CONTINUOUS ULTRASOUND BASED TONGUE MOVEMENT VIDEO SYNTHESIS FROM SPEECH

Jianrong Wang[†] *Yalong Yang*[⋆Ш] *Jianguo Wei*[‡*] *Ju Zhang* [†]

[†] School of Computer Science and Technology, Tianjin University
[⋆] Caulfield School of Information Technology, Monash University
[Ш] National ICT Australia (NICTA) Victoria
[‡] School of Computer Software, Tianjin University
{wjr, jianguo, juzhang}@tju.edu.cn, yalong.yang@monash.edu

ABSTRACT

The movement of tongue plays an important role in pronunciation. Visualizing the movement of tongue can improve speech intelligibility and also helps learning a second language. However, hardly any research has been investigated for this topic. In this paper, a framework to synthesize continuous ultrasound tongue movement video from speech is presented. Two different mapping methods are introduced as the most important parts of the framework. The objective evaluation and subjective opinions show that the Gaussian Mixture Model (GMM) based method has a better result for synthesizing static image and Vector Quantization (VQ) based method produces more stable continuous video. Meanwhile, the participants of evaluation state that the results of both methods are visual-understandable.

Index Terms— Multimodal interface, Movement synthesis, Vector Quantization, Gaussian Mixture Model, Ultrasound

1. INTRODUCTION

Speech synthesis has been studied by many researchers for a long time, but relatively little attention has been paid to use the speech signal to synthesis the movement of acoustic organs. Recently, there has been increasing interest in investigating the role of visual information in speech processing. Researchers have proofed that by adding visual information of acoustic organ movement to acoustic signals improves the speech intelligibility in noisy environment [1, 2] in early time. However, because of the limitation of both hardware and computation ability, no further studies or applications have been explored until 1990s. Morishima proposed a framework using VQ and Neural Networks to build the mapping from speech signals to 3D key points of a face [3, 4]. After that, GMM [5, 6] and Hidden Markov Model (HMM) [7] based method were used to map speech signals to 3D key lip positions. Furthermore, real-time talking head system has been developed to help language learning [8]. Meanwhile, with the development of Silent Speech Interface (SSI) [9], researchers realized not only the visual information of face images, but also the movement of other acoustic organs plays an important role in speech processing.

The movement of the tongue is critical to the pronunciation [10]. Moreover, the visual information of tongue movement will contribute to the speech intelligibility [11], and helps learning a second language [12]. Visualizing the movement of tongue can also be used in speech therapy for speech retarded children, of perception and production rehabilitation of hearing impaired children and of pronunciation training [13].

However, hardly any research has been explored on tongue movement synthesis. To the best of our knowledge, the only related work is using Deep Neural Network (DNN) to build a 2-way mapping between vowel speech and ultrasound image [14], but the authors are focusing on only 6 Chinese vowels which are stable in pronouncing process and they also admit that the synthesis of ultrasound images need to be improved.

This paper proposes a training and synthesis framework to build the mapping from acoustic speech signals to continuous tongue movement video. VQ-based and GMM-based mapping methods are applied separately in this framework. An ultrasound based synchronized visual-audio corpus used in the experiment is also described. Effectiveness of the proposed framework is verified by objective evaluation and subjective opinions.

2. FRAMEWORK

In this section, we will introduce the main parts of the framework. The framework consists of two parts: training part, which builds the mapping model; and synthesis part, which uses speech signal as input to synthesize continuous tongue movement video.

*Corresponding author

2.1. Training

In training part, we first extract the feature vectors from the acoustic speech signals and ultrasound tongue images. Then we use the feature vectors to build one-way mapping model. The training procedures are illustrated in Fig. 1.

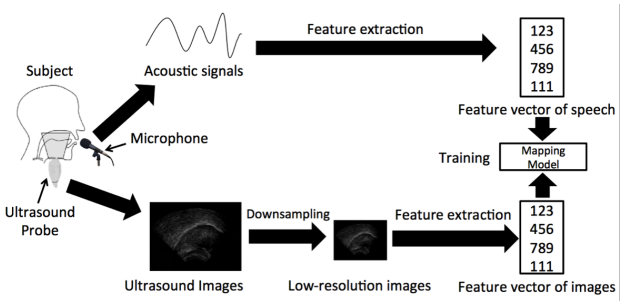

Fig. 1. Training procedures

In practical implementation, as the resolution of raw ultrasound images are very high, a down-sampling procedure is adopted to reduce the computation time while keeping the main feature of images.

