ESTIMATION OF VOCAL TRACT AREA FUNCTION FROM VOLUMETRIC MAGNETIC RESONANCE IMAGING

Zisis Iason Skordilis, Asterios Toutios, Johannes Töger, Shrikanth Narayanan

Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA, USA

skordili@usc.edu, toutios@usc.edu, toger@usc.edu, shri@sipi.usc.edu

ABSTRACT

The acoustic properties of speech signals are largely determined by the shaping of the vocal tract. Thus, measurements of vocal-tract area functions and their relationship to various properties of the speech signal have been of interest to the speech research community. Recent advances in Magnetic Resonance Imaging (MRI) allow direct 3D volumetric imaging of the upper airway during production of sustained sounds in as little as seven seconds, therefore allowing direct measurements of vocal-tract area functions for a variety of speech sounds, including fricative and liquid consonants. In this work we present a tool for semi-automatic vocal-tract area function estimation from such data and demonstrate its utility for estimation of the area function for various sustained sounds. Such estimations can be used to address the problem of sagittal-to-area conversion in order to allow inference of 3D vocal-tract shaping dynamics from mid-sagittal real-time MRI data.

Index Terms— vocal tract area function, volumetric MRI

1. INTRODUCTION

The human vocal tract is the main instrument of speech production. Humans are able to produce different sounds by controlling and modulating its shape. Speech researchers have been interested in characterizing the mapping between vocal tract shape and the acoustic properties of the speech signal. To this end, a representation of vocal tract shaping is needed. A commonly used such representation is the area function, namely the cross-sectional area of the airway as a function of distance from the glottis. This representation is motivated by the simplifying assumption of one-dimensional wave propagation in the vocal tract [1]. This assumption results in a model of the vocal tract as a stack of cylindrical tubes with varying cross-sectional areas. The area function gives the cross-sectional areas for the tubes in the model. The study of the relationship between area functions, which give a vocal tract shape representation, and the acoustic properties of the speech signal has received much attention in the literature, especially through speech simulation and articulatory synthesis [1–3]. The existence of an inventory of measurements of area functions from real-world data for various sounds is crucial for such studies [1].

Recent advances in Magnetic Resonance Imaging (MRI) allow direct volumetric imaging of the upper airway during production of sustained sounds [1, 4–9]. This enables direct measurement of area functions in three dimensions (3D). Area function measurements through midsagittal airway width from two dimensional (2D) midsagittal MRI have been reported earlier [10, 11]. However, midsagittal width does not fully determine the area function [2, 3, 6]. Volumetric MRI allows more accurate area function estimation.

Area function measurement from 3D MRI data requires airway segmentation from the surrounding soft tissue, which can be very challenging and time consuming to manually accomplish.

In this work, we propose an improved semi-automatic area function estimation method. Our method builds upon and further improves the automation of the one proposed by Kim et al. [6]. We use an analysis grid based on the methods of Öhman, Maeda, and Mermelstein [12–14]. We propose a method for vocal tract center-line estimation on the midsagittal plane. This was not included in the previous are function estimation method [6]. We use seeded region growing on each grid line slice to segment the airway and measure its area. We do not require manual seeding of each grid slice. We use a midsagittal airway segmentation to automatically place seeds on each grid line slice. This is an improvement in automation over the previously proposed method [6], which required the user to manually seed the airway on each slice, a quite cumbersome procedure due to the relatively large number of grid line slices.

We demonstrate the utility of our proposed method on a dataset consisting of volumetric MR images of sustained contextualized continuants. The dataset was collected using an accelerated protocol which requires $8\mathrm{s}$ to scan the entire volume [6, 7]. The fast scan time enabled collection of data for fricative and liquid consonants in addition to vowels. The dataset includes 17 subjects in total with 27 continuants recorded per subject. In this work, we use a subset of 4 subjects and 5 vowels per subject. We report the estimated area functions, which exhibit shapes similar to those previously reported [1]. We validate the estimated area functions by applying the articulatory synthesis technique proposed by Maeda [2]. We subjectively verify that the correct acoustics for each vowel were generated by the speech synthesizer.

