Tongue Segmentation with Geometrically Constrained Snake Model

Zhihua Su[1], *Jianguo Wei*[1], *Qiang Fang*[2,*], *Jianrong Wang*[3], *Kiyoshi Honda*[3]

[1]School of Computer Software, Tianjin University, Tianjin, China
[2]Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China
[3]School of Computer Science, Tianjin University, Tianjin, China
*Corresponding to: fangqiang@cass.org.cn

Abstract

Articulatory visualization aims at providing precise visual information of the speech organs (tongue, lips, and velum) that accompany with speech signals. It is often critical in fundamental studies and certain applications. To construct an articulatory visualization system, the profile of the speech organs must be segmented from images acquired by various types of medical equipments. In this paper, a geometrically constrained snake model is proposed to segment tongue profiles from mid-sagittal MRI to deal with the situation in which the tongue contacts with the surrounding structures and the target object with inhomogeneity nature. The result indicates that the proposed method improves segmentation performance significantly compared with the traditional snake model.

Index Terms: articulatory visualization, tongue segmentation, geometrically constrained snake model

1. Introduction

Most of the speech apparatus are hidden in the oral cavity, which makes it difficult to observe the profiles and movements of the speech apparatus directly. Therefore, the visualization of vocal tract and articulators are of great interest in speech visualization and pathological speech analysis. Magnetic resonance imaging (MRI) is the state-of-the-art modality for obtaining the vocal tract and articulators observation, due to its noninvasive and non-hazardous nature. In literature, a number of articulatory models are constructed based on the manually tracked articulators'/vocal tract's profiles from MR images. However, manual annotation of speech organs (vocal tract) is cumbersome and apt to introduce artifacts in the extracted profiles. Hence, the segmentation of the articulators/vocal tract from MR images is a very important topic in articulation visualization and related area.

The tongue, is one of the main articulators. Its modeling helps to study the mechanism of speech production and rehabilitation of person with speech impairment. A number of works have been conducted to segment the tongue out of MRI images automatically. Bresch et al. [1] employed the snake model to automatically extract the midsagittal vocal tract outline from real-time magnetic resonance images (RT-MRI). Peng et al. [2] proposed a shape-based (where the shape priors were obtained by principal component analysis on a data set of 39 manually delineated tongue contour of a reference speaker) variational framework to curve evolution for the segmentation of tongue contours from mid-sagittal images. The final contour of the tongue was obtained by minimizing the total energy function that includes both global and local image statistics. Eryildirim and Berger [3] improved Peng's work [2] by incorporating physical constraints on the extremities of the tongue into the segmentation process. Proctor et al. [4] described a method of segmentation of RT-MRI data for geometric analysis of vocal tract, where the tissue-airway boundary was estimated based on an estimated vocal-tract midline and the intensity profile on each grid line. However, the tissue-airway boundary extracted by that method had obvious visual errors when it was blurred or even missing (tongue contacts other articulators during articulation). Hewer et al. [5] proposed a hybrid approach to extract a 3D tongue from 3D or 2D MRI scans of the vocal tract during the speech, which combines unsupervised image segmentation with a mesh deformation technique. Since the mesh deformation can be applied even with a sparse point cloud, Hewer's method was possible to extract realistic 3D tongue shapes even from the 2D video frames of real-time MRI. Krishna et al. [6] introduced a convolutional neural network with an encoder-decoder architecture to jointly detect the relevant air-tissue boundaries. A greedy search algorithm to draw contours.

