Toward Automatic Vocal Tract Area Function Estimation from Accelerated Three-dimensional Magnetic Resonance Imaging

Yoon-Chul Kim[1], *Jangwon Kim*[1], *Michael Proctor*[2], *Asterios Toutios*[1], *Krishna Nayak*[1], *Sungbok Lee*[1], *Shrikanth Narayanan*[1]

[1] Department of Electrical Engineering, University of Southern California, Los Angeles, USA
[2]Department of Linguistics, University of Western Sydney, Sydney, Australia

yoonckim@usc.edu, jangwon@usc.edu, mike.i.proctor@gmail.com, toutios@sipi.usc.edu, knayak@usc.edu, sungbokl@usc.edu, shri@sipi.usc.edu

Abstract

Vocal tract area function estimation from three-dimensional (3D) volumetric dataset often involves complex and manual procedures such as oblique slice cutting and image segmentation. We introduce a semi-automatic method for estimating vocal tract area function from 3D Magnetic Resonance Imaging (MRI) datasets. The method was implemented on a custom MATLAB graphical user interface and computes the area function in a user-interactive way. The 3D MRI datasets were acquired with 1.25 mm isotropic resolution during 8-seconds sustained sound productions of vowels /IY/, /AA/, /UW/ by one male native speaker of American English at a 3 Tesla MRI scanner.

Index Terms: speech production, magnetic resonance imaging, image segmentation, area function, vocal tract shape.

1. Introduction

Direct 3D measurements of vocal tract shape by Magnetic Resonance Imaging (MRI) allow realistic estimation of the morphological structure of the vocal tract, which provides valuable information not only for speech production research but also for other application purposes such as more accurate articulatory-to-acoustic mapping and speech inversion. Detailed studies on the inter-speaker variations in vocal-tract shaping and vocal tract morphology are also possible. Previous studies have recorded 3D vocal airway data for sustainable speech sounds, such as vowels, fricatives, and liquids [1-6] and also have proposed a few methods for area function estimation based on the 3D measurements of the vocal tract shape [5,7,8,23]. They have also demonstrated their 3D measurement and area function estimation accuracy, and have shown that the acoustic characteristics, mostly examined with formant frequency, of the simulated vowel sounds by using the direct 3D measurements and estimated area function is close to those of naturally spoken vowel sounds [5]. The results suggested that the area function estimation from the 3D measurements can be a promising tool for articulatory analysis, modeling [24], synthesis, and inverse mapping [25].

In previous studies using 2D vocal tract image data which are easier to obtain than 3D data, the area function of the vocal tract was estimated by measuring mid-sagittal distance [9-11]. However, area function is not fully determined by the mid-sagittal distance, so direct measurement from 3D data can enhance the accuracy of area function estimation, and eventually can improve the quality of articulatory synthesis [12,13].

In most of the previous studies, area function estimation from 3D MRI data usually involved tedious and time-consuming manual segmentation of the air space surrounded by the soft tissue in the upper airway. The challenge for automatic segmentation presumably stems from the complex shapes and variations of the cross-sectional airway areas in the vocal tract. In this paper, we present a user-interactive and semi-automatic image analysis technique, which seeks to facilitate area function estimation with minimal user interactions. Our area function estimation method is based on the grid system proposed by Öhman [14] and Mermelstein [15], which has been recently adopted for developing an articulatory-acoustic analysis tool [16].

2. Methods

2.1. MRI data acquisition

Data collection was performed using a GE 3.0 Tesla HDxt scanner system at the Healthcare Consultation Center II, University of Southern California.

One male adult American English speakers was recruited and participated in the 3D MRI data collection. The subject was consented prior to the MRI study. A body coil was used for radiofrequency (RF) transmission, and a commercial 8-channel neurovascular receiver coil was used for RF signal reception. A prospective acquisition of undersampled three-dimensional Fourier Transform (3DFT) gradient echo (GRE) pulse sequence was employed [17]: Pulse repetition time (TR) = 4.2 ms, excitation slab thickness = 8 cm, FOV = 200 × 200 × 100 mm^3, spatial resolution = 1.25 × 1.25 × 1.25 mm^3, matrix size = 160 × 160 × 80. The readout was along the superior-inferior (S-I) direction. The phase encode was along the anterior-posterior (A-P) direction. The slice encode was along the right-left (R-L) direction. The k-space undersampling pattern was designed based on a Poisson disc sampling [18]. For the fully sampled acquisition, the scan time was 56 seconds. We performed a prospective acquisition with an acceleration factor of 7, which resulted in the scan time of 8 seconds.