2. DATASET

To demonstrate our method for direct area function estimation, we use a dataset consisting of volumetric MR images of sustained contextualized continuants. The data were collected using an accelerated imaging protocol that allows acquisition of the entire volume in $8\mathrm{s}$ [6, 7]. The data acquisition process, MRI protocol, and image reconstruction method have been previously described in Kim et al. [6]. The data were collected with a GE $3.0\,\mathrm{Tesla}$ HDxt scanner system at the Healthcare Consultation Center II, University of Southern California, using a standard 8-channel neurovascular receiver coil. The acquired data have isotropic spatial resolution of $1.5625\mathrm{mm}$ with an image size of $160\,(\mathrm{axial}) \times 160\,(\mathrm{coronal}) \times 80\,(\mathrm{sagittal})$. Data were collected for 17 subjects, all native speakers of American English. For each subject 27 scans were performed. During each scan the subject sustained a contextualized continuant for the scan duration ($8\mathrm{s}$). The following continuants were recorded: 13 vowels (abb**o**t, b**a**t, p**o**t, b**u**t, b**i**rd, b**ai**t, b**e**t, b**i**t, b**ee**t, b**oa**t, b**ou**ght,

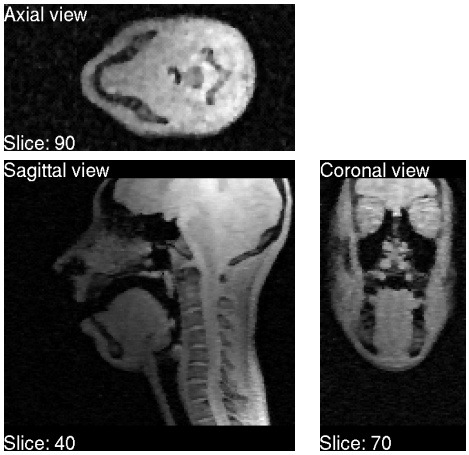

Fig. 1. *An example from the volumetric MRI dataset: an axial, a sagittal, and a coronal slice for a female subject for the stimulus b**a**t. A stack of slices forms the volume.*

b**oo**t, p**u**t), 9 fricatives (a**f**a, a**v**a, a**h**a, a**sh**a, a**g**a as in beige, a**s**a, a**z**a, a**th**a as in thing, a**th**a as in this), 3 nasals (a**m**a, a**n**a, a**ng**a), and 2 liquids (a**l**a, a**r**a).

The advantage of using an accelerated 3D protocol with 8s scan duration is the collection of data that is as ecologically valid as possible [15]: for most of the aforementioned continuants, the subjects were able to actively produce speech throughout the scan. This eliminates the need for repeated scans to acquire a particular vocal tract shape or for artificially holding tongue postures without active speech production for the purpose of collecting volumetric MRI data. The short scan duration allows for ecologically valid collection not only of vowels (which can be sustained longer) but also of fricative and liquid consonants.

In the present work, we will use a subset of this dataset to demonstrate our area function estimation method. We will consider 4 subjects and 5 vowels for each subject (b**a**t, b**e**t, b**ee**t, b**ou**ght, b**oo**t).

3. AREA FUNCTION ESTIMATION METHOD

The semi-automatic method for area function estimation from volumetric MRI data that we have developed builds upon and extends the previously proposed method by Kim et al. [6]. We improve the automation of the previous method and add a facility for midsagittal airway centerline estimation. Our method consists of the following stages: image preprocessing and enhancement; drawing of vocal tract grid lines on the midsagittal plane; midsagittal airway segmentation and center-line estimation; slice cutting along the grid line directions; automatic airway segmentation and area estimation on the grid line slices, and finally, manual inspection of the segmented airway cross-sections and correction where needed. For segmentation we use a modified version of seeded region growing which we describe first. We then describe the aforementioned stages of our area function estimation algorithm in detail.