Most of the above work deal with the tongue/vocal-tract profile in the midsagittal plane. However, the tongue is a flexible and active speech organ. It deforms and moves during speech production process. It frequently contacts with surrounding tissues when producing speech. In addition, the quality of acquired MRI image is usually not good, which means the boundary between the target object and background is blurred and the grayscale inside the target object is nonhomogeneous. The factors mentioned hereinbefore makes the task of tongue segmentation more difficult. Figure 1 gives an example of an MR image of articulation and the edge detection result by conventional gradient method. On the one hand, one can see that the gray scale of the region of interest is inhomogeneous. Several edges can be detected in the region of the tongue, and the magnitude of the gradient at some edges inside the tongue region is even stronger than the that of the target tongue profile edge. This makes it difficult to extract tongue contour with the conventional methods. On the other hand, tongue frequently contacts with surrounding soft speech organs (for example, the velum and the pharyngeal wall). When the tongue contacts with other tissues, the boundary between different speech organs is blurred. This makes it extremely challenging to segment tongue in this circumstance even for human annotator.

To tackle these issues, it is necessary to incorporate the anatomical knowledge of tongue into the tongue segmentation process. In this study, an improved snake model is adopted to segment the tongue from MR images, where the smoothness nature of the tongue is considered by the internal energy of the conventional model and the anatomical knowledge is taken

into account by incorporating geometrical constraints between tongue contour and anatomical landmarks.

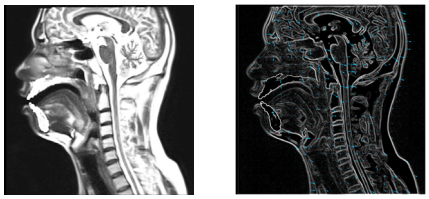

Figure 1: *The mid-sagittal MRI slice of vocal tract profile and the edge detection result*

2. Methods

2.1. Conventional snake model

SNAKE is a kind of model that defines a parametrical curve in an image domain, and evolves curve through minimizing a predefined energy function. The curve with the minimum energy is the ultimately extracted contour. The traditional SNAKE firstly defines [7] a curve as $\boldsymbol{v}(s) = (x(s), y(s)), s \in [0,1]$, and evolves contour through the spatial domain of an image to minimize the energy function:

$$E = \int_0^1 \frac{1}{2}\left[\alpha\left|\boldsymbol{v}'(s)\right|^2 + \beta\left|\boldsymbol{v}''(s)\right|^2\right] + E_{ext}(\boldsymbol{v}(s))ds \qquad (1)$$

where α and β are weighting parameters that control the snake's tension and rigidity, respectively, $|\boldsymbol{v}'(s)|$ and $|\boldsymbol{v}''(s)|$ denote the first and second derivative of $\boldsymbol{v}(s)$ with respect to s. The first term is called elastic energy and the second is bending energy. These two items are called internal forces, which are used to control the elastic deformation and maintain the continuity and smoothness of the contour. They are only related to the shape of the SNAKE rather than image. The external energy function E_{ext} is derived from the image and takes the smaller values at the positions of interest, such as edges. Given a grayscale image $I(x,y)$, viewed as a function of continuous position variables (x,y), a typical external energy that leads an active contour toward edges is designed as:

$$E_{ext}^1(x, y) = -\left|\nabla\left(G_{\sigma_1}(x, y) * I(x, y)\right)\right|^2 \qquad (2)$$

where $G_{\sigma_1}(x,y)$ is a two-dimensional Gaussian function with standard deviation σ_1, and ∇ is the gradient operator. It is easy to see that large σ_1 helps to increase the capture range of the external force, and its side effects are to blur the edges in the images. In real cases, a large σ_1 is often used.

2.2. Geometric constraint

To overcome the shortcomings of the conventional SNAKE, a new constraint, geometric constraint, is proposed and incorporated into the conventional SNAKE model. The geometric constraint is formulated as the distance between the contour and a set of geometric points (as shown in Eq. 3) [8].

$$d(x, y) = \prod_{i=1}^{n}\left(1 - e^{-\frac{(x-x_i)^2}{2\sigma_2^2}} e^{-\frac{(y-y_i)^2}{2\sigma_2^2}}\right), \forall (x, y) \in \Omega \qquad (3)$$

where (x_i, y_i) is the coordinate of the ith predefined landmark close to the boundary of the target object. By definition, in the whole image domain, $d(x,y)$ increases when the target contour deviates from the landmarks and decreases to zero when the target contour approximate the geometric landmarks. Therefore, this geometric constraint helps to attract the contour converge to target position globally. In the end, the external force is formulated as a combination of traditional gradient and geometric distance (as shown in Eq. 4).

$$E_{ext}^2 = -\gamma\left|\nabla\left(G_{\sigma_1}(x, y) * I(x, y)\right)\right|^2 + \delta d(x, y) \qquad (4)$$

where γ and δ are nonnegative weighting coefficients to balance these two terms. The model is able to find the contour by minimizing the energy E such that the boundaries of the region of interest are detected simultaneously around $d \approx 0$.

