MRI-based 3D Model of Spoken Lips

Yun Chen[1, a], Qiang Fang [2, b], Wenhuan Lu[3, *,c], Jianguo Wei[1, 3, d]

[1] School of Computer Science and Technology, Tianjin University, Tianjin, 300072, China

[2]Phonetics Lab., Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, 100732, China

[3]School of Computer Software, Tianjin University, Tianjin, 300072, China

[a]email: 381800185@qq.com, [b]email:fangqiang@cass.org.cn,

[c]email:wenhuan@tju.edu.cn, [d]email: jianguo@tju.edu.cn

Keywords: 3-Dimensional Articulatory Model; Control Parameter; Linear Component Analysis

Abstract. Modeling accurate speech organs are vital to speech synthesis. This paper introduces a construction method for geometric model of spoken lips based on a MRI database. Volume images of articulators were acquired by MRI on one Chinese subject uttering a corpus of sustained articulations. Then 3D models of each organ were annotated from relevant image slices. Finally linear component analysis was used to extract main control parameters from the 3D models of the spoken lips. The result shows that these control parameters could effectively represent the position and motion of spoken lips with a minor reconstruction error (less than 0.15cm).

Introduction

Articulatory speech synthesizer is promising for various studies and applications [1] [2] [3]. There are two common modeling strategies for constructing articulators' 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 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 present. Geometric modeling directly approximates the outline of the vocal tract and the surface of speech apparatuses using rule-based or statistical-characteristic-based methods [5] [6] [7]. The shape of speech apparatuses and the vocal tract can be driven directly by manipulating a set of predefined parameters which is gained from statistical data. Recently, Birkholtz et al. constructed a 3D geometric articulatory model using geometric primitives [8]. However, articulatory models constructed in this method heavily depended on pre-analysis of limited observations, and had a high degree of freedom. Engwall and Badin implemented semi-polar coordinate systems to construct a 3D tongue model [7] [9] [10]. The advantage of the semi-polar system is easily assessing coherence of the short sections of vocal tract among different articulations. However, this method will degrade the coherence of the representative vertices of articulators and inevitably introduce noise into following statistical analysis.

In this paper, we proposed a geometric modeling method in Cartesian coordinate system based on a Chinese MRI database. Spoken lips were modeled separately and control parameters of the lips were extracted by Linear Component Analysis (LCA). These parameters are proved of clear physical significance.

Dataset

The MRI data of a male subject were recorded using a SIEMENS Trio A Tim 3T system .We acquired 36 Chinese vowels(9 vowels with 4 different tones) and 73consonants 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].). All articulations were artificially

sustained during the 10s acquisition time, after removing the data acquisition in the process of incomplete pronunciation. Finally, a total of 104 sound pronunciations of organs under the MRI data have been collected.

