A NONLINEAR 3D GEOMETRIC TONGUE MODEL

Qiang Fang[1], Hequn Li[2], Jianguo Wei[2], Jianrong Wang[3], Xiyu Wu[4]

1. Institute of Linguistics, Chinese Academy of Social Sciences
2. School of Software, Tianjin University
3. School of Computer Science, Tianjin University
4. Peking University

ABSTRACT

This study describes a nonlinear geometric tongue model based on MRI and Cone-beam CT (CBCT) data. Comparing with the conventional geometric tongue model, the proposed tongue model is controlled by several prototype vertices, and the relationship between tongue mesh vertices and prototype vertices are modeled with quadratic functions. The results indicate that: i) quadratic models do improve the reconstruction performance of tongue mesh, especially in the tongue root region; ii) the quadratic model which use the cross-prototype-vertex information achieves the best performance of tongue mesh reconstruction; iii) the reconstruction performance can be further improved if an extra prototype vertex TP in the tongue root region is taken into account, even if TP is estimated from the measured prototype vertices.

Index Terms— 3D tongue model, anatomical landmark, nonlinear modeling

1. INTRODUCTION

Articulatory speech synthesizer is promising for various studies and applications. 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 these modules is possible to deteriorate sound quality synthesized by a articulatory synthesizer. Thus, accurate articulatory modeling is one of the important issues for articulatory synthesis and far from being resolved.

Geometric modeling is one of the important articulatory modeling techniques. It directly approximates the outline of the vocal tract or the surface of speech apparatus by using rule-based or statistic-based methods [1-3]. The shape of speech apparatus or vocal tract can be controlled by directly manipulating a set of predefined parameters of primitive geometric curves [4, 5] or factors extracted from collected data [1, 3, 6]. Most of the factor-based models are based on statistical analysis of the profiles of speech organ/vocal-tract obtained from static articulations. The articulators' kinematic information is usually obtained by EMA or ultrasound, which give kinematic information of part of the tongue only. To drive articulatory models generate continuous movements, the control parameters/factors should be estimated from kinematic data first, then the profile of speech organs is reconstructed from the estimated control parameters[3, 7]. This makes an inconsistent strategy to make continuous moveable articulatory models. In addition, our previous work[8], where a 3D tongue model is constructed with guide-PCA method, found that the reconstruction error was large in the region of tongue root. This indicates that linear model can't achieve good performance in the tongue root region. Hence, in this study, we attempt to construct a 3D tongue model which can be driven by several prototype vertices and more precise than previous linear tongue models.

In literature, several studies have been conducted to predict midsagittal tongue contour from the coordinates of several flesh points. Kaburgi et al. [9] applied a multivariable linear regression model to estimate the shape of midsagittal tongue contour from the position of coils attached to tongue surface based on a simultaneously measured database of EMA and ultrasound data. They found that the tongue contour was estimated (from four positions on the tongue) with an average estimation error of 1.24 mm. And the estimation error could be reduced to 0.84 mm when there was no measurement error between EMA and the ultrasonic data. They also found that the number of data frames for calculating the regression coefficients could be reduced, while maintaining the estimation accuracy by appropriately selecting data frames. Qin et al [10] proposed a radial basis function network to predict the midsagittal tongue contour from the locations of a few landmarks (metal pellets) on the tongue surface. They found that 3–4 landmarks are enough to achieve 0.3–0.2 mm error per point on the tongue. All those findings suggest that part of the 2D midsagittal tongue contours can be estimated from several flesh points on tongue surface. Nonetheless, no direct evidence shows that whole 3D tongue shape can be reconstructed from the coordinates of several flesh points.

In this study, we attempt to make a tongue model whose mesh can be reconstructed from several prototype vertices that are coils attached to tongue surface and jaw in EMA experiment. And the relationship between tongue mesh vertices and prototype vertices are modeled with nonlinear

functions. The advantage is that the proposed tongue model can be driven by measured data directly, and the reconstruction performance of the tongue mesh could be improved in comparison with linear models.