3.1. Seeded region growing

To segment an image into connected regions of homogeneous intensity, Adams and Bischof proposed the seeded region growing algorithm [16]. The algorithm begins with a manually specified seed, a small area inside the region of interest in the image. The seed is an initial estimate S_0 of the desired region. Region growing proceeds iteratively: at each iteration i, with current region estimate S_i, the algorithm considers the neighboring pixels $\mathrm{N}(S_i)$ of S_i and computes

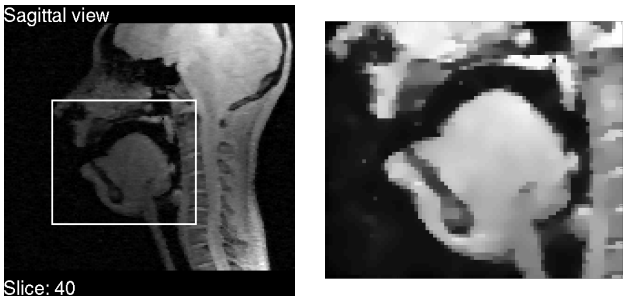

Fig. 2. *Image preprocessing and enhancement. Left: Raw data with manually specified bounding box around the region of interest. Right: Enhanced image after cropping, intensity correction, and denoising.*

the following intensity distance metric:

$$\delta(j) = |\phi(j) - \overline{\phi_{S_i}}|,\ j \in \mathrm{N}(S_i) \tag{1}$$

where $\phi(\cdot)$ is the image intensity function, and $\overline{\phi_{S_i}}$ the mean intensity of pixels in S_i. Based on their δ values, the neighboring pixels of S_i are enqueued in a priority queue Q. The algorithm updates the region estimate S_i by adding to the current region the pixel in the queue with intensity closest to the current region mean intensity:

$$S_{i+1} = S_i \cup \{\arg\min_{j \in Q} \delta(j)\} \tag{2}$$

The algorithm stops when the minimum intensity distance $\min_{j \in Q} \delta(j)$ of pixels available in Q exceeds a predefined threshold T. The result is a connected region of homogeneous intensity.

Since we are specifically interested in airway segmentation in MRI images, we use a modified seeded region growing algorithm previously proposed by Skordilis et al. [15]. If a pixel has lower intensity than the current airway region estimate, then we add it to the current region regardless of its absolute intensity distance from the region mean. This is equivalent to using the following modified intensity distance metric:

$$\delta(j) = \phi(j) - \overline{\phi_{S_i}},\ j \in \mathrm{N}(S_i) \tag{3}$$

This modification is motivated by the fact that the airway is a region of low (ideally zero) intensity in MR images. If a pixel has lower intensity than the mean intensity of the current airway region then, regardless of its exact intensity value, it is most likely part of the airway. Henceforth, we will use the term "seeded region growing" to refer to this modified version of the algorithm.

3.2. Image preprocessing and enhancement

First, we crop the MRI volume to the region surrounding the vocal tract by manually specifying a bounding box on the midsagittal plane (Fig. 2). Intensity correction is needed due to coil sensitivity roll-off in the anterior-posterior direction. We apply intensity correction by individually normalizing each coronal slice by its average tissue intensity [6] (we detect tissue by thresholding the intensity of the slice using Otsu's method [17]). Finally, we denoise the MRI volume using anisotropic diffusion on each sagittal slice individually [6, 18]. Anisotropic diffusion is used because it does not cause significant edge distortion. An example of an enhanced image is shown in Fig. 2.

3.3. Drawing of vocal tract grid lines

We use the vocal tract grid proposed by Bone, Proctor et al. [6,19,20] which is based on the methods of Öhman, Maeda, and Mermelstein [12–14]. The grid line configuration is motivated by the average shape of the human vocal tract: the vocal tract center-line can

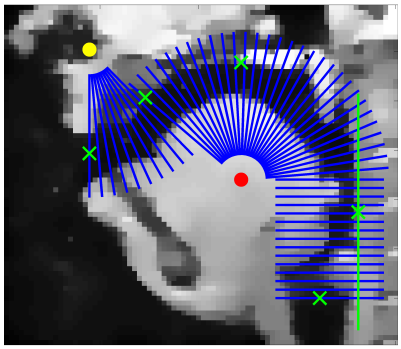

Fig. 3. *Drawing of the grid lines. The manually specified anatomical landmarks are shown with green crosses. The lingual center point is shown with a red bullet. The second center point is shown with a yellow bullet.*

