HMM BASED SPEECH-DRIVEN 3D TONGUE ANIMATION

Changwei Luo[1], *Jun Yu*[1], *Xian Li*[1], *Leilei Zhang*[2]

[1]Department of Automation, University of Science and Technology of China
[2] College of Information Science and Electronic Engineering, Zhejiang University

luocw@mail.ustc.edu.cn, harryjun@ustc.edu.cn

ABSTRACT

We propose a speech-driven 3D tongue animation system. Firstly, the input speech is analyzed to obtain the phoneme sequence. Next, articulatory movements are predicted from the phoneme sequence using a hidden Markov model (HMM) based framework. The HMMs are trained beforehand using a corpus of human articulatory movements, which are recorded by three electromagnetic articulograph (EMA) sensors glued on the tongue tip, tongue body, and tongue dorsum of a speaker respectively. Finally, the predicted articulatory movements are used to control the deformations of a 3D tongue model. The tongue model is a triangular mesh with three key vertices. The three vertices are chosen so that their positions are in correspondence with the three EMA sensors mentioned above. Our tongue model can achieve various tongue shapes with volume preservation. Experiments show that the generated tongue animations are realistic and synchronize well with the input speech.

Index Terms— speech-driven, tongue animation, hidden Markov model, articulatory movements

1. INTRODUCTION

Speech animation is a hot research topic in recent years. It has many applications, such as computer games and human-machine interaction. In human-machine interaction applications, a lip-sync talking head can attract the attention of a user, and make human-machine interaction more effective [1, 2].

A number of researchers have described techniques for synthesizing realistic speech animations (or talking heads) [3, 4]. Wang et al. [3] propose a system which renders a photo-real video of articulators in sync with the given speech by searching for the most plausible real image sample sequence. Deng et al. [5] first construct explicit speech co-articulation models from real human motion data. Then new speech animations are synthesized by blending a few 3D viseme shapes. A drawback of the system is that the tongue is not modeled, which makes the output speech animation less convincing. The human tongue is the most important speech organ. To increase the intelligibility of synthesized speech animation, a realistic 3D tongue model is highly required [6].

There exist some tongue models for speech animation or visual speech synthesis [7–9]. In [7], a parametric tongue model with six animation parameters is presented. The tongue model is composed of a B-spline surface with 60 control points. The weights of an animation parameter on the control points are manually determined. These weights are sensitive to the shape of tongue in its rest position. The work of [10] tracks the tongue contour in the X-ray images, and then animates a tongue model according to the tracking results.

Researchers have also proposed some finite element models (FEM) for the tongue [11, 12]. The FEM methods divide the tongue into small units, often tetrahedrons or prisms, and then define the strain and the elastic properties for these units. In most existing literature, the deformation of the tongue is driven by muscle activations. Although these methods are capable of generating realistic tongue motions, they are quite computationally expensive.

In [13], a 3D tongue model is developed using magnetic resonance (MR) images of a human subject producing 44 artificially sustained articulations. Based on the difference in tongue shape between articulations and a reference tongue model, six linear control parameters are obtained using linear component analysis.

In this paper, we propose a deformable tongue model for speech animation. The tongue model is a triangular mesh constructed from MR images. Three key vertices are selected on the tongue mesh, and are used to control the deformations of the tongue model. The three key vertices are respectively located at tongue tip, tongue body, and tongue dorsum of the tongue model. To generate accurate tongue motions from speech, we predict the movements of tongue tip, tongue body and tongue dorsum from speech based on HMMs [14]. The predicted movements are used to determine the position of the three key vertices of the tongue model. The positions of the rest vertices are determined by minimizing the deformation energy. Our tongue model is able to achieve various tongue shapes with exact volume preservation.

This work is supported by National Natural Science Foundation of China (61572450), Anhui Provincial Natural Science Foundation (1708085QF138), the Fundamental Research Funds for the Central Universities (WK2350000002) and the Open Funding Project of State Key Lab of Virtual Reality Technology and Systems, Beihang University (BUAA-VR-16KF-12). Corresponding author: Jun Yu.