2. ARTICULATORY DATA

2.1 The tongue mesh

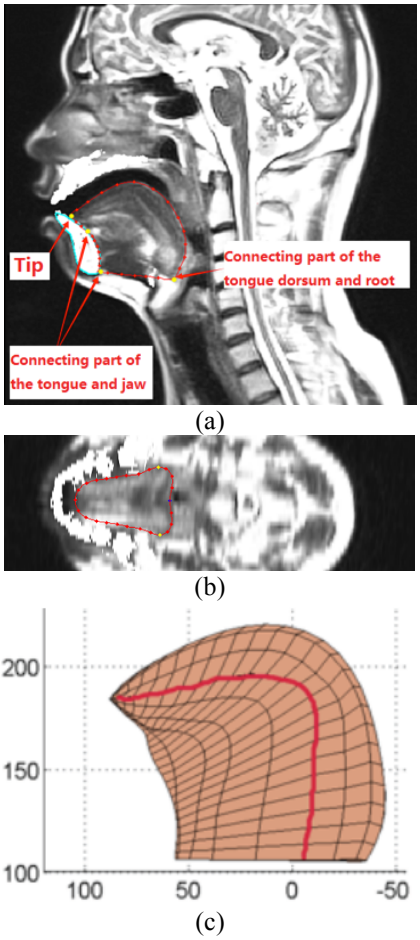

Figure 1. (a) The mid-sagittal slice of combined MRI-CBCT vocal tract profile of articulation [a]. (b) An example of axial slices of vocal tract profile. (c) The mean tongue surface mesh.

We acquired the articulation volumes of 36 Chinese vowels (9 vowels with 4 different tones) and 73 consonants in symmetric VCV (vowel-consonants-vowel) sequences by fusing MRI and CBCT images [8]. 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 is 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 hold 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.

The MRI data are annotated 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 three different regions: dorsum, ventral, and floor. These three regions are modeled with different meshes, respectively.

Finally, dorsum surface consists of 9 left-right symmetric longitudinal fibers (from fiber 1 to 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. In addition, 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.

To check the validity of the tongue mesh, 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 hypothesis of tongue volume.

2.2 EMA data

The same subject participated in a EMA experiment, where the NDI Wave system was employed to record acoustic signal and articulators' position simultaneously. 1108 phonetically balanced Chinese sentences in total were selected to serve as the recording prompts. In the EMA experiment, coils were glued to Tongue Rear (TR), Tongue Dorsum (TD), Tongue Blade (TB), Tongue Tip (TT), Lower Incisor (LI), Lower Lip (LL), and Upper Lip (UL) in the mid-sagittal plane. Another two coils were attached to the ridge of nose to serve as references (as shown in Figure. 2(a)). As a result, we could easily extract the global rigid body motion associated with head's movement. The sampling frequencies were 16 kHz for acoustic signal and 100 Hz for articulatory signal, respectively. A third-order Savitzky-Golay filter with the frame size of 21 was applied to smooth the trajectory of coils to suppress their jittery motions. Finally, the EMA data were aligned to the MRI image by translation and rotation with reference to the position of reference coils.

Since we aim to reconstruct the tongue mesh and drive the tongue realize continuous movement by using the data measured by EMA coil directly, we should define the correspondence between prototype tongue vertices and the measured EMA coils. To determine the vertices on tongue

surface which correspond to EMA coils, the mean of the tongue shape and the mean of coils' position are calculated, respectively. Then, the distances between TT and TB, TB and TD, TD and TR are calculated. At last, we correspond TT to the tongue tip vertex (the first vertex along the midsagittal plane), and the other 3 vertices along the tongue surface in the midsagittal plane according to the distance between TT and TB, TB and TD, and TD and TR. Fortunately, TB, TD and TR happen to be the 4th, 8th, and 10th vertex in the midsagittal plane, respectively.