be modeled with a vertical segment near the glottis, a circular segment centered at the tongue through the mid-oral vocal tract, and a circular segment centered above the lips from the alveolar ridge through the lips. The goal is to draw grid lines that are approximately normal to the vocal tract center-line. An example of grid line construction is shown in Fig. 3. The construction of the grid is done on the midsagittal slice and requires manual specification of the following anatomical landmarks: (1) the glottis, (2) a vertical line tangent to the posterior pharyngeal wall, (3) the highest point on the palatal contour, (4) the alveolar ridge, and (5) the middle point between the lips (Fig. 3). After manual placement of the anatomical landmarks, grid construction proceeds automatically.

A lingual center point is automatically placed on the tongue on the vertical line passing through the highest palatal point. The lingual center point is placed so that it is equidistant from the pharyngeal wall and the palate. Equidistant horizontal grid lines are drawn between the glottis and the lingual center point (Fig. 3). Equispaced radial grid lines centered at the lingual center point are drawn through the mid-oral vocal tract above the lingual center point and until the alveolar ridge (Fig. 3). A second center point is automatically placed at the intersection of the vertical line passing through the midlabial point and the line passing through the lingual center point and the alveolar ridge. Equispaced radial grid lines centered at the second center point are drawn though the anterior vocal tract from the alveolar ridge and until the midlabial point (Fig. 3).

We specify a fixed number of grid lines for each of the horizontal, mid-oral, and anterior-oral grid line groups, thus allowing the spacing between the lines to vary based on the anatomical characteristics of the subject. We draw 15 horizontal, 30 mid-oral, and 10 anterior-oral grid lines (55 grid lines in total).

3.4. Midsagittal vocal tract center-line estimation

We propose a method for vocal tract center-line estimation. This was not included in the previously proposed area function estimation tool by Kim et al. [6]. To estimate the vocal-tract center-line, we first segment the airway on the midsagittal slice. For this we require the manual specification (only on the midsagittal slice) of a seed inside the airway and a rough bound around the airway (Fig. 4). These can be quickly manually drawn as they do not need to be precise. The seeded region growing algorithm automatically segments the airway using the specified seed to yield an accurate midsagittal airway segmentation (Fig. 4). The manually specified bound around the airway is used to constrain the region growing to avoid leakage through the alveolar ridge and hard palate.

For each grid line we calculate the midpoint of its points of intersection with the boundary of the segmented airway. The collection of such midpoints on the grid lines yields our estimate of the vocal tract center-line (Fig. 4).

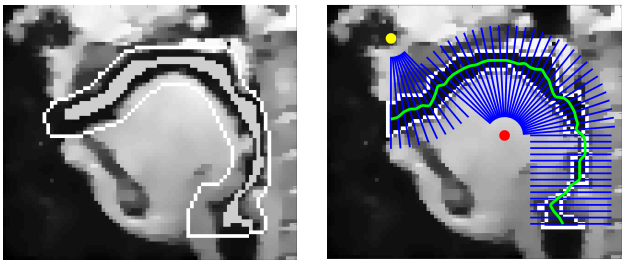

Fig. 4. *Center-line estimation. Left: Manually specified airway seed (gray) and outer bound (white). Right: Segmented airway boundary (white) and center-line (green).*

3.5. Slice cutting, airway segmentation, and area estimation

Horizontal grid lines are axially oriented so the corresponding slices can be readily extracted. Radial grid lines slice through the volume at a non-trivial angle. To extract the slice for such a grid line we rotate the volume with an affine 3D transformation [6] (using bicubic interpolation) until the grid line is coronally or axially oriented (we make the minimum possible rotation so grid line slices closer to vertical are rotated until coronally oriented and grid line slices closer to horizontal are rotated until axially oriented). Then, the corresponding slice can be readily extracted.

For each extracted slice, we use seeded region growing to segment the airway. We do not require manual specification of a seed for each slice, since we have the midsagittal airway segmentation available. We use the corresponding grid line profile of the segmented midsagittal airway to seed each extracted slice. This is an improvement in automation over the previously proposed method. The previous method required manual specification of seeds for each one of the extracted slices [6].

