q : ||p-q|| < d \quad (3)$$

$$\bar{I}_c = \max_{p\in P}(\alpha \cdot DR_p + (1-\alpha)\cdot CoI_p) \quad (4)$$

$$\hat{I}_p^j = b \cdot \hat{T}_p^j \quad (5)$$

2.3. Estimating Constriction Kinematics

Having located the region of greatest change in regional intensity for a sequence of images, the constriction degree in each frame can be estimated from the mean intensity of pixels in that region, since pixel intensities acquired from rtMRI reflect the density of soft tissue in a region of space. Local pixel intensity averaging (Eq. 1) therefore provides a good estimate of the kinematics of tongue movement into and out of alveolar, palatal, and dorsal regions (Fig. 2), or lip closure and opening, when the center pixel is chosen in labial regions.

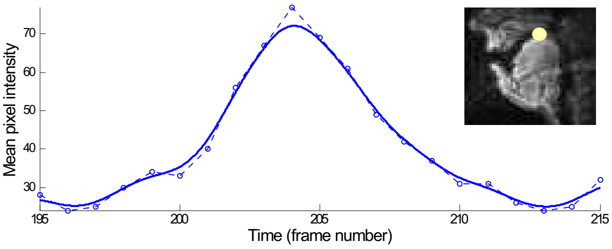

Figure 2: ***Change in intensity at constriction target**: intervocalic Italian stop [akːo]. Broken line: mean intensity of all pixels lying within circle (r = 3px) centered at point of maximum dorsal articulatory activity (39,20); Solid line: intensity function smoothed using locally-weighted linear regression.*

Because the effective sampling rate of rtMRI data is relatively low (< 40 frames/sec.), the resulting time series data can display discontinuities when derived from image sequences with low SNR. However, because the underlying articulator motion being estimated is characteristically smooth and intrinsically low frequency [15, 16], the intensity functions calculated using Equation 1 can be conditioned using a variety of techniques to remove noisy transients.

To reduce noise, and to facilitate estimation of constriction formation velocity, we fit an oversampled regression line $\hat{I}_p$ at samples $\hat{T}_p$, to the observed time series, $\bar{I}_p$ sampled at points T_p (Eq. 5), using locally-weighted linear regression [17]. As a weighting function, we use a gaussian kernel K having a standard deviation of h samples. The solid line in Fig. 2 illustrates the effect of smoothing the estimates of tongue body motion (broken line) into and out of a velar constriction during the production of an intervocalic Italian dorsal geminate stop. In this example, the kernel width parameter was h = 0.8 samples. Because samples lying more than $3h$ from the center of the gaussian kernel will receive weights near zero, this corresponds to a smoothing window width of approximately 90 msec for these data, which were sampled at 33.18 frames/second (sampling period = 30.1 msec).

2.4. Estimating Articulatory Activity

Because articulatory events appear to be more clearly identifiable from tissue velocity data (e.g., [10, 11, 12]), velocity metrics were derived from the estimates of constriction degree. First differences calculated directly from intensity functions $\bar{I}_p(\mathrm{f})$ were found to be insufficiently smooth to allow for the robust detection of gestural landmarks, so smoothed tissue velocity $d\hat{I}_p/dt$ was estimated from the regression coefficients b of the interpolated intensity functions $\hat{I}_p$ (Eq. 5).

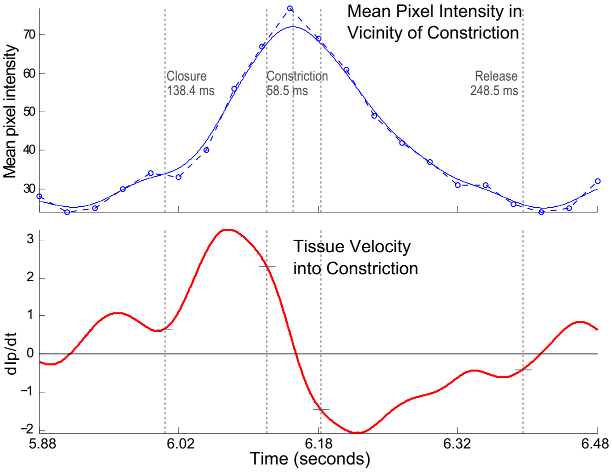

Figure 3: ***Tissue velocity estimated from intensity function**: intervocalic Italian stop [akːo]. Onset of closure gesture estimated at 20% of maximum positive tissue velocity (into constriction); Release gesture offset estimated at 20% of maximum negative velocity (away from constriction target).*

The timing of articulatory landmarks associated with the formation and release of consonant constrictions can be estimated using thresholds in the derived velocity functions (Table 1). For example, the beginning and end of the closure and release gestures hypothesized to underlie the production of an Italian intervocalic dorsal stop are indicated on the velocity and intensity functions shown in Fig. 3. The interval of maximal consonantal constriction may be estimated by locating thresholds around the negative-going zero-crossing of the velocity function (a 70% threshold constriction plateau is illustrated in Fig. 3).

