Vocal Tract Cross-Distance Estimation from Real-Time MRI using Region-of-Interest Analysis

Adam Lammert[1]*, Vikram Ramanarayanan*[1]*, Michael Proctor*[2] *and Shrikanth Narayanan*[1]

[1]University of Southern California, Los Angeles, CA, USA)
[2]University of Western Sydney, Penrith, NSW, Australia

lammert@usc.edu, vramanar@usc.edu, michael.proctor@uws.edu.au, shri@sipi.usc.edi

Abstract

Real-Time Magnetic Resonance Imaging affords speech articulation data with good spatial and temporal resolution and complete midsagittal views of the moving vocal tract, but also brings many challenges in the domain of image processing and analysis. Region-of-interest analysis has previously been proposed for simple, efficient and robust extraction of linguistically-meaningful constriction degree information. However, the accuracy of such methods has not been rigorously evaluated, and no method has been proposed to calibrate the pixel intensity values or convert them into absolute measurements of length. This work provides such an evaluation, as well as insights into the placement of regions in the image plane and calibration of the resultant pixel intensity measurements. Measurement errors are shown to be generally at or below the spatial resolution of the imaging protocol with a high degree of consistency across time and overall vocal tract configuration, validating the utility of this method of image analysis.

Index Terms: speech production data, real-time mri, analysis tools, vocal tract area functions

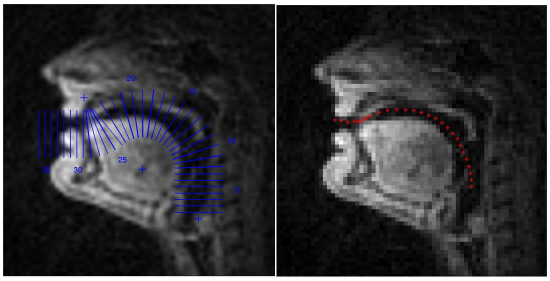
1. Introduction

Real-Time Magnetic Resonance Imaging (rtMRI) affords speech articulation data with good spatial and temporal resolution and complete midsagittal views of the moving vocal tract [1, 2]. Along with these useful characteristics, rtMRI brings many challenges in the domain of image processing and analysis. Vocal tract image sequences acquired using rtMRI are rich in information, and extracting relevant, low-dimensional representations is a non-trivial problem. Analysis techniques must take theoretical considerations into account regarding scientific/linguistic interpretability of the extracted variables, in addition to practical concerns, such as precision, robustness and efficiency.

Conventionally, the first step toward analyzing rtMRI data would be to extract edges corresponding to air-tissues boundaries. Edge extraction may be necessary in some instances, especially if the desired representation appeals directly to the postures of speech articulators (e.g., tongue body position) [3, 4]. Such boundary-tracing can be done in the spatial domain with a variety of standard edge detection algorithms (e.g., [5]). At least one algorithm has been developed to specifically operate in MRI's native frequency domain [6]. However, edge extraction tends to require a high computational cost, the task of robust edge detection from rtMRI data is further complicated by the nature of the data, which is intrinsically low SNR compared with images from standard structural MRI protocols.

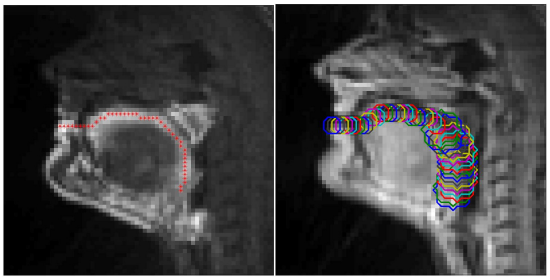
Alternative techniques for image analysis are possible if the desired representation appeals to either vocal tract constrictions (e.g., [7]) or midsagittal cross-distance functions (i.e., constriction degree at each point along the length of the vocal tract). Such representations can, of course, be determined from air-tissue boundaries, but other methods have been explored that directly use the mean pixel intensities within localized regions-of-interest (ROIs) of the vocal tract [8, 9, 10]. ROIs are commonly used in analyzing complex fMRI data, where they are used to quantify changes in brain activity in specific anatomical regions (e.g., [11]). In rtMRI data, pixel intensity values are closely related to vocal tract constrictions because they are a function of soft tissue density, and localized changes in tissue density along the vocal tract could be one definition of a vocal tract constriction. Moreover, ROI analysis boasts a high degree of robustness and very low computational cost.