After automatic segmentation, we allow for manual inspection of the airway segmentation on each slice and manual correction if needed. We observe that manual correction is always required for the slices in the anterior-oral vocal tract, as the airway near the lips is not fully surrounded by tissue and region growing inevitably leaks to the background. On average 10 to 12 slices require manual correction for each MRI volume.

With the corrected airway cross-section segmentation we can readily estimate the cross-sectional airway areas by multiplying the number of pixels in the airway by the area of each pixel (0.125^2cm^2). The conjunction of distance from the glottis (calculated using the estimated midsagittal airway centerline) with the airway cross-sectional areas yields the estimate of the area function.

4. RESULTS AND DISCUSSION

Using the proposed method we estimated area functions for the vowels b**a**t, b**ee**t, b**oo**t, b**e**t, b**ou**ght for 2 female (W1, W2) and 2 male (M1, M2) subjects from our 3D MRI dataset. The results are shown in Figures 5 and 6.

To validate the resulting area functions, we synthesized vowels from them using the articulatory synthesis method proposed by Maeda [2]. We evaluated the synthesized vowels perceptually. We found that each synthesized vowel was reasonably close to the expected actual vowel.

The estimated area functions exhibit the expected shape: constriction locations appear at the expected location along the vocal tract for each vowel. For example, for the front vowel b**ee**t, the constriction along the palate is evident in the estimated area function for

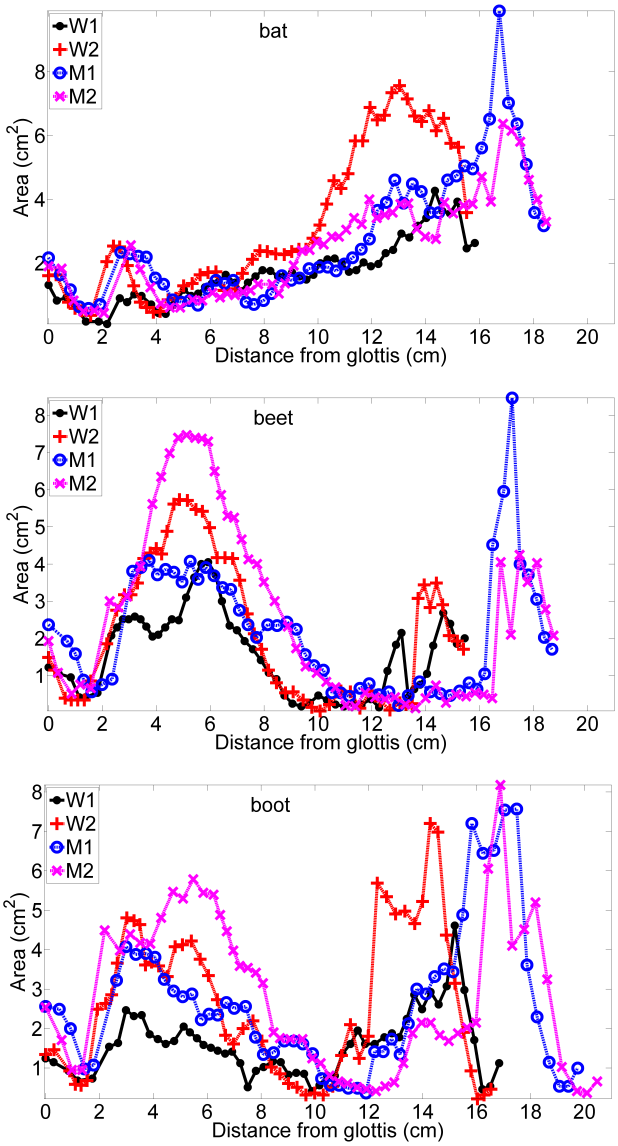

Fig. 5. Area function estimate for the 2 female (W) and 2 male (M) subjects for the vowels b**a**t, b**ee**t, and b**oo**t.

all subjects, while the pharynx is open yielding large cross-sectional areas. For the back vowel b**a**t, the cross sectional airway area is small at the back of the pharynx near the glottis and increases towards the lips. For the vowel b**oo**t, lip protrusion and lengthening of the vocal tract is observed for all subjects, along with labial constriction (observe the drop in estimated airway area near the lips). We also observe that the general trend of our estimated area functions for each sound is in agreement with the area functions previously reported [1].

