INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

An improved 3D geometric Tongue model

Qiang Fang[1], Yun Chen[2], Haibo Wang[3], Jianguo Wei[2], Jianrong Wang[4], Xiyu Wu[5], Aijun Li[1]

[1]Phonetics Lab., Institute of Linguistics, Chinese Academy of Social Sciences
[2] School of Computer Science, Tianjin University
[3]Institute of Ethnology and Anthropology, Chinese Academy of Social Science
[4]School of Computer Science, Tianjin University
[5]Peking University
fangqiang@cass.org.cn

Abstract

This study describes an improved geometric articulatory model based on MRI and CBCT（Cone Beam Computer Tomography）data. The basic idea is to improve the coherence of the vertices of tongue meshes so as to obtain more accurate tongue model. This is conducted in two aspects: i) The representative vertices of tongue surface are depicted in Cartesian coordinate system rather than in a semi-polar gridline coordinate system. ii) tongue surface meshes are modeled with reference to anatomical landmarks. Then, guided PCA is used to extract the control components based on MRI data. The average reconstruction error is less than 1.0 mm. Both qualitative and quantitative evaluation indicates that the proposed method surpasses the conventional semi-polar gridline system based method.

Index Terms—tongue model, Cartesian coordinate, anatomical landmark, guided PCA

1. Introduction

Articulatory speech synthesizer is promising for various studies and applications [1-3]. However, the quality of speech sound generated by articulatory synthesizers is degraded by various factors. Most articulatory synthesizers consist of three main modules: i) an articulatory model that imitates morphological structures of speech apparatus; ii) a coarticulation model that mimics the kinematic/dynamic behavior of speech apparatus; iii) and an acoustic model that simulates the aerodynamic process to generate corresponding speech signals. Any improper approximation in any of these modules is possible to deteriorate sound quality.

The current work describes an effort of constructing a more accurate tongue model to approximate the profile of tongue that is one of the most important articulator. This may lead to a better articulatory model for articulatory synthesis.

There are two modeling strategies for constructing tongue model: physiologic modeling, and geometric modeling. Physiological modeling implements finite element method (FEM) to simulate the biomechanical properties of soft tissues and embeds muscular structures to drive the FEM-based articulatory model [3, 4]. However, physiological articulatory models heavily depend on the anatomical and biomechanical properties of speech organ, which is not well understood at the current moment. Geometric modeling directly approximates the outline of the vocal tract or the surface of speech apparatus or by using rule-based or statistical-characteristic-based methods [5-7]. The shape of speech apparatus or vocal tract can be controlled by directly manipulating a set of predefined parameters of primitive geometric curves or parameters extracted from collected data. Unlike physiologic articulatory model, geometric articulatory model mimics neither the biomechanical properties of soft tissues nor the anatomical functions of the related muscular structure. In this way, it has a number of advantages, such as low computational cost and easy control of vocal tract shape. For above reasons, geometric articulatory model is the main stream for articulatory speech synthesis and speech animation.

In recent years, Birkholtz et al. constructed a 3D geometric articulatory model using geometric primitives [8]. However, articulatory models constructed in this manner depended heavily on pre-analysis of limited observations, and had high degree of freedom. These limitations make it difficult to ensure the validity of vocal tract shape generated using these models. Engwall and Badin implemented semi-polar coordinate systems to construct a 3D tongue model [7, 9, 10]. The advantage the semi-polar system is that it is easy to assess the coherence of short sections of vocal tract between different articulations. However, this degrades the coherence of the representative vertices of articulators and inevitably introduces noise into following statistical analysis, which will affect the precision of the constructed tongue model.

To overcome the shortcomings of 3D tongue models mentioned above, in this paper, an iterative method is used to construct a 3D geometric articulatory model in Cartesian coordinate system, and anatomic landmarks are used to improve coherence between representative vertices tongue.

2. Database

2.1. Data acquisition

The MRI data of a male subject were recorded using a SIEMENS Trio A Tim 3T system at Beijing Normal University. The parameters used in the MRI scans were as follows: 3 Tesla Magnetic Field Strength, 64 ms echo time (TE), 340 ms repetition time (TR), 31 sagittal slice planes, 3 mm slice thickness, 3.6 mm slice interval, averaged once, 256*256 mm field of view (FOV), and 192*192 pixel image size. The rightmost and leftmost planes are located at 54 mm from the mid-sagittal plane, respectively. During MRI data acquisition, the subject took supine position. The images of soft tissues (such as tongue, soft palate, pharyngeal wall) were obtained by using MRI (shown in Figure 1a). Since, the bony structures (teeth, jaw, and hard palate) were acquired by CBCT. The

http://dx.doi.org/10.21437/Interspeech.2016-901

CBCT data were translated and rotated to register with MRI images (as shown in Figure 1b).