Previous efforts to develop ROI analysis for rtMRI data have suggested that ROIs are a simple way to efficiently extract information about constriction degree, robustly preserving the relative degree of constriction and the timing of articulatory events. However, the accuracy of the extracted measurements has not been rigorously evaluated, and no method has been proposed to calibrate the pixel intensity values or convert them into absolute measurements of length. This work provides a validation of constriction degree measurements derived from ROI analysis in comparison to a more conventional, semi-automatic method which finds the distance between vocal tract outlines along semipolar gridlines superimposed



(a) (b)

Figure 1: *(a) Analysis grid and (b) mean vocal tract midline used for grid-based image analysis.*



(a) (b)

Figure 2: *Images show (a) Placement of ROI centers overlaid on the standard deviation image and (b) full ROIs overlaid on the mean image.*

on the vocal tract. Insights into the placement of ROIs in the image plane and calibration of the resultant pixel intensity measurements are also discussed. In addition, the efficacy of ROI analysis for extracting cross-distance functions along the entire length of the vocal tract is evaluated, an ability which hinges on the relative calibration quality measures extracted from all regions.

Section 2 provides a description of the rtMRI data used and the method of extracting cross-distances using grid-based and ROI analysis. Section 3 presents the results of rigorous quantitative comparison of cross-distances extracted using both methods. A discussion and interpretation of the results is presented in Section 4, and some concluding remarks are given in Section 5.

2. Method

Speech articulation by two male and two female speakers of American English was captured in the midsagittal plane using rtMRI [1, 2] with denoised audio [12]. These data were reconstructed into video sequences with a frame rate of 23.33 and a spatial image resolution of 68-by-68 pixels with 3mm width. The subjects read the following sentences from the well-known TIMIT corpus [13] aloud: “This was easy for us”, “Jane may earn more money by working hard”, “She is thinner than I am”, “Bright sunshine shimmers on the ocean” and “Nothing is as offensive as innocence”. This section describes methodologies for analyzing this video sequence using both ROI analysis and a more traditional grid-based, semi-automatic contour-tracking method, the latter serving as a standard for evaluating the quality of the proposed method.

All rtMRI data were subjected to an intensity correction procedure to compensate for the reduction in coil sensitivity at increasing spatial distance from the coil (e.g., moving posteriorly, from the lips toward the pharynx, for a coil located in front of the face). Differences in coil sensitivity result in lower mean and smaller dynamic range of intensity values for pixels at large distances from the coil. Intensity correction is a necessary first step for any image analysis technique relying directly on pixel intensity values as measurements to ensure that pixel intensity values at all spatial locations have the same interpretation. Coil sensitivity is relatively simple to model, and methods to correct it date back many years (e.g., [14]). For these data, a simple retrospective correction scheme was implemented, incorporating a nonparametric, monotonically increasing estimate of coil sensitivity derived from the all pixel values in the video sequence.

2.1. Grid Analysis

A traditional grid-based analysis of the vocal tract was performed to obtain cross-distances along its length. A composite semi-polar analysis grid was superimposed onto the image plane, extending from the glottis to the lips with gridlines spaced at approximately 5 mm intervals [15, 16]. An example grid for one subject can be seen in Figure 1. The grid was manually positioned relative to anatomical landmarks, namely the glottis, the highest point on the palate, the alveolar ridge and the lips. Proctor et al. [17] described this method, along with a technique for automatic tracing of vocal tract outlines in rtMRI data by identifying air-tissue boundaries intersecting with the gridlines. This method was used to produce traces of the midsagittal vocal tract outlines, which were subsequently inspected for accuracy and manually corrected when necessary.

Cross-distances were measured at gridlines from the base of the epiglottis to the most anterior point of the unprotruded lips, producing cross-distances, $d_{grid}[l, t]$, for each frame t and location l along the vocal tract. The location of these cross-distances along the vocal tract was taken to be the cumulative distance between successive gridlines, starting at the base of the epiglottis, along the mean vocal tract midline. Cross-distance measurements were later spatially resampled at 43 uniformly spaced locations along the vocal tract.