Further, we may observe inter-subject anatomical differences. The most evident difference is vocal tract length: the male subjects considered have vocal tracts about 2cm longer than the female subjects. We also observe that the area functions for the longer vocal tracts appear to be shifted versions of those for the shorter ones; they have the same overall trend and relative constriction locations, as expected.

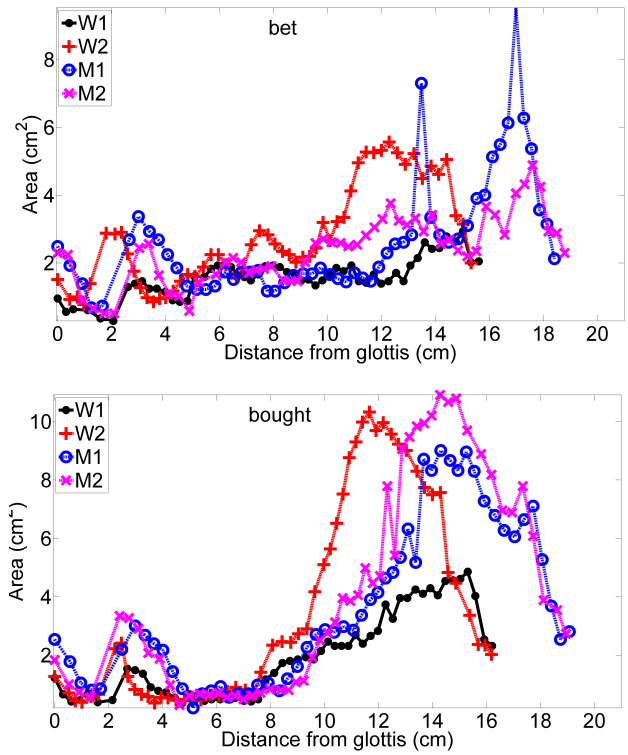

Fig. 6. Area function estimate for the 2 female (W) and 2 male (M) subjects for the vowels b**e**t and b**ou**ght.

5. CONCLUSIONS AND FUTURE WORK

We presented a semi-automatic method for area function estimation from 3D MRI data. Our method is based on the previously proposed method by Kim et al. [6]. We improved the automation of the previous method and added an airway center-line estimation facility. We demonstrated the utility of our proposed method by estimating area functions for 5 vowels for 4 subjects from a 3D MRI dataset collected with an accelerated protocol with 8s scan duration.

Our semi-automatic tool enables efficient area function estimation from 3D MRI data. We plan to estimate area functions for all recorded sounds from all 17 subjects in our dataset. This large 3D MRI dataset provides an unprecedented opportunity to capture a multitude of vocal tract shapes for many subjects. The estimated area functions can be used to address the sagittal-to-area conversion problem, for which previous studies consider only few subjects [1,21,22]. Having speaker specific conversion functions would enable accurate estimation of the dynamics of area functions from the dynamics of the midsagittal slice. Current MRI imaging techniques are not fast enough for real-time capture of the 3D dynamics of the vocal tract shape. However, 2D real-time MRI imaging techniques are available, and area function dynamics can be estimated by applying sagittal-to-area conversion on the midsagittal airway width measured from 2D real-time data [23,24]. Besides providing insight into 3D vocal tract dynamics, this information can also be used to improve articulatory synthesis [2].

6. ACKNOWLEDGEMENTS

This work was supported by NIH DC007124, NSF, and a USC Viterbi Graduate School PhD fellowship.

7. REFERENCES

[1] Brad H. Story, Ingo R. Titze, and Eric A. Hoffman, "Vocal tract area functions from magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 100, no. 1, pp. 537–554, 1996.

[2] Shinji Maeda, "A digital simulation method of the vocal-tract system," *Speech Communication*, vol. 1, no. 3, pp. 199–229, 1982.

[3] Man Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 7, pp. 955–967, 1987.

[4] Shrikanth S Narayanan, Abeer A Alwan, and Katherine Haker, "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I. The laterals," *The Journal of the Acoustical Society of America*, vol. 101, no. 2, pp. 1064–1077, 1997.