We acquired 36 Chinese vowels (9 vowels with 4 different tones) and 73 consonants in symmetric VCV (vowel-consonants-vowel) sequence (as shown in Table 1). The VCV sequences were produced with a consonant, surrounded by vowels, e.g. [a]-[t]+[a]. The subject practiced on all the VCV sequences beforehand, to ensure that the vowel context specification was followed. All articulations were artificially sustained during the 10s acquisition time. For the consonants, the subject made the initial VC transition before the acquisition, then held the articulation while breathing out very slowly (for fricatives) or holding his breath (for stops) and finally made the CV transition after the scan. Finally, 104 articulations are deemed good enough to be retained in the corpus.

Table 1. *The vowels and pseudo-consonants list of Chinese*

Vowel	[a], [i], [ɿ], [ʅ], [u], [ɛ], [ɤ], [o], [y]
Fricative	[s] + [a], [i], [u]; [ʂ] + [a], [i], [u]; [ɕ] + [i], [y]; [f] + [a], [ɛ], [u], [o]; [x] + [a], [ɛ], [u];
Stop	[t] + [a], [i], [u], [ɛ]; [k] + [a], [i], [u], [ɛ]; [p] + [a], [i], [u], [o]; [pʰ] + [a], [i], [u], [o]; [tʰ] + [a], [i], [u], [ɛ]; [kʰ] + [a], [i], [u], [ɛ];
Affricate	[ʦ] + [a], [i], [u], [ɛ]; [tʂ] + [a], [i], [u], [ɛ]; [ʨ] + [i], [y]; [ʦʰ] + [a], [i], [u], [ɛ]; [tʂʰ] + [a], [i], [u], [ɛ]; [ʨʰ] + [i], [y];
Nasal	[m] + [a], [i], [u], [o], [ɛ]; [n] + [a], [i], [u];
Lateral	[l] + [a], [i], [u], [y], [ɤ];
Approximant	[ɾ] + [i];

2.2 Data annotation

The meshes of bony structures were annotated from the configuration of articulation /a/ of CBCT data, and registered with MRI images of other articulations by translation and rotation. As for the profiles of tongue, their shapes vary among different articulations, and should be annotated individually.

Two factors influence the accuracy of articulatory models. Firstly, in some articulation, tongue contacts with surrounding soft tissues. In the contact portion, it is extremely difficult to discern the boundary between tongue and surrounding structures, and makes it apt to mis-annotate the profile. Hence, both sagittal and corresponding axial slices were combined to help the discerning of the profile where tongue contacts with surrounding soft tissues (as shown in Figure 1b).

Another important factor is the coherence of vertices of tongue meshes among different articulations. This is largely overcome by using Cartesian coordinate to depict the position of vertices, and by introducing several anatomic landmarks (the yellow spots that denotes the tongue tip, tongue root, start and end position where tongue connects to jaw in the sagittal planes, and the lateral edge of tongue dorsum, as shown in Figure 1a and Figure 1b). With the help of these landmarks, the sagittal and transversal profiles were fused and resampled to form 3D tongue surface meshes. Hence, the tongue surface is divided into 3 different regions: dorsum, ventral, and floor. These three regions are modeled with different meshes, respectively.