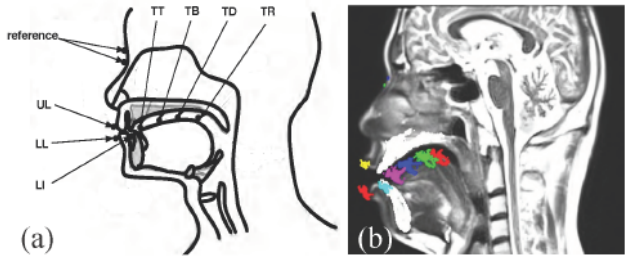

Figure 2. (a) EMA sensors' placement and (b) the aligned MRI-CBCT-EMA volume.

3. TONGUE MESH RECONSTRUCTION

3.1 The quadratic reconstruction model

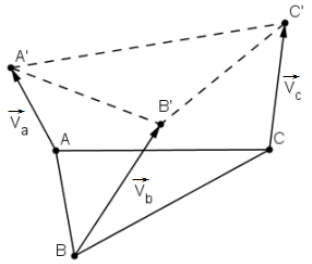

Figure 3. Illustration of linear component analysis [3].

To reconstruct the tongue mesh from a set of prototype vertices, we adopt the idea of Engwall [3], where the tongue mesh is estimated from the position of several prototype vertices. The basic idea is to estimate the position of tongue mesh vertices based on the displacement of prototype vertices. For example, as shown in Figure 3, vertex B is defined as the prototype vertex, then the displacement of vertex A and C can be estimate from that of vertex B by some function. In Engwall's work [3], they use a linear function. Due to the limitation of linear model, the reconstruction performance still need to be improved.

In our previous work, we found that the reconstructed tongue surface with linear model had large error in tongue root region. In this study, we resort to a quadratic function, which describe the relationship between tongue mesh vertices and prototype vertices, to further improve the performance of tongue mesh reconstruction in tongue root region. In the quadratic reconstruction model, the x-coordinate of the ith vertex can be estimated by using the following equations:

$$\Delta x_i = \boldsymbol{w}_{x_i}^T \boldsymbol{f} + b_{x_i} \quad (1)$$

$$x_i = \Delta x_i + x_{ref,i} \quad (2)$$

$$\boldsymbol{f} = (\boldsymbol{f}_1^T, \dots, \boldsymbol{f}_k^T, \dots, \boldsymbol{f}_N^T)^T \quad (3)$$

$$\boldsymbol{w}_{x_i} = (\boldsymbol{w}_{x_i1}^T, \dots, \boldsymbol{w}_{x_ik}^T, \dots, \boldsymbol{w}_{x_iN}^T)^T \quad (4)$$

where N is the number of prototype vertices, Δx_i is the displacement of the i^{th} vertex in x-direction, $x_{ref,i}$ is the x-coordinate of reference position of the i^{th} vertex, $\boldsymbol{f}_k$ is the feature vector derived from the displacement of the kth prototype vertex, $\boldsymbol{w}_{x_ik}$ is the corresponding weight vector associated with feature $\boldsymbol{f}_k$ for x-coordinate of the i^{th} tongue vertex. The y-coordinate and z-coordinate of the i^{th} vertex can be estimated in a similar way.

3.2 Features in quadratic model

The feature vector $\boldsymbol{f}$, of course, depends on the identities of the prototype vertices and how $\boldsymbol{f}_k$ is defined. In this study, we explore the effects of 7 types of prototype vertices combinations ([TT, TB, TD], [TT, TB, TR], [TT, TD, TR], [TT, TB, TD, TR], [LI, TT, TB, TD], [LI TT, TB, TR], [LI, TT, TD, TR]) and three types of feature definition of the i^{th} prototype vertex (Quadratic1: $[x_i^2, y_i^2, x_i, y_i]$; Quadratic2: $[x_i^2, y_i^2, x_i y_i, x_i, y_i]$; Quadratic3: $[x_i^2, y_i^2, \dots, x_i x_j, \dots, x_i y_j, \dots, y_i x_j, \dots, y_i y_j, \dots, x_i, y_i]$; $1 \le j \le N$).