2.2. Synthesis

In synthesis part, feature vectors of acoustic signals are used as input of the mapping model. The mapping model generates the related feature vector of ultrasound images. After that, reconstruction of image from the feature vector is applied to get the final tongue images. The synthesis procedures are presented in Fig. 2.

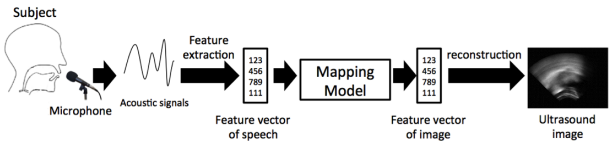

Fig. 2. Synthesis procedures

In this framework, special attention needs to be paid on the selection of feature vector of images. As we need to reconstruct images from the feature vectors in the synthesis procedure, we must choose features of image with the ability to reconstruct.

3. MAPPING MODELS

3.1. VQ-based model

VQ-based mapping model is using the VQ technology to build two codebooks for the input and output vectors respectively and using statistics to build the mapping function from the input to the output. The training and synthesis algorithm of VQ-based mapping model are described below:

A. Training:

1. Calculate VQ codebooks of N codeword vectors for both input vectors and output vectors respectively.
2. Calculate a correspondence histogram, $w_{i,j}$ represents the number of synchronized j_{th} codeword vector(C_j^V) in video codebooks for the i_{th} codeword vector(C_i^A) in audio codebooks.

B. Synthesis:

1. For a given new input vector f, find the codeword vector(C_f^A) that has the minimized distance of it.
2. Sort the list $w_{f,1}, w_{f,2}, \cdots, w_{f,N}$ that we can obtain from training procedure in the order from big to small $w_{f,x1}, w_{f,x2}, \cdots, w_{f,xN}$. For instance, $w_{f,x1}$ shows that (C_f^A) and (C_{x1}^V) synchronized the most in the training set.
3. We use the K largest count to calculate the expected output vector in Eq. (1).

$$C_f^A \rightarrow \frac{\sum_{k=1}^{K} w_{f,xk} C_{xk}^V}{\sum_{k=1}^{K} w_{f,xk}} \tag{1}$$

3.2. GMM-based model

The VQ procedure divides continuous data vector into discrete codebook, which could result in distortion of origin data. In the GMM-based model, we use the probability density function to map continuous inputs, which is expected to gain better results.

A. Training:

First we need to build a joint GMM model using the training set.

For a-dimensional input vector

$$C_t^A = [x_i(1), x_i(2), \cdots, x_i(a)]$$

and b-dimensional input vector

$$C_t^V = [y_i(1), y_i(2), \cdots, y_i(b)]$$

at frame t, the joint vector

$$z_t = [x_i(1), x_i(2), \cdots, x_i(a), y_i(1), y_i(2), \cdots, y_i(b)]$$

The joint GMM with M components will be calculated (m is the index of the component), the parameter set (λ_z) is presented as follow:

- Weight: w_m
- Mean: $v_m = \begin{bmatrix} v_m^x \\ v_m^y \end{bmatrix}$
- Covariance: $\Sigma_m = \begin{bmatrix} \Sigma_m^{xx} & \Sigma_m^{xy} \\ \Sigma_m^{yx} & \Sigma_m^{yy} \end{bmatrix}$

B. Synthesis : The conditional probability density of given input vector could be presented as Eq. (2).

$$P(C_t^V | f, \lambda_z) = \sum_{m=1}^{M} P(m|f, \lambda_z) P(C_t^V | m, f, \lambda_z) \tag{2}$$

The expected output $C_t^{V'}$ with the minimum mean-square error can be represented as Eq. (3) [6].

$$C_t^{V'} = E[C_t^V | f] = \sum_{m=1}^{M} p(m|f, \lambda_z) E_m^y \tag{3}$$

Where the mean vector and probability factor are calculated in Eq. (4) and Eq. (5) respectively.

$$E_m^y = v_m^y + \Sigma_m^{yx} (\Sigma_m^{xx})^{-1} (f - v_m^x) \tag{4}$$

$$p(m|f, \lambda_z) = \frac{w_m \mathcal{N}(f; v_m^x, \Sigma_m^{xx})}{\sum_{n=1}^{M} w_n \mathcal{N}(f; v_n^x, \Sigma_n^{xx})} \tag{5}$$

4. EXPERIMENTS AND DISCUSSION

4.1. Ultrasound based visual-audio corpus

Speech and visual ultrasound data for a male speaker of Chinese Mandarin were recorded using directional microphone and *Terason T3000* ultrasound system in a soundproof recording room. Audio was recorded at a sample rate of 44.1K and video was recorded at 90fps with 640 $\times$ 480 resolution.