2.2. ROI Analysis

After reconstruction, rtMRI data can be considered as a gray-level video sequence, denoted as $I[m, n, t]$, where m and n represent the vertical and horizontal position of a pixel in the image plane, respectively, and t is the time associated with a particular video frame. Circular ROIs were placed in the image plane along the length of the vocal tract from near the base of the epiglottis to the lips. The centers of these ROIs were placed along the vocal tract midline by selecting pixels with the highest standard deviation across time. This method of placement ensures that regions cover as much temporal pixel intensity fluctuation as possible, which is crucial to the proposed method. Figure 2 shows an image representing the standard deviation of each pixel across time with the centers of the 43 selected ROIs overlaid, along with the outlines of the ROIs themselves.

Radius r of regions was determined by inspection of the anatomy. Radii were chosen so that regions were just wide enough to fill the cavity – i.e., so that the outer edge of the circle just touched the upper outline of the vocal tract (i.e., the palate and posterior pharyngeal wall). Accordingly, the regions in the pharyngeal cavity were nearly twice as wide as those in the oral cavity (mean 12.6 mm versus 6.9 mm, respectively, across subjects).

For a region centered at (a, b) in the image plane, the mean intensity of that region at time t is:

$$\mu[a, b, r, t] = \frac{1}{|R_{a,b,r}|} \sum_{(x,y)\in R_{a,b,r}} I[x, y, t] \qquad (1)$$

where r is the radius of the circular region and $R_{a,b,r} = \{(x, y) : r > \sqrt{(x-a)^2 + (y-b)^2}\}$ is the set of all pixels within the circle.

Calibration of the mean intensities is accomplished by normalizing individual values, μ as a function of all extracted values from a subject, $\boldsymbol{\mu}$ in the following way:

$$d_{roi}[a, b, t] = \frac{2r(max(\boldsymbol{\mu}) - \mu[a, b, r, t])}{range(\boldsymbol{\mu})} \qquad (2)$$

which functions as an estimate of the vocal tract cross-distance at location (a, b) at time t. To be consistent with measures d_{grid}, these cross-distance estimates were expressed, alternatively, as a function of location along the vocal tract midline (i.e., $d_{roi}[l, t]$). Location of these cross-distances along the vocal tract was taken to be the cumulative distance between successive region centers. Cross-distances were later spatially resampled at 43 uniformly spaced locations along the vocal tract.

3. Results

Both root-mean-square (RMS) error and Pearson's product-moment correlation coefficient were used to quantify the accuracy of measurements extracted from ROI analysis. The overall RMS error and correlation coefficient for each subject across all spatial location and temporal measurements can be seen in Table 1. RMS error values range between 2.96 and 3.18 mm. Correlation coefficients ranged between r = 0.83 and r = 0.88 ($p \ll 0.01$).

Subject	**RMSE**	**Pearson's r**
M1	3.18	0.85
M2	2.96	0.88
F1	3.07	0.86
F2	3.10	0.83

Table 1: *Overall RMS error and correlation for each subject, across all spatial locations and temporal samples.*

To further illustrate the performance of ROI analysis the utterance "This was easy for us" from subject M2 was analyzed in more detail. In particular, RMS error and Pearson's r were calculated both across time, assessing the accuracy of a particular ROI, and across space, looking at the accuracy of all regions at a particular time. For instance, correlation over all spatial locations at time t would involve the correlation between all values $d_{grid}[*, t]$ and $d_{roi}[*, t]$, whereas the correlation over all frames at location l would involve the correlation between all values $d_{grid}[l, *]$ and $d_{roi}[l, *]$. RMS error was calculated by considering the same sets of measurements. Figure 3 shows the evolution of RMS error and correlation coefficients for each temporal sample (i.e., image frame), while Figure 4 shows the RMS error and correlation associated with individual spatial locations along the vocal tract, looking across time.

4. Discussion

The overall accuracy of ROI analysis was consistently high across subjects. Errors were on the order of a single pixel's width in the current rtMRI protocol, which may be considered a reasonable lower bound on error measurements. The remainder of this section discusses the detailed analysis of one sentence from speaker M2, which provides deeper insight into the strengths and limitations of the proposed method.