In some circumstances, the medical image contains both strong and weak edges, but the real boundary of the interest locates at one the weak edges. The conventional SNAKEs, such as the distance snake [9] and the GVF snake [10], always step through the weak edge and stop at the strong edge, and give undesired boundary. This issue can be largely alleviated by the proposed geometric constraint since the positions of the landmarks are given near the actual boundary. And the contour eventually converges to the neighborhood of these landmarks to approximate the real boundary even if the edge is weak.

3. Experiments

3.1. Data Description

Three datasets are used in our experiment to evaluate the performance of the proposed method. Dataset 1, 2 and 3 are MR image data recorded with different scanning parameters. The first dataset is MR image database of Japanese vowel production recorded at the Brain Activity Imaging Center, ATR-Promotions [11]. The MRI data were obtained using a Shimadzu Marconi MAGNEX ECLIPSE 1.5 T Power Drive 250 system. An atlas array coil was used for acquiring MRI data of the subject's head and neck regions. The imaging sequence was a sagittal fast spin echo series with 2.0 mm slice thickness, no slice gap, no averaging, 256×256 mm field of view (FOV), 512×512 pixel image size, 51 slices, 11 ms echo time (TE) and 3,000 ms repetition time (TR). During the scan, the subject was asked to repeat steady phonation 64 times for each vowel, which took approximately seven minutes.

The second dataset is recorded at the Beijing Normal University with a SIEMENS Trio A Tim 3T system. The parameters used in the MRI scans were as follows: 64 ms TE, 340 ms TR, 31 sagittal slice planes, 3 mm slice thickness, 3.6 mm slice interval, averaged once, 256×256 mm FOV, and 192×192 pixel image size. The rightmost and leftmost planes are located at 54 mm from the midsagittal plane. 36 Chinese vowels (9 vowels with 4 different tones) and 73 consonants in a symmetric VCV (vowel-consonants-vowel) sequence are acquired [12]. All the articulations were artificially sustained during the 10s acquisition time.

The third dataset is USC-TIMIT database [13] recorded for a subject M1 on a Signa Excite HD 1.5T scanner with gradients capable of the 40mT/m amplitude and 150mT/m/ms slew rate. The main parameters used in the MRI scans were TR= 6.164 ms, FOV=200×200 mm. Each frame shows the mid-sagittal slice of a single speaker and consists of 68×68 pixels with a pixel size of 2.9× 2.9 mm. Since the image size

is too small, the data actually used in the analysis is interpolated to 256×256 pixels with cubic spline interpolation.