The subject lay supine and was able to read the stimuli relying on our mirror-projector setup, and sustained each vowel or consonant for 8 seconds. A whole set of the stimuli was: 13 American English sustained vowel sets (“b**ee**t”, “b**i**t”, “b**ai**t”, “b**e**t”, “b**a**t”, “p**o**t”, “b**u**t”, “b**ou**ght”, “b**oa**t”, “b**oo**t”, “p**u**t”, “b**ir**d”, “abb**o**t”), 9 American English sustained fricatives (“a**f**a”, “a**v**a”, “a**th**a” as in thing, “a**th**a” as in this, “a**s**a”, “a**z**a”, “a**sh**a”, “a**g**e**a**” as in beige, “a**h**a” as in happy), 3 American English sustained nasals (“a**m**a”, “a**n**a”, “a**ng**a”), 2 American English sustained liquids (“a**l**a”, “a**r**a”), and 6 non-speech tongue/velum gestures (“breathe normally with your mouth closed”, “clench your teech”, “stick your tongue out as far as you can”, “pull back your tongue as far into the mouth”,

"raise your tongue tip to the palate", "hold your breath"). The 3D MRI data collection of the 33 stimuli was remarkably time-efficient compared to the imaging time reported in the literature [1-8] and only took 20 to 30 minutes to record 2 repetitions of the whole stimuli when excluding the time for the positioning of the patient in the scanner. A fiberoptic microphone was close to the subject's mouth, and audio recordings were made simultaneously with the MRI scans.

2.2. MRI image reconstruction

Image reconstruction of the 3D Fourier dataset was based on the combined use of compressed sensing and parallel imaging. The 3D Fourier dataset was first 1D Fourier transformed along the readout direction so that the dataset was transformed to the (x, ky, kz) domain. For each x, a compressed sensing parallel imaging reconstruction was performed in the (ky, kz) domain. The reconstruction was based on conjugate gradient optimization and was terminated at the 20th iteration. The reconstruction was effective in removing incoherent spatial aliasing artifacts resulting from variable-density Poisson disc undersampling. The 2D reconstructed images covering the vocal tract region of interest were stacked to result in a 3D volumetric image data (see Fig 1).

2.3. Image processing

Intensity correction of the 3D volumetric image data was performed by the normalization of each coronal slice image individually. This approach is based on the fact that there is significant coil sensitivity roll-off in the anterior-posterior direction with the use of the 8-channel neurovascular receiver coil. For example, intensity drop from the lips to the pharyngeal wall was observed in a mid-sagittal slice image when an image reconstruction considered data only from the anterior coil elements.

After the intensity correction, 2D image de-noising was performed using anisotropic diffusion [19] and was performed on each sagittal slice image. Resulting images are shown in Figure 1, which reveal excellent contrast between the air and soft tissue in all three orthogonal views of the 3D volumetric image data.

2.4. Drawing of the gridlines

For drawing of the grid lines, we adopted the algorithm that Proctor et al. [16] had previously developed for the analysis of mid-sagittal 2D real-time MRI data, and then incorporated it into our existing static 3D GUI analysis tool [20]. Drawing the grid lines requires the user's selection of four anatomical landmarks: 1) glottis, 2) maximum height in the hard palate, 3) alveolar ridge, and 4) middle point between the upper and lower lips. The center point (see the red dot in Fig. 1) in the polar grid line group is determined based on the locations of manually selected landmarks. Similarly, the center point (see the yellow dot in Fig. 1) in the reverse polar grid line group is determined. Eventually, the algorithm automatically results in the formation of the grid lines as shown in Fig. 1.