Accuracies vary somewhat at different locations along the vocal tract, as can be seen in Figure 4, with the worst results being in the upper pharynx and velar region (centered around 0.3/L mm from the glottis, where L is vocal tract length), as reflected in both the RMS error and correlation coefficients. This degraded performance likely results from interference of the velum impinging on the ROIs. Difficulty was also observed in the region nearest to the teeth (centered around 0.8/L mm from the glottis) in terms of RMSE, although the correlations were still very high. This is most likely due to the general difficulty of dealing with the teeth in MRI data, since they do not image within this modality.

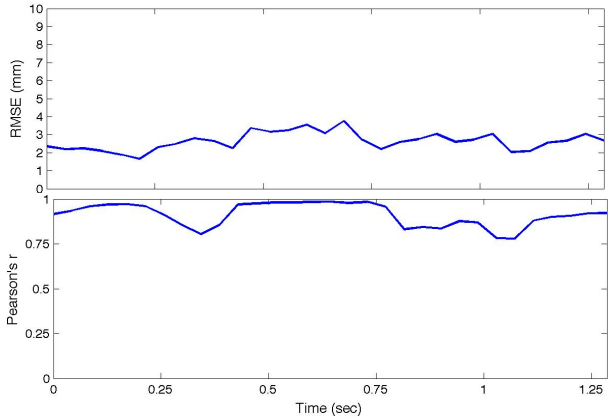


Figure 3: *Accuracy of the cross-distance function (RMS error above, correlation below) across time during subject M2's utterance "This was easy for us".*

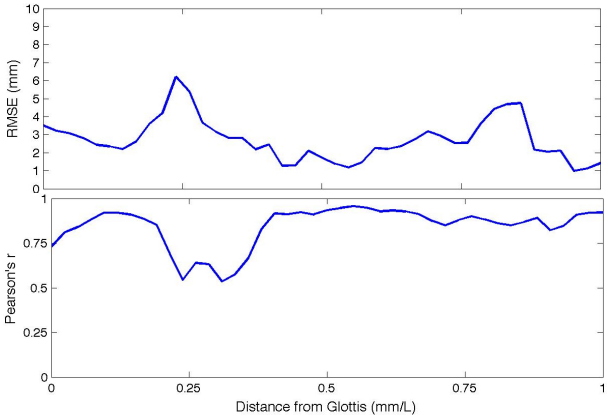


Figure 4: *Accuracy of cross-distances by location within the vocal tract (RMS error above, correlation below) during subject M2's utterance "This was easy for us".*

Accuracies were very consistent across time, both in terms of RMS error and correlation. Some difficulty was associated with relatively neutral vocal tract postures. Local minima in correlation values can be seen in Figure 3 at approximately 0.33 and 1.1 seconds, corresponding to the /ʌ/ in the words "was" and "us". For such postures, which display a relatively flat cross-distance function, correlation may be a poor measure of performance. Note that RMS error does not increase at those same times.

Figure 5 shows that cross-distances estimated with ROI analysis have a highly linear relationship with those from grid-based analysis, indicating that the proposed conversion of ROI values into millimeter measurements (see Equation 2) has the correct functional form. Only a slight deviation from slope of unity can be observed in the relationship. Regression analysis suggests that an adjustment to ROI cross-distances conforming to $1.2d_{roi} - 3.5$ would provide an additional 11% decrease in RMS error.

5. Conclusion

ROI analysis is a simple method that allows for accurate extraction of constriction degree information and cross-distance functions directly from pixel intensity values when applied to rtMRI video sequences of the vocal tract. Realtime MR image sequences contain rich information about the movement and coordination of speech articulators within the vocal tract. Future work will apply ROI analysis to a variety of rtMRI data that has been collected to facilitate modeling of speech articulation time series.