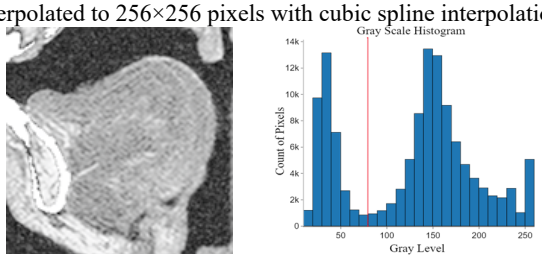

Figure 2: *Tongue image in dataset 1 and grayscale histogram*

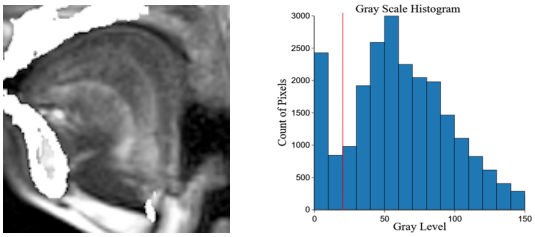

Figure 3: *Tongue image in dataset 2 and grayscale histogram*

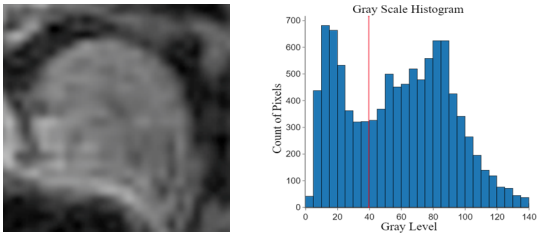

Figure 4: *Tongue image in dataset 3 and grayscale histogram*

The regions of interest of representative images in three datasets are shown in the left panels of Figure 2, Figure 3, and Figure 4, respectively. As shown in the left panels of the above figures, the qualities of the images of different data sets are quite different. The difference between tongue and airway is very clear, and the pixels in the tongue region is almost homogeneous. As shown in Figure 3, the pixels in the tongue region is inhomogeneous, which results in several edges the in the tongue region. Another issue is that the contrasts between background and tongue in images some dataset are difficult to discern (as shown in the left panel of Figure 3 and Figure 4). The corresponding histograms of pixels are shown in the right panels. The red lines indicate the mean gray scale of the pixels at the border of the target object (tongue). The latter two situations are very common in collected MR image datasets. These result in that the profile obtained by conventional methods is likely to converge to the unwanted position rather than the real boundary. Hence, it is necessary to consider these circumstances when conducting tongue segmentation.

3.2. Experimental setting

In order to assess the performance of the geometrically constrained SNAKE model, the following analyses were carried out on these three types of images. First, the region of interest of an MR image is selected. Then, geometric landmarks are specified at the places where the tongue contact with surrounding tissues or the curvature of the tongue boundary changes rapidly. The same initial contour of two models is obtained by interpolating between geometric landmarks. Next, the parameters of the energy function are tuned. Specifically, the initial parameter settings are manually selected based on the experience on each dataset. The parameters are adjusted one by one sequentially. When a parameter is under adjusting, the other parameters are fixed. To objectively evaluate the performance, the Dice similarity coefficient (DSC) [14], is adopted as the quantitative measure (as shown in Eq. 5).

$$DSC = \frac{2|G \cap I|}{|G| + |I|} \tag{5}$$

where G is the region enclosed by the ground truth profile, and I is the region of interest segmented out by the SNAKE. The DSC measures the amount of overlap between two segmentations. It equals 1 when the regions contained inside by the both contours coincide and 0 when they are completely different.

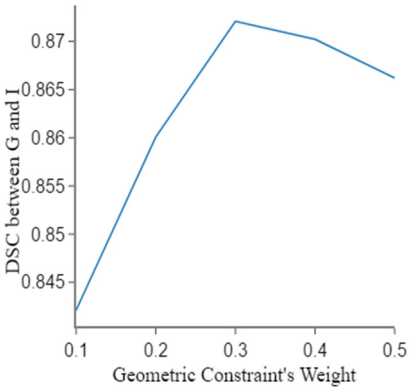

Figure 5: *DSC Line Chart*

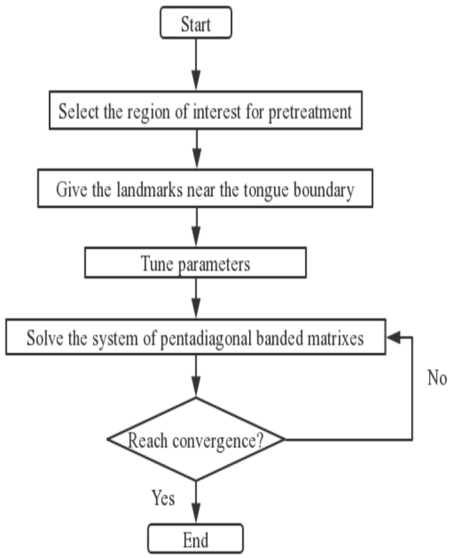

Figure 6: *Algorithm flowchart*

For example, to empirically determine appropriate value for the geometric constraint weight δ, the DSC is calculated for an image randomly selected from the second dataset (α = 0.1, β = 0.1, σ_1 = 8, σ_2 = 8, γ = 5). Figure 5 shows the performance of the proposed model by varying the parameter

δ. The best performance of the proposed model is obtained when δ equals 0.3.