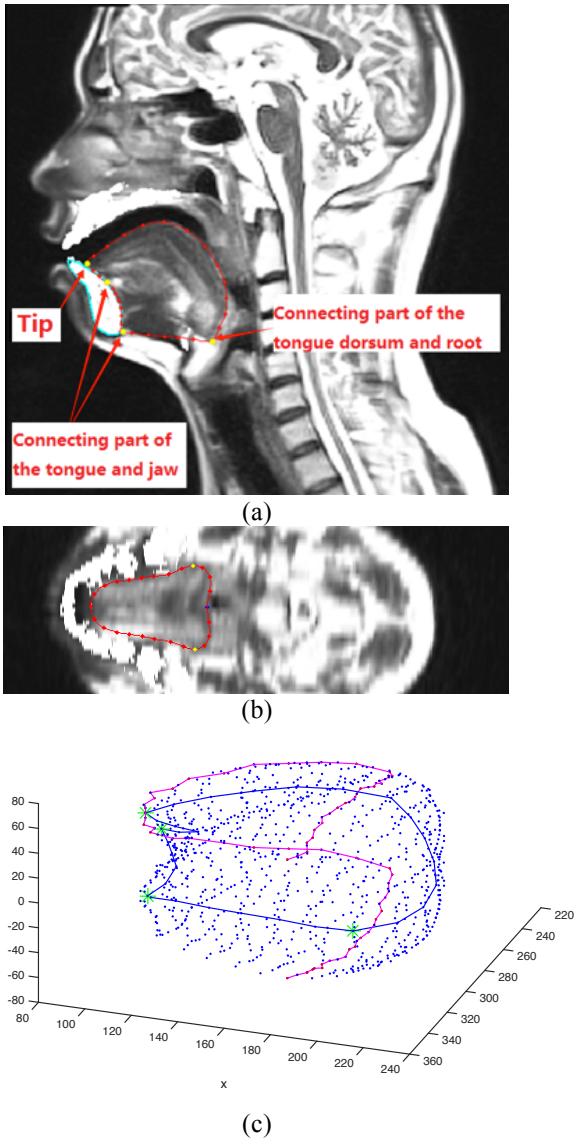

Figure 1*: (a) The mid-sagittal slice of combined MRI-CBCT vocal tract profile of articulation [a]. (b) An example of axial slices of vocal tract profile. (c) The annotated tongue surface points of articulation [a].*

3. The Tongue Mesh

Based on the annotated surface points and anatomic landmarks. An iterative algorithm (shown in Table 2) is implemented to resample the tongue surface with reference to the anatomical landmarks and produce the final tongue meshes.

Finally, dorsum surface consists of 9 left-right symmetric longitudinal fibers (form fiber 1 tow fiber 5 and fiber 15 to 18), which start from tongue tip and end at the tongue root. And 25 vertices evenly span on each tongue dorsum fiber. The 1st vertex corresponds to tongue tip, and the 25th vertex corresponds to tongue root. The ventral surface also consists of 9 left-right symmetric longitudinal fibers (from fiber 6 to fiber 14) with 25 vertices on each fiber. The ventral surface is divided

into two portions: one portion connected to jaw (from the 22th to 25th) that would not deform, and the other portion ((from the 1st to 22th) that deforms freely. The tongue floor consists of 18 fibers that start from the dorsum and ventral fibers and converge at the center of the tongue floor with 5 vertices span evenly on each floor fiber.

Table 2. *Algorithm for resampling the tongue surface.*

step 1:	resample the mid-sagittal dorsum curve to 25 points resample the mid-sagittal ventral to 25 points resample the dorsum edge curve to 25 point
step 2:	generate cutting plane according to corresponding points on mid-sagittal dorsum curve and edge curve generate cutting plane according to corresponding points on mid-sagittal dorsum curve and mid-sagittal ventral curve
step 3:	project the nearest annotated point to the plane, and create initial curves that connect corresponding resampled anatomical landmarks along tongue surface
step 4:	generate initial evenly span points on the curve created in previous step
step 5:	connect the corresponding points in longitudinal direction, and resample to make new points evenly distribute along the longitudinal direction
step 6:	connect the corresponding points in transversal direction, and resample to make new points evenly transversal along the longitudinal direction
step 7:	repeat step 5 and step 6, until the position of resampled points do not change

To check the validity of the tongue model, the corresponding volumes of tongue for different articulations are analyzed based on the resampled tongue mesh. The mean volume of the tongue is 105.102cm^3, the std. is 2.067cm^3, and the maximum deviation is 3.100cm^3. This is consistent with the hydro-elastic property of tongue, which makes the volume of tongue almost constant.

4. Tongue shape analysis

The guided-PCA method is used to extract a set of meaningful articulatory parameters that control the shape of tongue[5]. This method relates the shape of speech organ, s, to the average shape vector, $\mathbf{s_m}$, the transformation matrix, **M** whose columns each correspond to one basic shape vector, and the vector of loadings factors **a**.

$$\boldsymbol{s} = \boldsymbol{s}_m + \boldsymbol{Ma} \qquad (1)$$

The loading factors and basic shape vectors are determined in light of the following procedures: (i)The data describing the jaw movement is fed to PCA to extract the component for jaw movement. (ii)The influence of jaw movements on tongue is approximated and by using linear regression on the extracted jaw components. (iii) The active tongue movements are obtained by remove the approximated tongue movement from tongue data. (iv)The residue obtained in step 3 is fed to PCA to extract active tongue components.