There are several ways to improve the accuracy of the proposed method. Better calibration of pixel intensities into absolute length measurements may be possible. It was shown in Section 4 that a slight linear adjustment to the current calibration may be necessary. Coil sensitivity estimation and subsequent pixel intensity correction – considered an open problem in the medical imaging community – can also be improved over the fairly simple scheme used in this work. Finally, it was implicitly assumed that the subjects' heads were stationary throughout the scan, ensuring that the midsagittal and image planes coincide. Indeed, head motion was not observed with the current subjects, but for the minority of subjects who do show appreciable head motion during scans – usually on the order of a few millimeters – correction is necessary to ensure accurate measurements. Note that most of these improvements affect the conversion of pixel intensities to absolute measurements, and therefore will not substantially affect the evaluation results using correlation presented in this work.

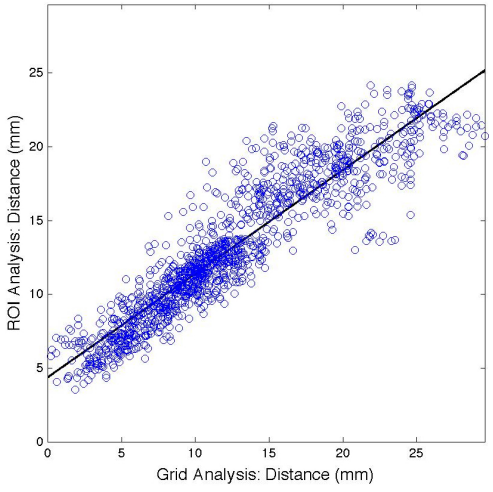


Figure 5: *Cross-distances from grid-based and ROI analysis, from all spatial locations and temporal samples during subject M2's utterance "This was easy for us".*

6. Acknowledgements

Work supported by NIH Grant DC007124 and a graduate traineeship NIH 5T32DC009975.

7. References

[1] 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.

[2] Narayanan, S., Bresch, E., Ghosh, P., Goldstein, L., Katsamanis, A., Kim, Y., Lammert, A., Proctor, M. , Ramanarayanan, V. and Zhu, Y., "A Multimodal Real-Time MRI Articulatory Corpus for Speech Research", in Proceedings of INTERSPEECH, 2011.

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

[4] Maeda, S., "Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model.", in Speech Production and Speech Modelling [Hardcastle, W.J. and Marchal, A., eds.], Kluwer: Netherlands, 131–149, 1990.

[5] Canny, J., "Computational Approach To Edge Detection", IEEE Transactions Pattern Analysis and Machine Intelligence, 8(6):679–698, 1986.

[6] 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.

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

[8] Bresch, E., Katsamanis, A., Goldstein, L. and Narayanan, S., "Statistical multi-stream modeling of real-time MRI articulatory speech data," in INTERSPEECH: 1584–1587, 2010.

[9] Lammert, A., Proctor, M. and Narayanan, S., "Data-Driven Analysis of Realtime Vocal Tract MRI using Correlated Image Regions", in INTERSPEECH: 1572–1575, 2010.

[10] Proctor, M., Lammert, A., Katsamanis, A., Goldstein, L., Hagedorn, C., and Narayanan, S., "Direct Estimation of Articulatory Kinematics from Real-time Magnetic Resonance Image Sequences", In INTERSPEECH:281-284, 2011.

[11] Engel, S.A., Rumelhart, D.E., Wandell, B.A., et al., "fMRI of human visual cortex", Nature 369:525, 1994.

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

[13] Wrench, A. and Hardcastle, W.J., "A multichannel articulatory speech database and its application for automatic speech recognition", in Proceedings of 5th SSP, 305–308, 2000.

[14] Axel, L., Constantini, J. and Listerud, J., "Intensity Correction in Surface-Coil MR Imaging", Journal of Roentgenology, 148:418–420, 1987.

[15] Öhman, S.E.G., "Numerical model of coarticulation", Journal of the Acoustical Society of America, 41(2):310–320, 1967.

[16] Maeda, S., "Un modèle articulatoire de la langue avec des composantes lineaires", in 10ème Journées d'Etude sur la Parole, 1979.

[17] Proctor, M.I., Bone, D., Katsamanis, A. and Narayanan, S.S., "Rapid Semi-automatic Segmentation of Real-time Magnetic Resonance Images for Parametric Vocal Tract Analysis", in Proceedings of INTERSPEECH, 2010.