In the linear model, there exists large error in tongue root region. Therefore, we also investigate whether adding an extra prototype vertex in tongue root region improves the reconstruction performance. Since coil can't be glued to tongue in that region in the EMA experiment, the position of the prototype vertex in tongue root region, TP, is estimated from the above prototype vertices combinations first, then TP is appended to the end of the corresponding prototype vertices combinations to form a new prototype vertices combination.

In order to determine the optimal prototype vertex TP, we make a preliminary investigation on the reconstruction performance of the vertex on fiber 1 between the 12th and 25th vertex in different prototype vertex combination. The results show that the reconstruction errors are not significantly different when the TP is chosen between 12th and 25th tongue vertex on fiber 1 for each prototype vertices combination. Therefore, the 20th vertex is chosen as the prototype tongue vertex TP.

4. RESULTS

To evaluate the performance of the proposed 3D tongue model, the reconstruction error of all the tongue mesh vertices are calculated by using Eq. 5.

$$err_v = \sqrt{\| v_r - v \|^2} \quad (5)$$

4.1 General reconstruction performance

Table 1 gives the results obtained with different prototype vertices combination and different quadratic feature definition.

If we look at Table 1 in column direction, we can found that: i) the more prototype vertices we use in the reconstruction model, the better performance we obtain, ii)

the reconstruction performance is better in the situation that prototype vertex of jaw (LI) is involved than in the situation that only prototype vertices of tongue is consider. And T-test analysis indicates that the difference between the reconstruction error in 3-prototype-vertex models ([TT, TB, TD], [TT, TB, TR], [TT, TD, TR]) and 4-prototype-vertex models ([TT, TB, TD, TR], [LI, TT, TB, TD], [LI TT, TB, TR], [LI, TT, TD, TR]), model with Jaw prototype vertex ([LI, TT, TB, TD], [LI TT, TB, TR], [LI, TT, TD, TR]) and without ([TT, TB, TD, TR]) Jaw prototype vertex and models are statistically significant.

Table 1. Mean reconstruction error of the tongue mesh vertices (Quadratic#+pre means the position of TP is estimated and TP is concatenated to the corresponding prototype combinations).

	TT TB TD	TT TB TR	TT TD TR	TT TB TD TR	TT TB TD LI	TT TB TR LI	TT TD TR LI
LCA	1.9	1.8	1.8	1.7	1.3	1.3	1.3
Quadratic1	1.6	1.5	1.5	1.4	1.2	1.2	1.2
Quadratic1+pre	1.5	1.5	1.5	1.4	1.2	1.1	1.1
Quadratic2	1.5	1.5	1.5	1.4	1.1	1.1	1.1
Quadratic2+pre	1.4	1.4	1.4	1.3	1.1	1.1	1.1
Quadratic3	1.3	1.3	1.3	1.1	0.9	0.8	0.9
Quadratic3+pre	1.0	1.1	1.1	0.7	0.6	0.6	0.6

If we look at Table 1 in row direction, we can find that: i) the reconstruction performance of quadratic model is better than that of linear model, and the difference are statistically significant; ii) the performance of quadratic models is similar if the feature vector of each prototype vertex is derived from its own information alone; iii) the performance is significantly improved if the feature vector of each prototype vertex is derived from the information of both its own and other prototype vertices; iv) the performance improvement of model Quadratic1 and Quadratic2 are not significant when the position of TP, estimated from corresponding prototype vertices combinations, is used; v) the performance of model Quadratic3 is improved significantly when the position of TP, estimated from corresponding prototype vertices combinations, is used. This indicates that prototype tongue vertex information are useful only if a proper model is selected.