The script of the corpus is a combination of Microsoft corpus and Speechocean corpus with 1000 sentence. After cleaning the data, 931 sentences are used for our experiments with 6,732 seconds audio and 606,133 ultrasound tongue images.

4.2. Feature extraction

The 1st to 13th Mel Frequency Cepstral Coefficient, its differential and acceleration coefficients are combined to 39 dimensional feature vectors to present the speech signal with 10ms shift frames using a 25ms Hamming window.

We choose the *EigenTongue* feature that has the ability to reconstruct and encode the maximum amount of relevant information in the ultrasound images [15]. As described in the section of framework, for the ultrasound images, we first down-sample the resolution to 160 $\times$ 120. In order to be compatible (albeit artificially) with a more standard frame rate for speech analysis, the sequences of EigenTongue coefficients are oversampled from 90 to 100Hz using linear interpolation. The effective frame size thus corresponds to 10 ms[16]. The first 40 EigenTongue coefficients are used to present each ultrasound image.

4.3. EigenTongue reconstruction

450,000 ultrasound images were chosen randomly to build the EigenTongue model. In Fig. 3, we present the result of reconstruction for a sentence from 40 EigenTongue coefficients to image (images are picked in $1s$ interval with $4s$ in total).

As can be seen from Fig. 3, the reconstructed images have differences from the origin images, but the most relevant information is enhanced, mainly the tongue position.

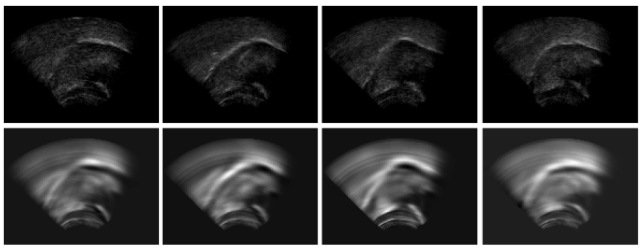

Fig. 3. EigenTongue reconstruction; the up row is the origin image and the bottom row is the result of reconstruction from EigenTongue

4.4. Synthesis evaluation

The synthesized EigenTongue coefficients were evaluated by time-averaged Euclidean error distance E and time-averaged differential error ΔE between the synthesized EigenTongue coefficients $[x_1^s, x_2^s, \cdots, \mathbf{x}_d^s]$ and origin EigenTongue coefficients $[x_1^o, x_2^o, \cdots, \mathbf{x}_d^o]$ (d is the number of dimension of EigenTongue coefficients, which in this experiment is 40).

$$E = \frac{1}{T} \sum_{t=1}^{T} \sqrt{\sum_{n=1}^{d} (x_n^s - x_n^d)^2}$$

$$\Delta E = \frac{1}{T} \sum_{t=1}^{T} \sqrt{\sum_{n=1}^{d} (\Delta x_n^s - \Delta x_n^d)^2}$$

For the VQ-based method, we construct 256 codework for each codebook of input and output. The largest count K of VQ-based method and the number of component M of GMM-based method are investigated. The result of the evaluation is illustrated in Table 1.

Method	E	ΔE
VQ($K = 1$)	7.27	4.61
VQ($K = 4$)	6.51	2.97
VQ($K = 16$)	6.21	2.08
VQ($K = 64$)	6.13	1.53
VQ($K = 256$)	6.10	1.31
GMM($M = 1$)	6.12	1.17
GMM($M = 4$)	6.10	1.42
GMM($M = 16$)	6.08	1.60
GMM($M = 64$)	6.07	1.77
GMM($M = 256$)	6.16	2.21

Table 1. Evaluation result of VQ-based and GMM-based methods

For the increasing K, E and ΔE are decreasing gradually in VQ-based method. In GMM-based method, with the increasing M, E remains in a relatively low magnitude and

only changes in a small scale. However, ΔE is increasing which is out of our expectation.