2.5. Slice cutting and image segmentation

Slice cutting from the horizontal grid lines is straightforward. The polar grid lines are oblique and the same is true for the reverse polar grid lines. With the angle information of the grid

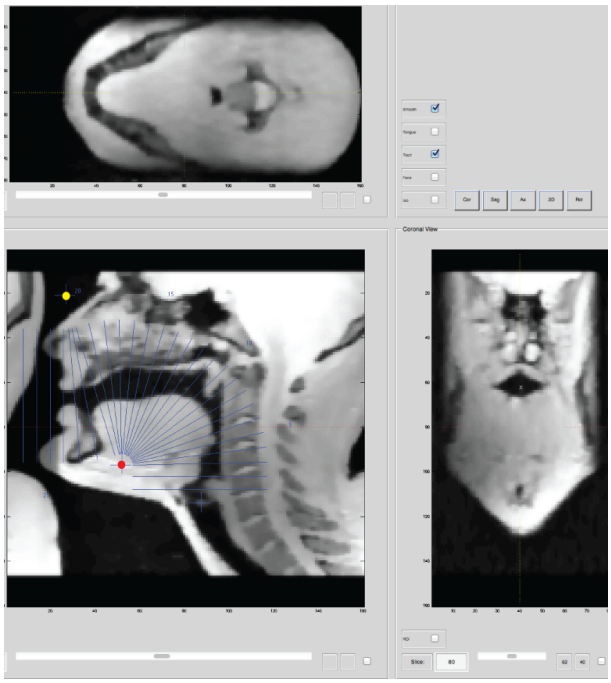

Figure 1: *A snapshot of the proposed MATLAB graphical user interface. Three orthogonal views (axial, sagittal, coronal) and grid lines in the mid-sagittal slice are shown. High quality isotropic resolution MRI data were available from our accelerated 3D upper airway MR imaging during a single sustained sound production of 8-seconds. The graphical user interface analysis tool, with the high quality isotropic 3D data loaded, facilitates a semi-automatic extraction of vocal tract cross-sectional area in any arbitrary orientation given the grid lines.*

lines, we made the rotation of the 3D volume via an affine transformation. Then, the slice of interest from either the polar or reverse polar grid lines was axially oriented.

We took the axial slice and performed a seeded region growing [21] to segment cross-sectional airway. Seed points were selected manually. In case that the cross-sectional airway was not fully surrounded by the soft tissue, the segmentation algorithm propagated to the background region, and it overly estimated areas (for example, near the lip region indicated by the yellow arrows in Fig. 2). In that case, the algorithm was automatically switched to the standard Matlab function `roipoly()`.

2.6. Cross-sectional area and mid-sagittal width

Cross-sectional area was defined to be the multiplication of the number of segmented pixels by pixel resolution (i.e., 0.125 × 0.125 cm^2). Mid-sagittal width was defined to be the multiplication of the number of segmented pixels in the mid-sagittal slice image of the 3D volume by pixel size (i.e., 0.125 cm).

3. Results and Discussion

The 3D MRI datasets provided excellent contrast between the air and soft tissue, which facilitated extraction of vocal tract cross-sectional area via region growing segmentation. The