3. Results

For the data analyzed so far, automatic constriction location was best achieved over a search space extending 2 to 3 pixels below the upper limit defined by the passive articulators (Fig. 1), calculating local intensity functions within a 4-pixel neighborhood (Eq. 1: $r = 1$), examining intensity correlations over an 8-pixel nearest-neighbors grid (Eq. 3: $d = 1$), and selecting for greater

Articulatory Event	Temporal Landmark
Closure gesture start	$d\hat{I}_p/dt > \beta \cdot \max(d\hat{I}_p/dt)$
Closure gesture end	$\max(\hat{I}_p)$
Release gesture start	$\max(\hat{I}_p)$
Release gesture end	$d\hat{I}_p/dt > \beta \cdot \min(d\hat{I}_p/dt)$

Table 1: *Temporal location of gestural landmarks in estimated tissue velocity function.*

changes in local intensity, rather than more highly correlated pixel activity over wider regions (Eq. 4: $\alpha < 0.2$).

Constriction kinematics were estimated by calculating mean regional intensity within a 3-pixel radius (Eq. 1: $r = 3$). Intensity function smoothing and tissue velocity $d\hat{I}_p/dt$ estimation was found to be most effective when $\hat{I}_p$ was interpolated using a gaussian kernel width parameter in the range $0.6 < h < 1.4$. Constriction closure gestural onset, and constriction release gestural offsets were estimated to occur within a 10% to 40% threshold of maximum and minimum tissue velocities, respectively (Table 1: $0.1 < \beta < 0.4$).

Maximal constriction (labial contact, or some part of the tongue in contact with the passive articulators) was typically observed while tissue velocity remained in the range $0.8 \min(d\hat{I}_p/dt) < d\hat{I}_p/dt < 0.8 \max(d\hat{I}_p/dt)$ – a result which is consistent with the hypothesis that the target constriction degree for stop consonants is negative [18, 13].

3.1. Validation

The accuracy with which articulatory activity is estimated using this technique can be assessed by comparing derived intensity functions with direct measurements of constriction degree taken from the same images. For the intervocalic dorsal stop sequences examined in §2.2–2.3, tissue boundaries defining the midsagittal airway in each frame were identified using a semi-automatic segmentation algorithm [7], and corrected manually where necessary.

For each frame, the shortest distance between the tongue and the velar constriction target c was calculated from the tissue boundaries, to produce a timecourse of constriction formation and release (Eq. 6). Constriction degree was also estimated by inverting the intensity function $\bar{I}_c(\text{f})$ calculated at the target, and scaling by the ratio of maximum intensity DR_c and maximum constriction degree CD_m (Eq. 7). Aperture-measured $CD_m(\text{f})$ and intensity-estimated $CD_e(\text{f})$ constriction functions for the intervocalic dorsal are compared in Fig. 4.

$$
\begin{aligned}
CD_m(\text{f}) &= \min(||c - \mathit{tongue}(\text{f})||) && (6)\\
CD_e(\text{f}) &= \max(CD_m)/DR_c \cdot (DR_c - \bar{I}_c(\text{f})) && (7)
\end{aligned}
$$

4. Applications

Because it allows for rapid, automatic characterization of constriction kinematics, the method described here is beginning to provide important insights into the temporal and spatial properties of consonant production.

4.1. Characterizing Coronal Place of Articulation

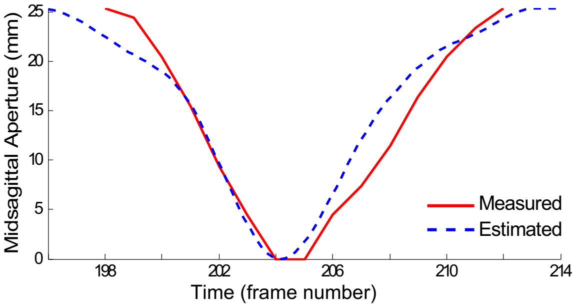

Figure 4: ***Change in dorsal constriction degree over time***: intervocalic Italian stop [ako]. (i) midsagittal aperture measured between velar target and closest point on tongue (solid line); (ii) estimated aperture calculated from mean pixel intensity in vicinity of velar target pixel (broken line).