Using the above method, a set of components are obtained for the purpose of controlling the shape of tongue and jaw. The physical meaning of extracted components is shown in Figure 2. The first two lines demonstrate the effects of jaw movement on tongue. The left three lines demonstrate the active movement of tongue itself. The 1st component JH (Figure 2a and Figure 2b) moves tongue upward-downward obliquely accompanied by tongue tip elevation-depression. The 2nd component JA (Figure 2c and Figure 2d) moves tongue forward-backward accompanied by tongue body depression. The 3rd component TA (Figure 2e and Figure 2f) moves tongue body forward and backward. The 4th component TT (Figure 2g and Figure 2h) controls the elevation of tongue tip accompanied by grooving at the posterior part of tongue dorsum. The 5th component TD (Figure 2i and Figure 2j) controls the benching-flattening of tongue dorsum accompanied by grooving at the anterior of tongue dorsum.

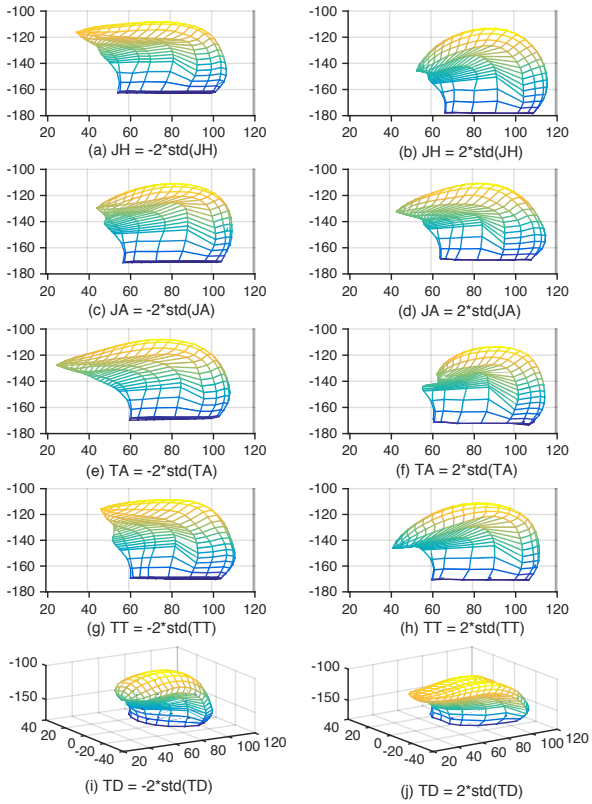

Figure 2: *Physical meaning of the extracted components for tongue movements.*

The components JH and JA explained 96.39% of the variance of jaw movement. Table 2 shows the variance of tongue explained by the extracted components. Components JH, JA, TA, TT, and TD explain 6.7%, 5.7%, 47.7%, 16%, and 8.8% of the variance of tongue, respectively.

Table 3. *The Variance of tongue explained by extracted components.*

Component	Explained variance
JH	6.7%
JA	5.9%
TA	47.7%
TT	16%
TD	8.8%
Total	85.1%

The RMSE of a specific vertex is calculated using the following equation:

$$RMSE = \sqrt{\frac{1}{N_s}\sum_{i=1}^{N_s} \| v_r^i - v^i \|^2} \qquad (2)$$

where N_s is the number of the tongue shape samples, $\mathbf{v_r}$ is the coordinates of a reconstructed tongue vertex, and $\mathbf{v}$ is that of the corresponding original tongue vertex. The mean RMSE of reconstructed jaw was 0.05cm with the standard deviation of 0.03cm. The RMSE of reconstructed tongue was 0.09cm with the standard deviation of 0.05cm.