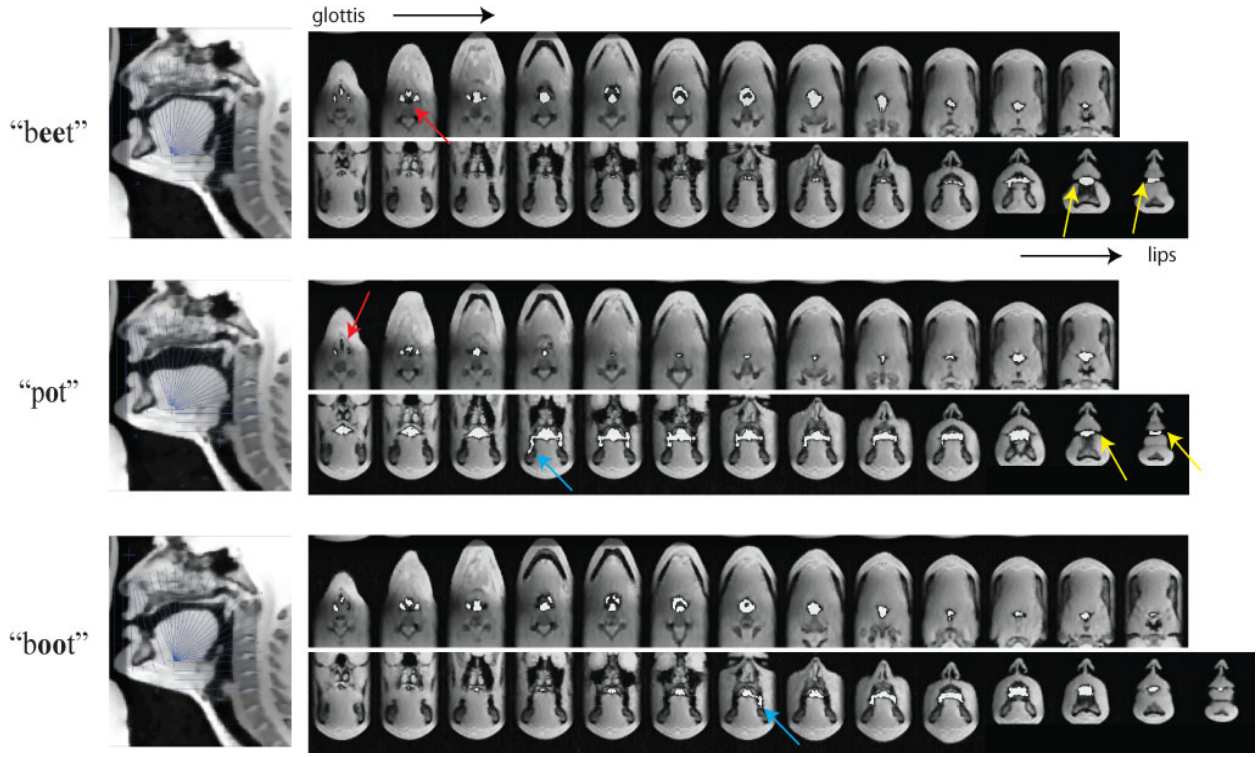

Figure 2. *Region growing segmentation results for sustained speech sounds /IY/, /AA/, /UW/. Cross-sectional airways were cut from the grid lines. Segmented airways overlaid onto cross-sectional airway images are shown from (upper left) the glottis to (lower right) the lips for each speech sound.*

proposed technique provides a convenient way to estimate vocal tract area function from a 3D dataset using a custom Matlab graphical user interface script.

The region growing algorithm works based on similarity in neighboring pixel intensity, and was problematic in the segmentation of a few slices. First, region growing stopped before it reached the air-tissue boundaries (e.g., see the red arrows in Fig 2.). Second, "bleeding" resulting from the region growing segmentation was observed in some oblique coronal slices especially near the lateral portions in the teeth or the air cavity region (e.g., see the blue arrows in Fig. 2).

Figure 3 compares cross-sectional areas from the grid lines with squared mid-sagittal widths from the same grid lines for /IY/, /AA/, /UW/ during the sustained vowel speech sounds of "**b**ee**t**", "**p**o**t**", "**b**oo**t**", respectively. Although the simple squared mid-sagittal width plot appears to have a similar pattern to the estimated area function plot for /IY/, /AA/, and /UW/, the discrepancy suggests that one may need a more elaborated conversion method for vowels. For instance, a better area conversion function should consider not only the mid-sagittal widths but also the lateral dimension of the vocal tract as well as the position in the vocal tract as effective functional parameters.

4. Conclusions and Future Works

We have demonstrated a novel way of extracting the vocal tract area function from 3D MRI vowels dataset. The method was implemented in a custom MATLAB graphical user interface for minimal user interactions: manual selections of four anatomical landmarks and seed points for image segmentation, and drawing of the air-tissue boundary with `roipoly()` in a few problematic slices where region growing algorithm fails due to open airway near the lips.