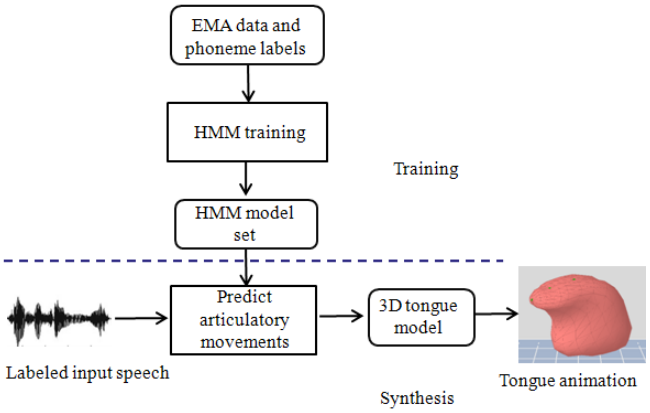

Fig. 1. Overview of our HMM based speech-driven tongue animation system

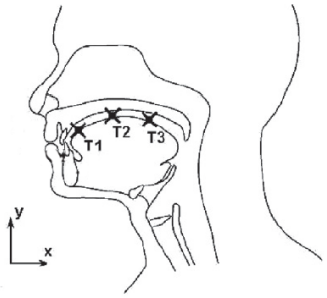

Fig. 2. The placement of the three EMA sensors used to record the articulatory movements of the tongue

2. SYSTEM OVERVIEW

An overview of our speech-driven tongue animation system is show in Figure 1. It consists of a training stage and a synthesis stage. In the training stage, we first build a data set comprised of articulatory movements recorded concurrently with the corresponding acoustic waveforms. The articulatory movements are recorded using electromagnetic articulograph (EMA). Three EMA sensors are used, and respectively glued on the tongue tip (T1), tongue body(T2), and tongue dorsum (T3) of a speaker. This is illustrated in Figure 2. We use a HMM based framework to model the relationship between phonemes and articulatory features (i.e., articulatory movements). The articulatory features are composed of the positions of the three EMA sensors.

In the synthesis stage, the phoneme sequence associated with input speech is used to predict articulatory features based on the trained HMMs. Since the three EMA sensors which record the articulatory features are in correspondence with the three key vertices of our tongue model, the predicted articulatory features can be directly used to determine the positions of the three key vertices during speech. The deformations of other vertices of the tongue model are calculated based on the proposed deformable tongue model.

3. PREDICTING ARTICULATORY MOVEMENTS FROM PHONEME-ANNOTATED SPEECH

To synthesize speech-driven tongue animation, we first need convert the speech into articulatory movements. This procedure is called audio-to-visual conversion. We perform the conversion using a HMM-based method. To train the HMMs, a data set comprised of articulatory movements and waveforms is required. Here EMA is used to record the continuous motions of the tongue during speech. Three EMA sensors are used and placed in the midsagittal plane. Each sensor records spatial location in 3 dimensions at a 200 Hz sample rate. The movements in z-axis are very small. Therefore, only the x- and y-coordinates of the three sensors are used in our experiments, making a total of 6 static articulatory features at each sample instant. The speech waveforms are aligned with phoneme labels using force alignment [14]. Then the articulatory features share the same alignments with speech waveforms.

let x_t be the static articulatory feature at frame t. To consider the dynamics of the articulatory movements, the observation feature vector $X_t = [x_t, \Delta x_t, \Delta^2 x_t]$ is composed of the static features x_t, the velocity Δx_t and acceleration $\Delta^2 x_t$,

$$\Delta x_t = 0.5x_{t+1} - 0.5x_{t-1} \tag{1}$$

$$\Delta^2 x_t = x_{t+1} - 2x_t + x_{t-1} \tag{2}$$

Let $x = [x_1, x_2, ..., x_T]$ be static feature vector sequence. The observation feature vector sequence $X = [X_1, X_2, ..., X_T]$ can be written as

$$X = Wx \tag{3}$$

where W is a transformation matrix described in [14].

Except articulatory features, acoustic features such as Mel-frequency cepstral coefficients can also be used as observation features [11]. However, the acoustic features differ greatly for different speakers. To make our system user independent, only the articulatory features are used as observe vectors of HMM. In the training of HMM, a 5-state, left-to-right model structure with no skips is adopted. A single Gaussian distribution with diagonal covariance is used for each HMM state.