Table1 the vowels and pseudo-consonants list of Chinese

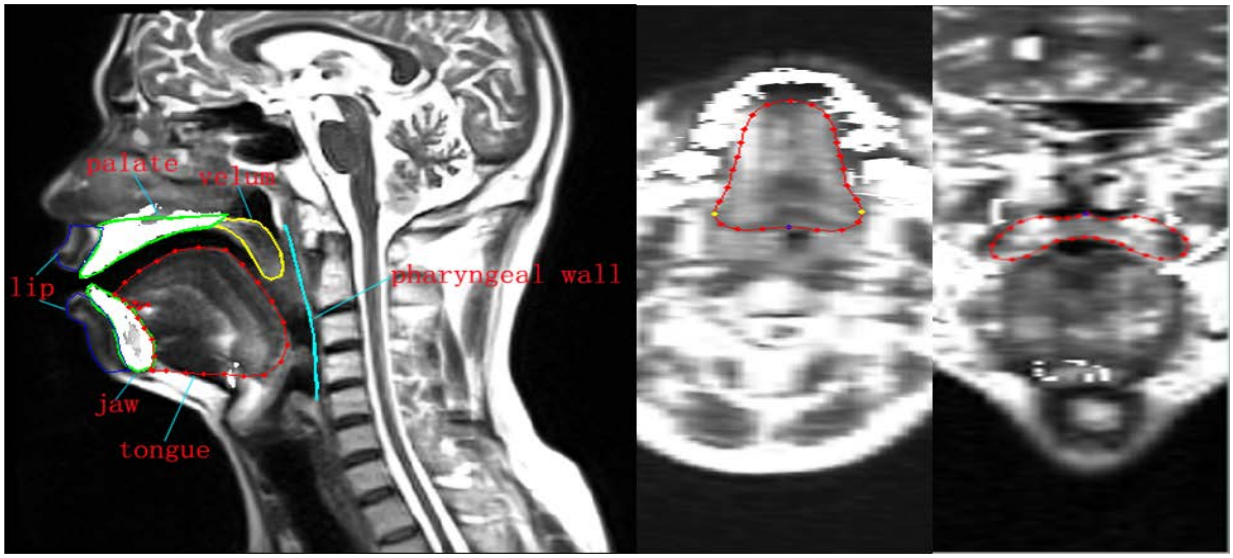
Vowel	[a], [i], [ɿ], [ʅ], [u], [ɛ], [ɤ], [o], [y]
Fricative	[s]+[a],[i],[u]; [ʂ]+[a],[i],[u]; [ɕ]+[i],[y]; [f] + [a], [ɛ], [u], [o]; [x] + [a], [ɛ], [u];
Stop	[t]+[a], [i], [u], [ɛ]; [k]+[a], [i], [u], [ɛ]; [p] + [a], [i],[u],[o]; [p^h] + [a],[i],[u],[o]; [t^h] +[a], [i] , [u], [ɛ] ; [k^h]+[a], [i], [u], [ɛ];
Affricate	[ʦ]+[a], [i], [u], [ɛ]; [tʂ]+[a], [i], [u], [ɛ]; [tɕ]+ [i], [y]; [$ʦ^h$]+[a], [i], [u], [ɛ] ; [$tʂ^h$]+[a], [i], [u], [ɛ]; [$tɕ^h$]+[i], [y];
Nasal	[m] + [a], [i], [u], [o], [ɛ]; [n]+[a], [i], [u];
Lateral	[l]+[a], [i], [u], [y], [ɤ];
Approximant	[r] + [i];

In the data collection process, the subject maintained the supine posture. 31 head parallel slices was adopted to record the corresponding state of the pronouncing participant, and all these slices are combined together to reconstruct the three-dimensional head model. Due to bone structure of the upper (lower) jaw and impossibility of directly imaging in the MRI, we adopt CBCT (Cone Beam Computer Tomography) to collect the participant's data of the bone structure of upper and lower jaw. The bone structure was added to the MRI data by rotating and translating the CBCT data.

3-Dimensional Model

The 3D modeling of articulatory: Based on the 104 MRI data collected from the database, we rebuild 3 complementary perspectives of sagittal plane, transverse plane and coronal plane (Figure 1a, 1b and 1c, respectively). These 3 perspectives will make us easily referring the boundary information of the articulatory organs, thus improving the modeling precision.

Spoken lips are represented by generic surface meshes fitted to the 3D contours extracted from MRI for each articulation. Due to the low-resolution image of MRI and the difficult division of organ boundary, the manual extraction of boundary is used. Then we fit these slice-parallel contours based on a specific rule and rebuild the 3D surface of the lips. Similarly, we build a 3D jaw model by using above method, and we get 104 jaw data by rotating and translating this model.



(a) Sagittal (b) transverse (c) coronal

Fig.1. The experimental results

The annotation of lips: The lips (located in the area which is surrounded by the blue line in the Figure 1(a)) can extend the vocal tract while pronunciation. In our work, annotations were made in the sagittal plane, and anchor points were marked in each image to maintain the continuity of articulators. For the upper lip, the anchor points are at the nasolabial angle and the anterior nasal spine. For the lower lip, the anchor point is at the intersection of the picture horizon and lowest point of the jaw. And then the entire contours of lips could be sketched. As showed in Figure 2, the upper lip contains 12*20 points and lower lip contains 18*20points.