When the parameters are determined, the Eq.4 is discretized. At last, two equations with pentadiagonal banded matrixes are constructed and solved iteratively [7]. The proposed contour segmentation process is summarized in Figure 6.

4. Result

The geometric landmarks are indicated by red circles in the left panels of Figure 7, Figure 8 and Figure 9. The results obtained by the conventional SNAKE (red contour) and the proposed SNAKE (green contour) are shown in the right panels.

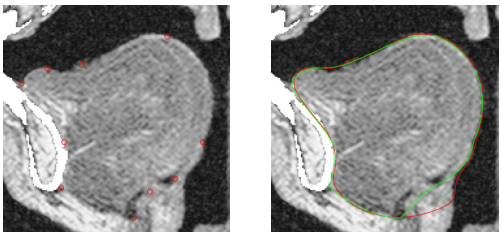

Figure 7: *Geometric landmarks on the image in dataset 1 and the obtained tongue contours*

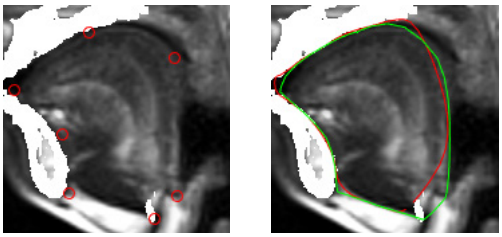

Figure 8: *Geometric landmarks on the image in dataset 2 and the obtained tongue contours*

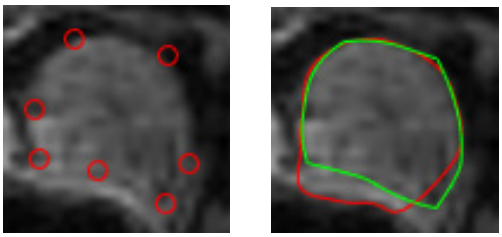

Figure 9: *Geometric landmarks on the image in dataset 3 and the obtained tongue contours*

One can see that the proposed model achieves better segmentation performance compared with the traditional one. Specifically, in Figure 7, the traditional SNAKE cannot distinguish the epiglottal cartilage from the tongue despite of the good image quality, while the geometrically constrained ones successfully delineate the boundary between tongue and epiglottis. In the right panel of Figure 8, at the anterior part, the traditional SNAKE converges to the boundary to between teeth and airway, rather than the boundary between the tongue tip and airway. At the posterior part, the tongue traditional snake converges to a relatively stronger edge in the tongue region rather than the boundary between the tongue root and the airway. The proposed model can successfully deal with these situations. Figure 9 illustrates that the traditional SNAKE cannot separate the tongue and the chin due to blurred boundaries and similar gray level of the pixels. However, this problem is effectively dealt with the geometrically constrained SNAKE model.

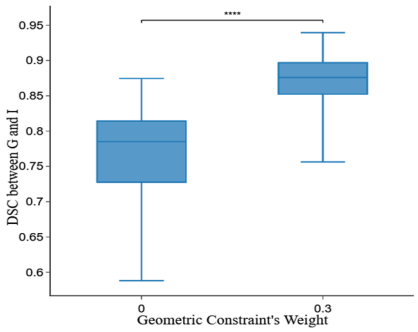

Figure 10: *Boxplot of DSC values with or without the geometric constraint*

A further quantitative measurement was conducted based on 50 images selected from the second dataset, respectively. The DSC values of the model with or without the geometric constraint are calculated for all the selected images. The results are shown in Figure 10. The T-test (p=0.05) justified that the difference of the performance between traditional SNAKE model (δ = 0) and geometrically constrained model (δ = 0.3) are statistically significant. Similar results are also obtained in the other two datasets. In summary, the proposed model has an obvious advantage over the conventional snake model.