After HMM training, articulatory features can be predicted from input speech. Firstly, speech is aligned with phoneme sequence using force alignment. Secondly, a sequence of HMM are concatenated based on the phoneme sequence. Finally, articulatory features are generated by using the maximum likelihood parameter generation algorithm [15],

$$x^* = \arg\max P(Wx|\lambda, l) \tag{4}$$

where x^* is the optimal static observation vector sequence (i.e., the articulatory feature sequence). λ is the trained HMMs. $l = [l_1, l_2, \cdots, l_T]$ is the phoneme sequence.

4. THE DEFORMABLE TONGUE MODEL

We construct a 3D tongue model from MR images. To animate the tongue model, we select three key vertices on the tongue model. The three vertices are in one-to-one correspondence with the three EMA sensors glued on the tongue of the speaker. By using the method of Section 3, we can generate ariticulatory features from input speech. Actually, the ariticulatory features are the predicted positions of three EMA sensors. So the ariticulatory features can be directly used to determine the positions of the three key vertices. Our tongue model can deform effectively given only the positions of the key vertices.

4.1. Tongue model reconstruction

We reconstruct a 3D tongue model using a set of MR images. Firstly, the tongue contours are manually extracted from each MR image, and 18 points are sampled on each contour. Then, the surface mesh of the tongue model is created by connecting each vertex to its neighbor vertices. Details of the construction procedure can be found in [13]. In this paper, we only need to reconstruct the 3D shape of the tongue in its rest position, while [13] needs to reconstruct 44 3D shapes for different articulations.

4.2. Deformations of the tongue model

To control the deformations of the 3D tongue model, we select three key vertices on the tongue model. These three key vertices are in correspondence with these three EMA sensors. For our speech-driven tongue animation system, the positions of the three key vertices can be predicted using the trained HMMs. The remaining problem is how to deform the tongue model according to the three key vertices.

We represent the deformation as a collection of affine transformations. Let p_1, p_2 and p_3 be the undeformed vertices of a triangle of the tongue model, and $\hat{p}_1$, $\hat{p}_2$ and $\hat{p}_3$ be the deformed vertices of this triangle. An affine deformation is defined by a 3×3 matrix Q and a 3×1 translation vector d,

$$Qp_i + d = \hat{p}_i \tag{5}$$

where $i = 1, 2, 3$. The three vertices of a triangle before and after affine deformation could not fully determine the affine transformation $[Q, d]$ since they do not give how the space perpendicular to the triangle deforms. To solve this issue, we add a fourth vertex p_4 to the triangle in the direction perpendicular the triangle.

$$p_4 = p_1 + (p_2 - p_1) \times (p_3 - p_1) / ||(p_2 - p_1) \times (p_3 - p_1)|| \tag{6}$$

After the definition of the fourth vertex, the transformation matrix Q can be determined as follows [16]:

$$Q = \hat{P}P^{-1} \tag{7}$$

where

$$\hat{P} = [\hat{p}_2 - \hat{p}_1, \hat{p}_3 - \hat{p}_1, \hat{p}_4 - \hat{p}_1] \tag{8}$$

$$P = [p_2 - p_1, p_3 - p_1, p_4 - p_1] \tag{9}$$

$\hat{p}_4$ is the deformed vertex for p_4.

Based on the transformation matrix Q, we define the deformation energy as follow:

$$E = \frac{1}{2} \sum_{i=1}^{M} \|Q_i - I\|^2 \tag{10}$$

where M is the number of triangles of the tongue model. Q_i is the matrix for the ith triangle. I is a 3×3 identity matrix. The positions of the deformed vertices $\hat{p}_j$ can be solved by minimizing the deformation energy E. Besides, the tongue is mainly composed of muscle fibers and its volume is nearly unchanged during movements. Thus, given the positions of the three key vertices, we obtain the deformed positions of all the mesh vertices by solving the following minimization problem:

$$\min E(\hat{p}_1, \hat{p}_2, ..., \hat{p}_N) = \frac{1}{2} \sum_{i=1}^{M} \|Q_i - I\|^2 \tag{11}$$

subject to:

$$\hat{p}_{c_k} = v_k,\ k = 1, 2, 3. \tag{12}$$

$$V - V_0 = 0 \tag{13}$$

where $\hat{p}_1, \hat{p}_2, ..., \hat{p}_N$ is the deformed vertices. N is the number of vertices of the tongue model. c_k is the index for the three key vertices. v_k is the predicted positions of the three key vertices. V_0 and V are the volumes before and after deformation respectively. V_0 can be calculated as follows:

$$V_0 = \frac{1}{6} \sum_{i=1}^{M} p_{i,1} \cdot (p_{i,2} \times p_{i,3}) \tag{14}$$