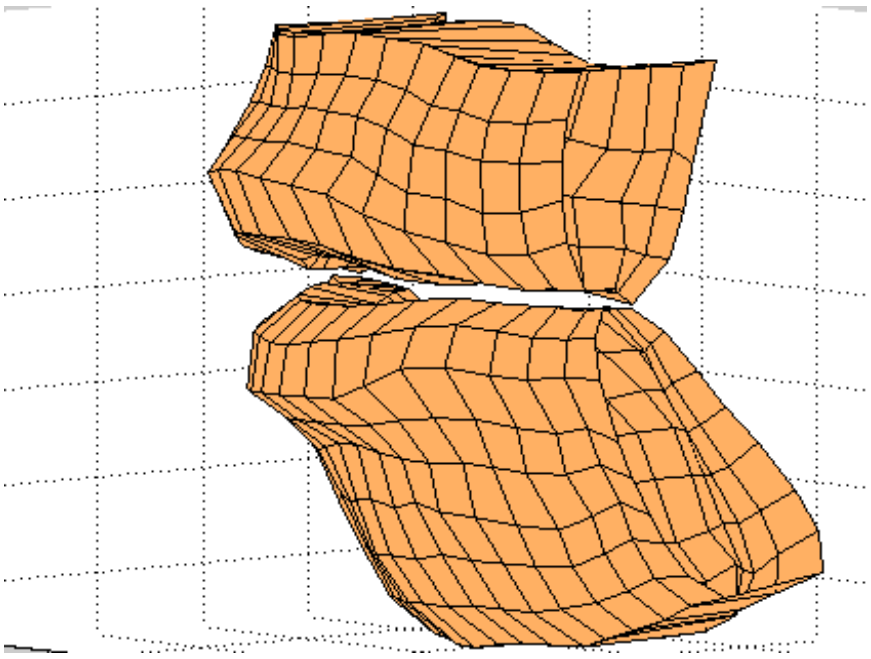


Fig.2.The upper and lower lips

Methodology

The Linear Component Analysis(LCA) is used to extract a set of meaningful articulatory parameters that control the shape of spoken lips [11].The articulatory parameters are determined in light of the following procedures(Figure 3): (i)The data describing the jaw movement is fed to PCA to extract the component for jaw movement. (ii) The active spoken lips movements are obtained by using linear regression to remove the influence of jaw movements on the extracted jaw components. (iii)The residue obtained in step 2(ii) is fed to PCA to extract active articulatory components.