There is a lot of room for improvement in our current area function estimation method. First, the automatic and accurate identification of the seed points from a cross-sectional airway image will reduce time and effort for the region growing segmentation task. Second, a routine for automatic drawing of the mid-line on the mid-sagittal upper airway image may be necessary. The mid-line can be useful for estimating the vocal tract length and standardizing the setting of the grid-line for area function estimation. For example, one can maintain the same distance along the mid-line between every two adjacent grid lines. Currently we are working for such improvements in the software.

It is important to evaluate the accuracy of our area function estimation method. We will compare the acoustic characteristics, e.g., formant frequencies, of simulated speech generated by a tube-model based synthesizer, especially for vowels, and naturally spoken speech.

Although it is a challenging problem, an important potential application of the 3D MRI data is more accurate estimation of the vocal tract area function from mid-sagittal images obtained using (2D) real-time MRI [22]. It is important to note that the 3D MRI data is acquired during sustained speech while the 2D real-time MRI data is acquired during running speech. To the best of our knowledge, this problem has received only ad hoc solutions in the past.

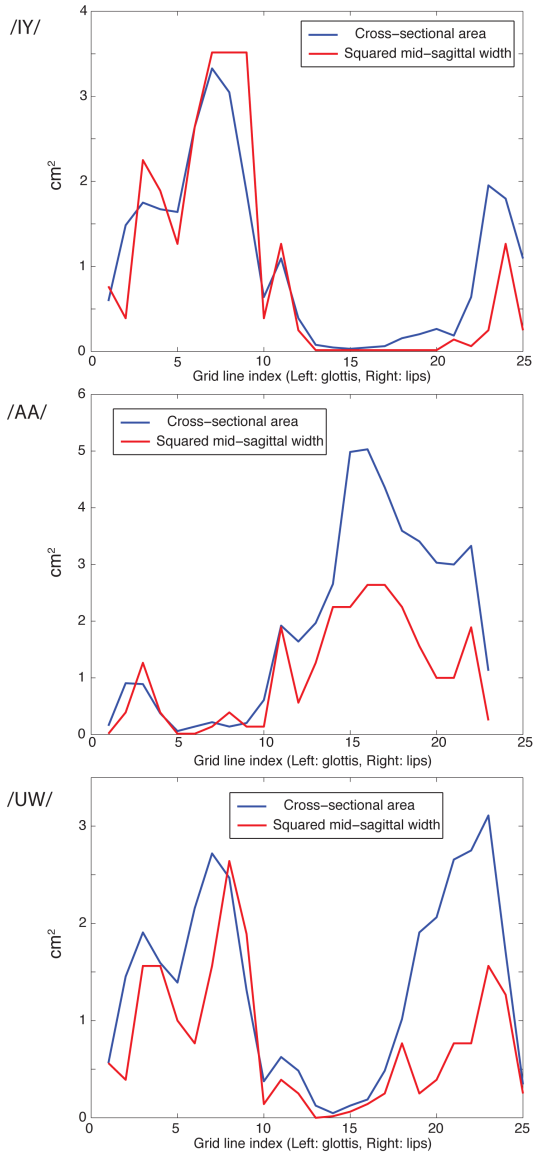

Figure 3. *Area function (blue) and corresponding squared mid-sagittal width (red) measurements for (top) /IY/, (middle) /AA/, (bottom) /UW/ from the male American English speaker.*

5. Acknowledgements

We would like to thank the grant support MIT-Lincoln Lab, NIH NIDCD R01-DC007124, and NSF IIS-1116076.

6. References

[1] Baer, T., Gore, J.C., Gracco, L.C, Nye, P.W., "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels," J Acoust Soc Am, 90(2), p799-828, 1991.

[2] Alwan, A., Narayanan, S., Haker, K., "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics," J Acoust Soc Am, 101, p1078–1089, 1997.

[3] Narayanan, S., Alwan, A., Haker, K., "An articulatory study of fricative consonants using magnetic resonance imaging", J Acoust Soc Am, 98:1325-1347, 1995.

[4] Narayanan, S., Alwan, A., Haker, K., "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I. The laterals," J Acoust Soc Am, 101(2), p1064–1077, 1997.