[5] Abeer Alwan, Shrikanth Narayanan, and Katherine Haker, "Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics," *The Journal of the Acoustical Society of America*, vol. 101, no. 2, pp. 1078–1089, 1997.

[6] Y.-C. Kim, J. Kim, M.I. Proctor, A. Toutios, K.S. Nayak, S. Lee, and S.S. Narayanan, "Toward automatic vocal tract area function estimation from accelerated three-dimensional magnetic resonance imaging," in *Proc. ISCA Workshop on Speech Production in Automatic Speech Recognition (SPASR)*, France, 2013.

[7] Yoon-Chul Kim, Shrikanth S. Narayanan, and Krishna S. Nayak, "Accelerated three-dimensional upper airway MRI using compressed sensing," *Magnetic Resonance in Medicine*, vol. 61, no. 6, pp. 1434–1440, 2009.

[8] A. Soquet, V. Lecuit, T. Metens, and D. Demolin, "Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with MRI," *Speech Communication*, vol. 36, no. 34, pp. 169–180, 2002.

[9] T. Baer, J. C. Gore, L. C. Gracco, and P. W. Nye, "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels," *The Journal of the Acoustical Society of America*, vol. 90, no. 2, pp. 799–828, 1991.

[10] Pascal Perrier, Louis-Jean Boë, and Rudolph Sock, "Vocal tract area function estimation from midsagittal dimensions with CT scans and a vocal tract castmodeling the transition with two sets of coefficients," *Journal of Speech, Language, and Hearing Research*, vol. 35, no. 1, pp. 53–67, 1992.

[11] Denis Beautemps, Pierre Badin, and Rafael Laboissiére, "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*, vol. 16, no. 1, pp. 27 – 47, 1995.

[12] Sven EG Öhman, "Numerical model of coarticulation," *The Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 310–320, 1967.

[13] Shinji Maeda, "Un modéle articulatoire de la langue avec des composantes lineaires," *Actes 10emes Journées dEtude sur la Parole*, pp. 152–162, 1979.

[14] P. Mermelstein, "Articulatory model for the study of speech production," *The Journal of the Acoustical Society of America*, vol. 53, no. 4, pp. 1070–1082, 1973.

[15] Zisis Iason Skordilis, Vikram Ramanarayanan, Louis Goldstein, and Shrikanth S Narayanan, "Experimental assessment of the tongue incompressibility hypothesis during speech production," in *Proc. Interspeech*, Dresden, Germany, September 2015.

[16] R. Adams and L. Bischof, "Seeded region growing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 6, pp. 641–647, Jun 1994.

[17] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[18] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 629–639, 1990.

[19] Daniel K Bone, Michael I Proctor, Yoon Kim, and Shrikanth S Narayanan, "Semi-automatic modeling of tongue surfaces using volumetric structural MRI," *The Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. 2549–2549, 2011.

[20] Michael I Proctor, Daniel Bone, Athanasios Katsamanis, and Shrikanth S Narayanan, "Rapid semi-automatic segmentation of real-time magnetic resonance images for parametric vocal tract analysis.," in *Proc. Interspeech*, Makuhari, Japan, 2010.

[21] Christine Ericsdotter, "Detail in vowel area functions," in *Proc. International Congress of Phonetic Sciences (ICPhS)*, Saarbrücken, Germany, 2007.

[22] Richard S McGowan, Michel TT Jackson, and Michael A Berger, "Analyses of vocal tract cross-distance to area mapping: An investigation of a set of vowel images," *The Journal of the Acoustical Society of America*, vol. 131, no. 1, pp. 424–434, 2012.

[23] Shrikanth Narayanan, Krishna Nayak, Sungbok Lee, Abhinav Sethy, and Dani Byrd, "An approach to real-time magnetic resonance imaging for speech production," *The Journal of the Acoustical Society of America*, vol. 115, no. 4, pp. 1771–1776, 2004.

[24] Erik Bresch, Yoon-Chul Kim, Krishna Nayak, Dani Byrd, and Shrikanth Narayanan, "Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 123–132, 2008.