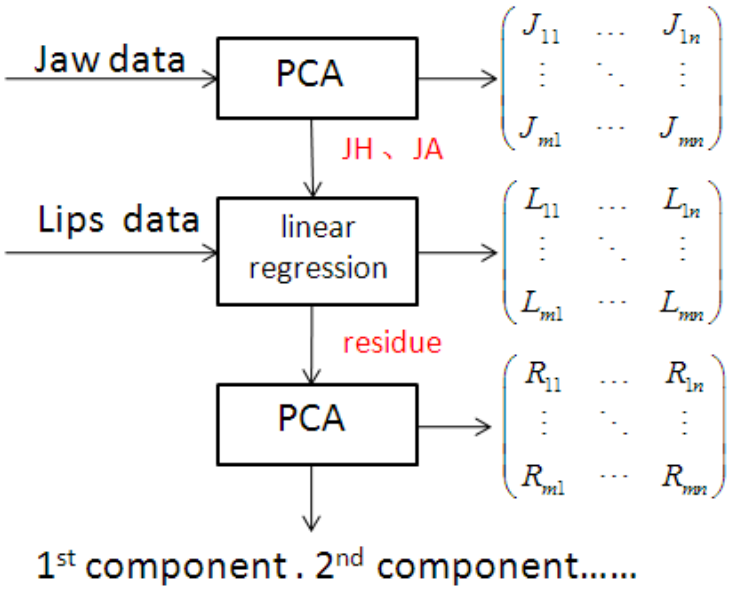


Fig.3. The procedure of extracting articulatory parameters

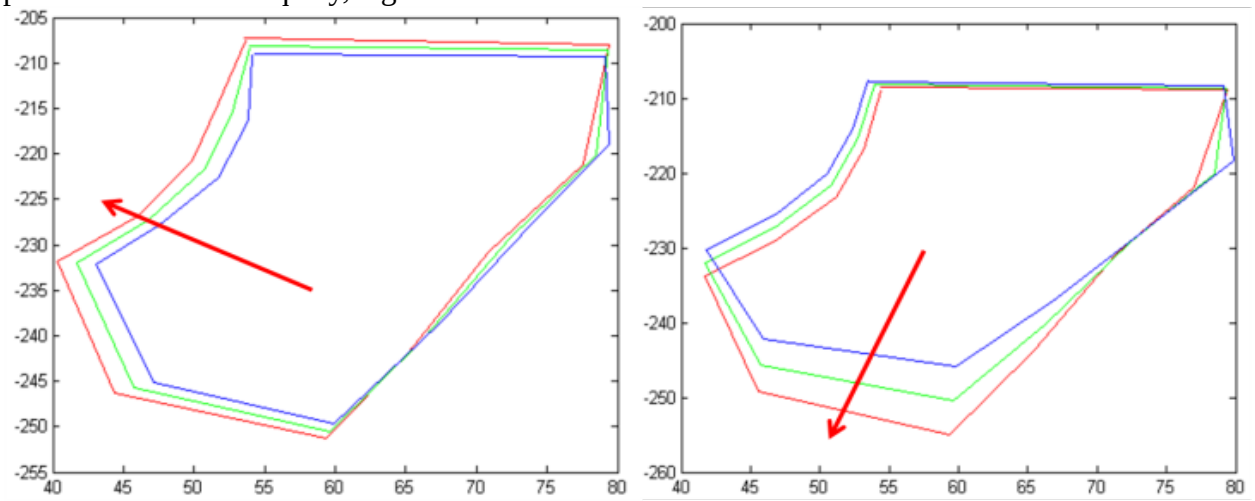
The RMSE of a specific vertex is calculated using the following equation:

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

Where Ns is the number of the articulatory shape samples, vr is the coordinates of a reconstructed articulatory vertex, and v is that of the corresponding original articulatory vertex.

Results

Upper lip: Linear Component Analysis (LCA) applied to extract the control component of the upper lip. Although the two components could only account for about 46% of total variance, the RMSE of reconstruction is 0.1043cm, and the reconstruction error can be acceptable. The physical meaning of extracted components is shown in Figure 4. Green line is the standard deviation of the original shape, the red and blue are ±2 times more than original shape respectively. The 1st component (Figure 4(a)) moves lip left-upward obliquely. The 2nd component (Figure 4(b)) moves lip left-downward obliquely, e.g.