[5] Story, B.H., Titze, I.R., Hoffman, E.A. "Vocal tract area functions from magnetic resonance imaging," J Acoust Soc Am, 100, 537-554, 1996.

[6] Narayanan, S., Alwan, A., Song, Y., "New results in vowel production: MRI, EPG, and acoustic data." Proc. of European Speech Proc Conf.; Rhodes, Greece. Sept. pp. 1007–1009, 1997.

[7] Demolin, D., Metens, T., Soquet, A., "Three-dimensional measurement of the vocal tract by MRI," Spoken Language, 1996. ICSLP 96. Proceedings, Fourth International Conference on , vol.1, no., pp.272,275 vol.1, 3-6 Oct., 1996.

[8] Story, B. H. "Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002," J Acoust Soc Am, 123(1), 327-335, 2008.

[9] Perrer, P., Boe, L.J., Sock, R., "Vocal tract area function estimation from midsagittal dimensions with CT scans and a vocal tract cast: modeling the transition with two sets of coefficients," J of Speech and Hearing Research, 35:53-67, 1992.

[10] Beautemps, D., Badin, P., Laboissiere, R., "Deriving vocal-tract area functions from midsagittal profiles and formant frequencies: a new model for vowels and fricative consonants based on experimental data," Speech Communication, 16:27-47, 1995.

[11] Sundberg, J., Johanson, C., Widbrand, H., Ytterbergh, C., "From sagittal distance to area: A study of transverse, vocal tract cross-sectional area," Phonetica, 44:76-90, 1987.

[12] Maeda, S., "A digital simulation method of the vocal tract system," Speech Communication, 1(3-4): 199-229, 1982.

[13] Sondhi, M.M. and Schroeter, J., "A hybrid time-frequency domain articulatory speech synthesizer," IEEE Transactions on Acoustics, Speech, and Signal Processing, 35(7):955-967, 1987.

[14] Öhman, S.E.G., "Numerical model of coarticulation," J Acoust Soc Am, 41(2), p310-320, 1967.

[15] Mermelstein, P., "Articulatory model for the study of speech production," J Acoust Soc Am, 53(4), p1070-1082, 1973.

[16] Proctor, M.I., Bone, D., Narayanan, S.S., "Rapid semi-automatic segmentation of real-time magnetic resonance images for parametric vocal tract analysis," in Proc. Interspeech, Makuhari, Japan, p1576-1579, 2010.

[17] Kim, Y.C., Narayanan, S.S., Nayak K.S., "Accelerated three-dimensional upper airway MRI using compressed sensing", Mag Reson in Med., 61(6), p1434-1440, 2009.

[18] Lustig, M., Alley, M., Vasanawala, S., Donoho, D., Pauly, J.M., "L1-SPIRiT: Autocalibrating parallel imaging compressed sensing", Proc. ISMRM, p379, 2009.

[19] Perona, P., Malik, J., "Scale-space and edge detection using anisotropic diffusion," IEEE Trans. PAMI, 12(7), p629-639, 1990.

[20] Bone, D.K., Proctor, M.I., Kim, Y.C., Narayanan, S.S., "Semi-automatic modeling of tongue surfaces using volumetric structural MRI", J. Acoust. Soc. Am., 130(4), p2549, 2011.

[21] Adams, R., Bischof, L., "Seeded region growing," IEEE Trans. PAMI, 16(6), p641-647, 1994.

[22] Narayanan, S., Nayak, K., Lee, S., Sethy, A., Byrd, D., "An approach to real-time magnetic resonance imaging for speech production," J Acoust Soc Am 115, p1771–1776, 2004.

[23] Kröger, B.J., Winkler, R., Mooshammer, C., Pompino-Marschall, B., "Estimation of vocal tract area function from magnetic resonance imaging: preliminary results," Proceedings of the 5th seminar on speech production, p333-336, 2000.

[24] Yehia, H., Tiede, M., "A parametric three-dimensional model of the vocal-tract based on MRI data," Proc. ICASSP, 3, p1619-1622, 1997.

[25] Dang, J., Honda, K., "Estimation of vocal tract shapes from speech sounds with a physiological articulatory model," J of Phonetics, 30(3), p511-532, 2002.