4.2 Tongue surface reconstruction performance

Figure 4 and Figure 5 illustrate the reconstruction performance of 4-prototype-vertex models ([LI, TT, TB, TR]) for the tongue vertices on midsagittal contour (Fiber 1) and on the tongue surface edge (Fiber 5). The results indicate that: i) the reconstruction error from 1st vertex to 10th vertex is very small, while the error grows rapidly from the 11th to the 25th on the midsagittal tongue surface curve; ii) quadratic models significantly reduce the reconstruction error in tongue root region; iii) the reconstruction error of the vertices on the lateral side are relative larger than that of the vertices on the midsagittal curve; iv) the same trend is found as in the analysis of general reconstruction error: LCA gives the worst performance, the quadratic model which uses cross-prototype-vertex feature gives the best performance, and the quadratic models which do not use cross-prototype-vertex feature achieve performance in between.

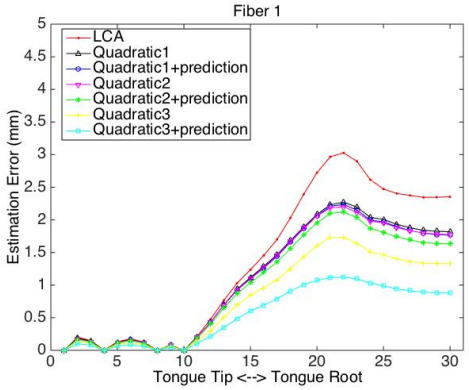

Figure 4. Mean reconstruction error of Fiber 1

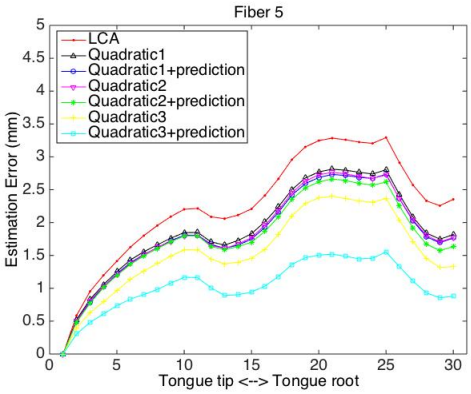

Figure 5. Mean reconstruction error of Fiber 5.

5. CONCLUSION

In this study, we construct a quadratic 3D tongue model. In the model, the tongue mesh is estimated from several prototype vertices that correspond to some of the EMA coils in EMA experiments. The results show that: i) quadratic models do improve the performance of the tongue model, especially in tongue root region; ii) the quadratic model which use the cross-prototype-vertex information achieves the best performance for tongue mesh reconstruction; iii) the performance can be further improved if we take a prototype vertex TP in tongue root region into account, although TP is estimated from the measured prototype vertices.

6. ACKNOWLEDGEMENT

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. Maeda, S., *Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory mode. Speech production and modeling*. Speech production and modeling. 1990: Kluwer Academic Publishers.
2. 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.
3. Engwall, O., *Combining MRI, EMA and EPG measurements in a three-dimensional tongue model.* Speech Communication, 2003. 41: p. 303-329.
4. Mermelstein, P., *Articulatory model for the study of speech production.* J. Acoust. Soc. Am., 1973. 53: p. 1070-1082.
5. Birkholz, P., D. Jackèl, and B.J. Kröger, *Construction and control of a three-dimensional vocal tract model*, in *ICASSP*. 2006. p. 873-876.
6. Badin, P., et al., *A threedimensional linear articulatory model based on MRI data*, in *The 3rd ESCA/COCOSDA International Workshop on Speech Synthesis*. 1998. p. 249-254.
7. Badin, P., E. Baricchi, and A. Vilain, *Determining tongue articulation: from discrete fleshpoints to continuous shadow*, in *EuroSpeech*. 1997. p. 47-50.
8. Fang, Q., et al., *An improved 3D geometic tongue model*, in *InterSpeech*2016. p. 1104-1107.
9. Kaburagi, T. and M. Honda, *Determination of sagittal tongue shape from the positions of points on the tongue surface.* J. Acoust. Soc. Am., 1994. 96(3): p. 1356-1366.
10. Qin, C., et al. *Predicting tongue shapes from a few landmark locations*. in *Interspeech*2008.