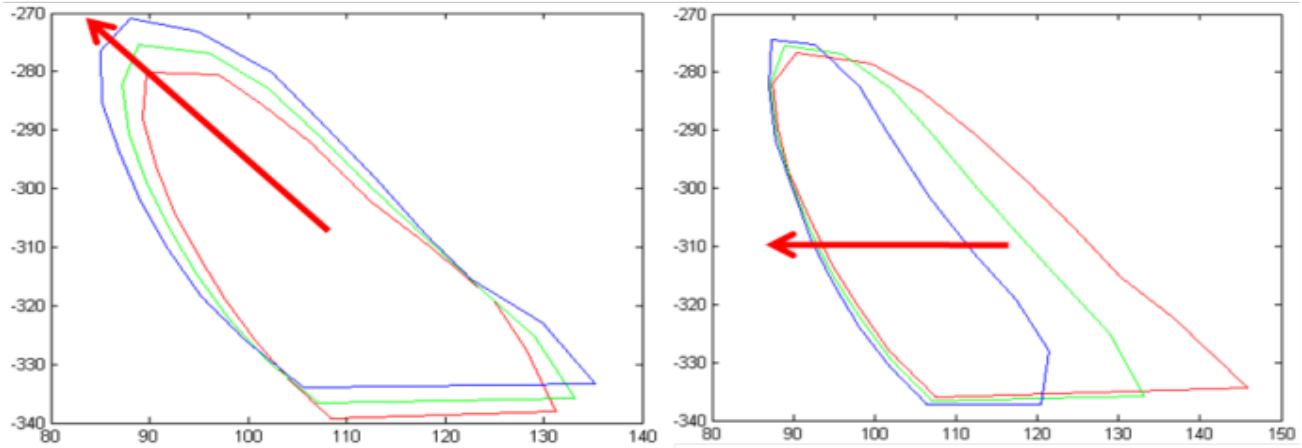


(a) First component: upper left (b) second component: lower left

Fig.4.The movement of the upper lip

Lower lip: Linear Component Analysis (LCA) applied to extract the control component of the lower lip. Similarly the two components could only account for about 46% of total variance, the

RMSE of reconstruction is 0.1374cm, standard deviation is 0.0246cm, and the max error of lower lip surface of the vertex is 0.7205cm.The physical meaning of extracted components is shown in Figure 5. Green line is the standard deviation of the original shape, the red and blue are ±2 times more than original shape respectively. The 1st component (Figure 5(a)) moves lip left-upward obliquely. The 2nd component (Figure 5(b)) moves horizontal direction lightly.



(a) First component: upper left (b) second component: left

Fig.5.The movement of the lower lip

Conclusion

Based on 104 speech MRI database, we built an articulatory model. The Linear Component Analysis (LCA) is applied to extract the control components to control the 3D model of the vocal tract (lips), and reconstruction of the articulatory by these components, the error is less than 0.15cm. At present, we only modeled and analyzed the static state of the articulatory, lack of the continuous speech data. Therefore, in the future, we will use EMA to drive the articulatory to simulate the progress of speech production.

Acknowledgement

In this paper, the research was supported in part by the grants from the National Natural Science Foundation of China (Program No. 61304250 and 61471259).

References

[1] Fang, Q., et al. Investigation of the functional relationship of tongue muscles for the control of a physiological articulatory model [C].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 [J]. TSINGHUA SCIENCE AND TECHNOLOGY, 2009. 14 (5).

[3] Dang J, Honda K. Construction and control of a physiological articulatory model [J]. Journal of the Acoustical Society of America, 2004, 115(2):853-70.

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

[5] Maeda S. Compensatrory Articulation During Speech: Evidence from the Analysis and Synthesis of Vocal Tract Shapes using an Articulatory Model [M].Speech Production and Speech Modelling. Springer Netherlands, 1990:91 - 100.

[6] Badin P, Serrurier A. Three-dimensional modeling of speech organs : Articulatory data and models [J]. Ieice Technical Report Speech, 2006, 106(177):29-34.

[7] Engwall O. Combining MRI, EMA and EPG measurements in a three-dimensional tongue model [J]. Speech Communication, 2003, 41(2-3):303-329.

[8] Birkholz P, Jackel D, Kroger B J. CONSTRUCTION AND CONTROL OF A THREE-DIMENSIONAL VOCAL TRACT MODEL [C] International Conference on Acoustics, Speech, and Signal Processing. 2006:I.

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

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

[11] Fang Q, Liu J, Song C, et al. A novel 3D geometric articulatory model [C]. International Symposium on Chinese Spoken Language Processing. IEEE, 2014:368-371.