Moreover, the reconstruction errors of the vertices on dorsum fibers are analyzed. Since the tongue model is left-right symmetric, only the errors of the vertices on fiber 1 to fiber 5 are analyzed. The reconstruction error of a specific tongue vertex is computed by using Eq. 3.

$$err_v = \sqrt{\| v_r - v \|^2} \qquad (3)$$

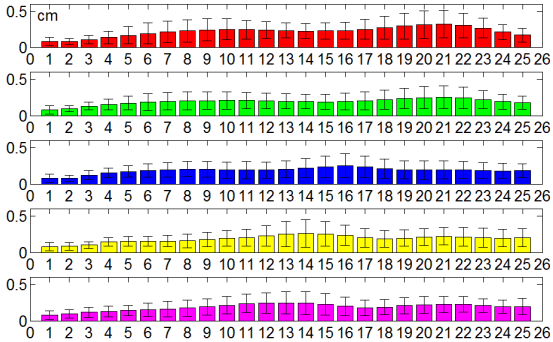

Figure 3: *Mean and standard of reconstruction error of the vertices on dorsum fibers 1-5.*

The results are shown in Figure 3. The x-axis is the vertex index along dorsum fibers from tongue tip to bottom (1 indicates tongue tip, and 25 indicates the tongue bottom). The top panel demonstrates the mean reconstruction error of vertices on the mid-sagittal fiber (fiber 1). The bottom panel demonstrates the mean reconstruction error of the vertices on the edge fiber (fiber 5). And the panels in between demonstrate the error of vertices on the fiber 2 to 4. For all the dorsum fibers, error increases gradually from tongue tip to dorsum. And there is no significant difference between the errors of vertices on difference dorsum fibers. The largest error occurs at the posterior part of tongue dorsum while produce articulation [s] and [ɛ] whose crucial articulator are both tongue apexes.

Table 4. *Comparison of the accuracy of these two models*

	Accumulating contribution rate	RMSE	Std.	Max error
Proposed	85.1%	0.09	0.05	0.36
Engwall	85.4%	0.13	0.09	0.75

To evaluate the accuracy of the proposed method, we reconstructed the tongue model according to semi-polar coordinate system of KTH tongue model. Then, we compared the precision of these two models. As shown in Table 4, the reconstruction error for the tongue in the current method is 0.09cm with the standard variation of 0.05cm, both of which are lower than the results of the conventional method (0.13 cm and 0.09 cm), respectively.

5. Conclusion

This paper attempts to improve the precision of a 3D tongue model by improve the coherence of vertices of tongue models from two aspects: i) in annotation phase, images from sagittal and axial perspective were combined to help the discerning of unclear profile of tongue, and anatomical landmarks were explicitly labeled; ii) in modeling phase, an iterative method is applied to model different anatomical regions of tongue surface. After that, the conventional guided-PCA is applied to extract the control component of the tongue model, and serve as the basis for evaluating the reconstruction error of the tongue model. The reconstruction error between tongue model and real data was less than 1.0 mm. This indicates that, with the extracted components, various positions and deformations of tongue with could be generated with high accuracy. In addition, quantitative results indicated that the proposed method surpassed the conventional methods.

6. Acknowledgements

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

7. References

[1]. Fang, Q., et al. Investigation of the functional relationship of tongue muscles for the control of a physiological articulatory model. *in The 8th national conference of Phonetics. 2008. Beijing, China.*

[2]. Fang, Q., A. Nishikido, and J. Dang, Feedforward Control of a 3D Physiological Articulatory Model for Vowel Production. *TSINGHUA SCIENCE AND TECHNOLOGY, 2009. 14(5).*

[3]. Dang, J. and K. Honda, Construction and control of a physiological articulatory model. *J. Acoust. Soc. Am., 2004. 115(2): p. 853-870.*

[4]. Perrier, P., L. Ma, and Y. Payan. Modeling the production of VCV sequences via the inversion if a biomechanical model of the tongue. *in INTERSPEECH 2005. 2005. Lisbon, Portugal.*

[5]. Maeda, S., Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory mode. *Speech production and modeling. 1990: Kluwer Academic Publishers.*

[6]. Badin, P. and A. Serrurier, Three-dimensional modeling of speech organs: Articulatory data and models. *Transactions on Technical Committee of Psychological and Physiological Acoustics, 2006. 36 5 (H-2006-77): p. 421-426.*

[7]. Engwall, O., Combining MRI, EMA and EPG measurements in a three-dimensional tongue model. *Speech Communication, 2003. 41: p. 303-329.*

[8]. Birkholz, P., D. Jackèl, and B.J. Kröger, Construction and control of a three-dimensional vocal tract model, *in ICASSP. 2006. p. 873-876.*

[9]. Badin, P., et al., A three dimensional linear articulatory model based on MRI data, *in The 3rd ESCA/COCOSDA International Workshop on Speech Synthesis. 1998. p. 249-254.*

[10]. Badin, P., et al., Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. *Journal of Phonetics, 2002. 30(3): p. 533-553.*