5. Conclusions

This paper presents a geometrically constrained snake model to extract the tongue contour from MR images with various qualities. The proposed approach can effectively deal with the issues such as the difficulty of segmentation when the tongue is in contact with surrounding tissues and has complicated or blurred boundaries. In addition, the quantitative evaluation indicates that the proposed model outperforms the traditional SNAKE model in the sense of Dice similarity measure (DSC). In the future, other MRI slices rather than the mid-sagittal slices will be considered for three- dimensional tongue segmentation.

6. Acknowledgements

This work was supported by the National Natural Science-Foundation of China (No.61175016, No.61304250), Key Fund projects (No.61233009), Key Project of National Social Science Foundation of China (No.15ZDB103), and CASS Innovation Project “Articulatory model for pronunciation training”.

7. References

[1] Bresch, Erik, et al. "Semi-automatic processing of real-time MR image sequences for speech production studies." Proceedings of the 7th international seminar on speech production. 2006.

[2] Peng, Ting, Erwan Kerrien, and Marie-Odile Berger. "A shape-based framework to segmentation of tongue contours from MRI data." Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. IEEE, 2010.

[3] Eryildirim, Abdulkadir, and Marie-Odile Berger. "A guided approach for automatic segmentation and modeling of the vocal tract in MRI images." Signal Processing Conference, 2011 19th European. IEEE, 2011.
[4] Proctor, Michael I., et al. "Rapid semi-automatic segmentation of real-time magnetic resonance images for parametric vocal tract analysis." Eleventh Annual Conference of the International Speech Communication Association. 2010.
[5] Hewer, Alexander, Ingmar Steiner, and Stefanie Wuhrer. "A hybrid approach to 3D tongue modeling from vocal tract MRI using unsupervised image segmentation and mesh deformation." Fifteenth Annual Conference of the International Speech Communication Association. 2014.
[6] Somandepalli, Krishna, Asterios Toutios, and Shrikanth S. Narayanan. "Semantic Edge Detection for Tracking Vocal Tract Air-tissue Boundaries in Real-time Magnetic Resonance Images." Proc. Interspeech 2017 (2017): 631-635.
[7] Kass, Michael, Andrew Witkin, and Demetri Terzopoulos. "Snakes: Active contour models." International journal of computer vision 1.4 (1988): 321-331.
[8] Le Guyader, Carole, and Christian Gout. "Geodesic active contour under geometrical conditions: theory and 3D applications." Numerical algorithms 48.1-3 (2008): 105-133.
[9] Cohen, Laurent D., and Isaac Cohen. "Finite-element methods for active contour models and balloons for 2-D and 3-D images." IEEE Transactions on pattern analysis and machine intelligence 15.11 (1993): 1131-1147.
[10] Xu, Chenyang, and Jerry L. Prince. "Snakes, shapes, and gradient vector flow." IEEE Transactions on image processing7.3 (1998): 359-369.
[11] Kitamura, Tatsuya, et al. "Transfer functions of solid vocal-tract models constructed from ATR MRI database of Japanese vowel production." Acoustical science and technology 30.4 (2009): 288-296.
[12] Fang, Qiang, et al. "An Improved 3D Geometric Tongue Model." INTERSPEECH. 2016.
[13] Narayanan, Shrikanth, et al. "Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC)." The Journal of the Acoustical Society of America 136.3 (2014): 1307-1311.
[14] Popovic, Aleksandra, et al. "Statistical validation metric for accuracy assessment in medical image segmentation." International Journal of Computer Assisted Radiology and Surgery 2.3-4 (2007): 169-181.

[This paper was published at Interspeech 2018]