$p_{i,1}$, $p_{i,2}$, $p_{i,3}$ are the three vertices of the ith triangle. V can be calculated using a similar equation. Substituting (12) to (11), then (11) can be reformulated in matrix form:

$$\min E(U) = \frac{1}{2} \|A \cdot U - b\|^2 \tag{15}$$

subject to

$$g(U) = V - V_0 = 0 \tag{16}$$

where U is a vector of unknown positions of the deformed vertices. A is a matrix determined by the undeformed tongue shape (i.e., the initial tongue shape). b is a vector determined by the three key vertices. We solve the optimization problem by applying Lagrange multipliers with Newton's method [17].

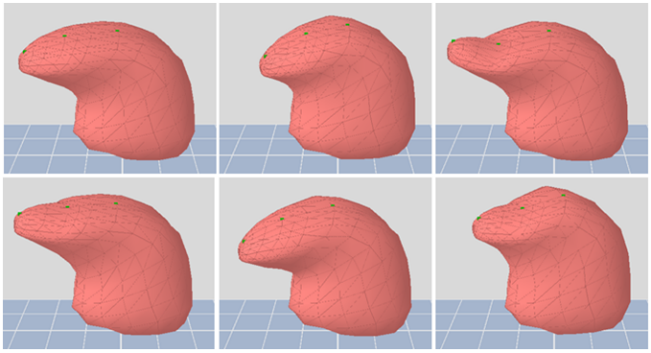

Fig. 3. Example tongue animations from profile view.

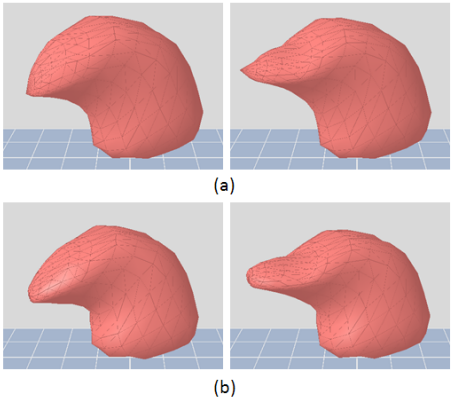

Fig. 4. Comparison between tongue animations. (a) Animations without volume preservation. (b) Animations with volume preservation.

5. EXPERIMENTS

5.1. Articulatory feature prediction

In our experiments, we use a mandarin Chinese corpus containing 380 sentences. 360 sentences are used for training and the rest for testing. The audio data are recorded at 16KHz, and the articulatory data are sampled at 200Hz. Only articulatory features and phoneme labels are used to train the HMMs, and no acoustic features are used.

The HTS Toolkit [18] is used to train the HMMs. To evaluate the accuracy of articulatory feature prediction, two HMM systems are trained, one with monophone HMM models and the other with triphone HMM models. The output distributions of the HMMs are represented by single Gaussian densities. There are a total of 61 monophone models (21 initials and 38 finals plus two phonemes for silence and short pause). There are 4398 triphone models in the training set. Decision tree clustering is performed on these triphone models to ensure adequate training data for the HMMs. In our experiments, there are 5 states for each HMM, and each state is clustered independently.

Table 1. RMS error of predicted articulatory movements (mm)

	monophone HMM	triphnone HMM
T1_x	2.56	2.06
T1_y	2.47	1.91
T2_x	2.61	1.88
T2_y	2.32	2.02
T3_x	2.41	2.12
T3_y	3.11	2.41
Average	2.58	2.06

Table 1 shows the root mean square (RMS) error of predicted articulatory movements in millimeter. T1_x and T1_y are the x, y coordinates of T1 (tongue tip). It is shown that triphone models outperform monophone model in terms of RMS error. The RMS error is very small. Thus, we can accurately estimate the articulatory movements from speech.

5.2. Speech-driven tongue animation

To synthesize speech-driven tongue animation, the articulatory movements predicted from input speech are directly used to determine the positions of three key vertices of the tongue model. The tongue model is then animated according to the positions of the three key vertices.