When languages contrast multiple coronal consonants, these segments are typically characterized by fine articulatory differences [19, 20]. A proper understanding of coronals requires detailed knowledge about which part(s) of the tongue come into contact with different regions of the teeth, alveolar ridge, and palate; however, comprehensive information about place of articulation has been difficult to obtain from data acquired with sensing modalities other than palatography.

The utility of direct image analysis as a means of examining place of articulation is demonstrated in the data illustrated in Fig. 5. Place of articulation was calculated automatically, using the method described in §2.2, for 5 utterances of the intervocalic coronal stop [ada] and 10 utterances of the intervocalic lateral [ala], by the same speaker of Italian. For the coronal stop /d/, the mean center of articulatory activity occurs at p_d = (24,27); for the coronal lateral /l/, the mean center was located at p_l = (25.4,26.3), a place of articulation approximately 6 mm posterior to that of the stop. These data are consistent with previous characterizations of Italian /d/ as dental, and /l/ as alveolar [21]; further analysis of articulatory differences in Italian stops using this method is reported in [22].

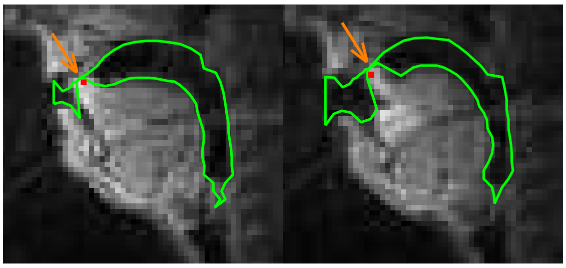

Figure 5: ***Characterization of Place of Articulation: Italian Coronals***. *Left: intervocalic stop [ada]; Right: intervocalic lateral [ala], Tissue boundaries indicate tongue position at point of maximal constriction. Red point indicates mean center of articulation estimated over multiple utterances of each consonant.*

4.2. Analyzing Consonant Duration

Because it provides information about the temporal evolution of constrictions in the vocal tract, the method described here can provide insights into timing differences in consonant production. Smoothed intensity functions of labial singleton and geminate stops are compared in Fig. 6. The total duration of the labial gesture (onset of closure to offset of release) for the geminate consonant is estimated to be 520 msec – approximately 30% longer than the singleton consonant (401 msec). Duration differences for Italian stops have been analyzed in detail using this method in [22].

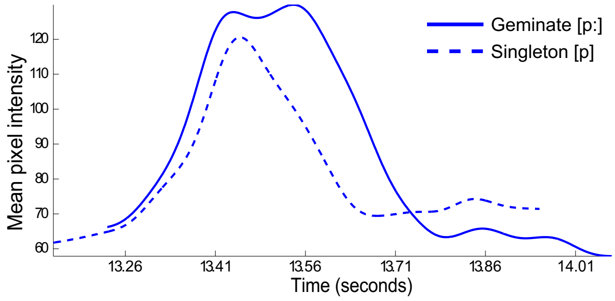

Figure 6: ***Estimation of Constriction Duration: Italian Labials***. *Solid line: intensity function for intervocalic geminate stop* [opːi]*; Broken line: intensity function for singleton stop* [opi]*, both produced by adult male speaker of Standard Italian.*

5. Discussion

The method described here represents a further contribution to the set of tools being developed for direct image analysis of rtMRI data [23, 24]. These approaches offer important advantages over other methods of analysis, being relatively noise robust and immune to segmentation inconsistencies between frames. As such, they allow for rapid experimentation and data discovery approaches to the study of speech production, and facilitate the use of phonetic data acquired using rtMRI, without the need for extensive post-processing of image sequences.

The major limitation of the current technique is that it depends on the ability to reliably detect correlated intensity fluctuations in a specified region, and is therefore restricted in application to phonetic phenomena which produce such a change in the vocal tract, such as intervocalic consonant production.

The utility of this method will be extended with the development of more sophisticated methods of validation, and the use of richer datasets, including multi-modal data acquired from the same speakers. Higher frame rates and improved SNR will afford more accurate estimation of constriction kinematics. While all of the examples shown in this paper involve midsagittal articulation, the same method can be applied to the analysis of MR data acquired from other imaging planes, for example in the coronal analysis of tongue grooving.

6. Conclusion

Analysis of regional correlated pixel intensity variation has been shown to be a viable method of estimating articulatory activity from rtMRI data. Both place of articulation and constriction kinematics of intervocalic consonantal production can be robustly estimated from MR image sequences, without the need for tissue segmentation. The approach is beginning to provide new insights into aspects of consonant production which are difficult to study using traditional phonetic methodologies.

7. Acknowledgements

Research supported by NIH Grant R01 DC007124-01.

8. References

[1] E. Bresch, Y.-C. Kim, K. Nayak, D. Byrd, and S. Narayanan, “Seeing speech: Capturing vocal tract shaping using real-time MRI,” *IEEESPM*, vol. 25, no. 3, pp. 123–132, 2008.

[2] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, “An approach to real-time Magnetic Resonance Imaging for speech production,” *JASA*, vol. 115, no. 4, pp. 1771–1776, 2004.

[3] D. Demolin, T. Metens, and A. Soquet, “Three-dimensional Measurement of the Vocal Tract by MRI,” in *Proc. ICSLP*, Philadelphia, 1996, pp. 272–275.

[4] A. Soquet, V. Lecuit, T. Metens, B. Nazarian, and D. Demolin, “Segmentation of the airway from the surrounding tissues on MRI: A comparative study,” in *Proc. ICSLP*, Sydney, 1998.

[5] J. Behrends and A. Wismüller, “A segmentation and analysis method for MRI data of the human vocal tract,” *FIPKM*, vol. 37, pp. 179–189, 2001.

[6] E. Bresch and S. Narayanan, “Region segmentation in the frequency domain applied to upper airway real-time MRI,” *IEEE Trans. Medical Imaging*, vol. 28, no. 3, pp. 323–338, 2009.

[7] M. I. Proctor, D. Bone, and S. S. Narayanan, “Rapid semi-automatic segmentation of rtMRI for parametric vocal tract analysis,” in *Proc. Interspeech*, Makuhari, Japan, Sep 2010.

[8] J. R. Westbury, G. Turner, and J. Dembowski, “X-Ray microbeam speech production database user’s handbook,” University of Wisconsin, Tech. Rep., 1994.

[9] J. S. Perkell, M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta, and M. T. Jackson, “Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements,” *JASA*, vol. 92, no. 6, pp. 3078–3096, Dec 1992.

[10] S. Fuchs, P. Perrier, and C. Mooshammer, “The role of the palate in tongue kinematics: an experimental assessment from EPG & EMMA,” in *Proc. Eurospeech*, Aalborg, 2001, pp. 1487–1490.

[11] M. K. Tiede, “MVIEW: software for visualization and analysis of concurrently recorded movement data,” 2005.

[12] A. Löfqvist, “Lip kinematics in long and short stop and fricative consonants,” *JASA*, vol. 117, no. 2, pp. 858–878, 2005.

[13] E. L. Saltzman and K. G. Munhall, “A dynamical approach to gestural patterning in speech production,” *Haskins Laboratories Status Report on Speech Research*, vol. 99-100, pp. 38–68, 1989.

[14] E. Bresch, J. Nielsen, K. Nayak, and S. Narayanan, “Synchronized and noise-robust audio recordings during realtime MRI scans,” *JASA*, vol. 120, no. 4, pp. 1791–1794, 2006.

[15] B. Tuller and J. A. S. Kelso, “The timing of articulatory gestures: Evidence for relational invariants,” *JASA*, vol. 76, no. 4, pp. 1030–1036, 1984.

[16] K. G. Munhall, “An examination of intra-articulator relative timing,” *JASA*, vol. 78, no. 5, pp. 1548–1553, 1985.

[17] C. Atkeson, A. Moore, and S. Schaal, “Locally weighted learning,” *AI Review*, vol. 11, pp. 11–73, 1997.

[18] C. P. Browman and L. M. Goldstein, “Articulatory gestures as phonological units,” *Phonology*, vol. 6, no. 2, pp. 201–251, 1989.

[19] P. Ladefoged and I. Maddieson, *The sounds of the world’s languages*. Oxford; Cambridge, MA: Blackwell, 1996.

[20] T. A. Hall, *The phonology of coronals*. Amsterdam, Philadelphia: John Benjamins, 1997, vol. 149.

[21] P. M. Bertinetto and M. Loporcaro, “The sound pattern of Standard Italian,” *JIPA*, vol. 35, no. 02, pp. 131–151, 2005.

[22] C. Hagedorn, M. Proctor, and L. Goldstein, “Automatic analysis of geminate consonant articulation using real-time MRI,” in *Proc. ISSP*, Montreal, Canada, 2011.

[23] A. Lammert, M. I. Proctor, and S. S. Narayanan, “Data-driven analysis of realtime vocal tract MRI using correlated image regions,” in *Proc. Interspeech*, Makuhari, Japan, 2010.

[24] A. Lammert, M. Proctor, L. Goldstein, M. Pouplier, and S. Narayanan, “Automatic identification of stable modes and fluctuations in a repetitive task using real-time MRI,” in *Proc. ISSP*, Montreal, Canada, 2011.