Figure 3 shows a few examples of generated animations. We can see that our tongue model can achieve various tongue shapes. More speech-driven tongue animations can be seen in the accompanying video.

Volume preservation is also necessary for realistic tongue animations. Figure 4 makes a comparison between the animations without volume preservation and those with volume preservation. It is shown that tongue animations with volume preservation look more realistic.

6. CONCLUSION

A speech-driven tongue animation system is presented. Given the input speech and its phoneme sequence, we predict the articulatory moments using a HMM based system with triphone HMM models. The predicted articulatory movements are used to animate a deformable 3D tongue model. The tongue model is a triangular mesh with three key vertices, which are respectively located at tongue tip, tongue body, and tongue dorsum of the tongue model. The key vertices are used to control the deformations of the tongue. Our tongue model is capable of approximating various tongue shapes with exact volume preservation. In the future, we would like to integrate our tongue model into a speech facial animation framework.

7. REFERENCES

[1] J Ostermann and A Weissenfeld, "Talking faces-technologies and applications," in *International conference on pattern recognition*, 2004.

[2] J Yu and Z Wang, "A video, text and speech-driven realistic 3-d virtual head for human-machine interface," *IEEE Transactions on Cybernetics*, vol. 45(5), pp. 977–988, 2015.

[3] L Wang, X Qian, W Han, and F Soong, "Synthesizing photo-real talking head via trajectory-guided sample selection," in *interspeech 2010*, 2010.

[4] C Luo, J Yu, and Z Wang, "Synthesizing real-time speech-driven facial animation," in *International Conference on Acoustics, Speech, and Signal Processing*, 2014, pp. 4568–4572.

[5] Z. Deng, U. Neumann, J. Lewis, T. Kim, M. Bulut, and S. Narayanan, "Expressive facial animation synthesis by learning speech coarticulation and expression spaces," *IEEE Trans. on visualization and computer graphics*, vol. 12(6), pp. 1523–1534, 2006.

[6] J Yu, C Jiang, R Li, C Luo, and Z Wang, "Real-time 3d facial animation: From appearance to internal articulators," in *IEEE Transactions on Circuits and Systems for Video Technology*, DOI: 10.1109/TCSVT.2016.2643504, 2016.

[7] S King and R Parent, "Creating speech-synchronized animation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 11(3), pp. 341–352, 2005.

[8] Z Chen, X Zhang, and Z Wu, "A new tongue model based on muscle-control," in *IEEE International Conference on Granular Computing*, 2011.

[9] R Li, J Yu, C Jiang, C Luo, and Z Wang, "A mass-spring tongue model with efficient collision detection and response during speech," in *The 9th International Symposium on Chinese Spoken Language Processing*, 2014, pp. 354–358.

[10] C Luo, R Li, L Yu, J Yu, and Z Wang, "Automatic tongue tracking in x-ray images," *Chinese Journal of Electronics*, vol. 24(4), pp. 767–771, 2015.

[11] Q Fang, S. Fujita, X Lu, and J. Dang, "A model-based investigation of activations of the tongue muscles in vowel production," *Acoustical Science and Technology*, vol. 30(4), pp. 277–287, 2009.

[12] Yin Yang, Xiaohu Guo, Jennell Vick, Luis Torres, and Thomas Campbell., "Physics-based deformable tongue visualization," *IEEE Transaction on Visualization and Computer Graphics*, vol. 19(5), pp. 811–823, 2013.

[13] Olov Engwall, "Combining mri, ema and epg measurements in a three-dimensional tongue model," *Speech Communication*, vol. 41, pp. 303–329, 2003.

[14] Zhenhua Ling, Korin Richmond, and Junichi Yamagishi, "An analysis of hmm-based prediction of articulatory movements," *Speech Communication*, vol. 52(10), pp. 834–846, 2010.

[15] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *International Conference on Acoustics, Speech, and Signal Processing*, 2000, pp. 1315–1318.

[16] R Summer and J Popovic, "Deformation transfer for triangle meshes," *ACM Transactions on Graphics*, vol. 22(3), pp. 399–405, 2004.

[17] J Huang, X Shi, X Liu, and et al., "Subspace gradient domain mesh deformation," *ACM Transactions on Graphics*, vol. 25(3), pp. 1126–1134, 2006.

[18] H Zen, K Tokuda, and A Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51(11), pp. 1039–1154, 2009.