Fig. 4 shows the numeric synthesis results of two methods with $K = 64$ and $M = 64$. The results were compared with the original EigenTongue coefficients using one sentence. It could be seen from the figure that while the results of two methods both follow the trend of the original coefficients roughly, GMM-based method has a closer tracking, but also a much greater fluctuation than VQ-based method. GMM-based method is more sensitive to changes.

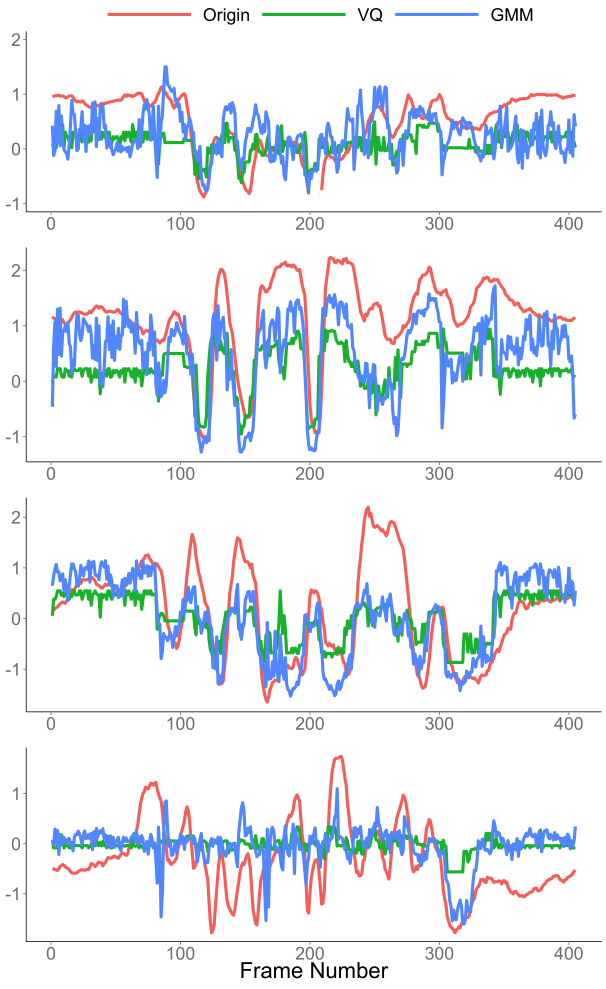

Fig. 4. Numeric synthesis results of one sentence, from top to bottom are the results of $1st$, $2nd$, $3rd$, $4th$ EigenTongue coefficients

Fig. 5 presents the visual synthesis result of the methods. We can observe the general trajectory of tongue from both results. Meanwhile, the result of GMM-based methods shows more details than the VQ-based method.

We also investigate 20 students with computer science background for general opinions of the visual result. The result could be concluded as:

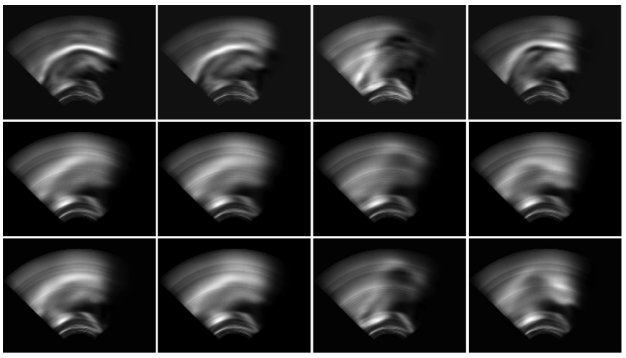

Fig. 5. Visual synthesis result of one sentence ($1s$ interval, $4s$ in total), top row is EigenTongue reconstructed image; middle row and the bottom row are synthesized by QV-based method and GMM-based method respectively

- Result of both methods is visual-understandable;
- For static images, the GMM-based method has a better visual-satisfied result;
- For continuous video, the VQ-based method performs better, as the tongue is trembling in GMM-based method, while VQ-based method is more stable.

5. CONCLUSION AND FUTURE WORKS

This paper proposed a promising framework to synthesis continuous ultrasound tongue movement video through speech. We also applied two different mapping models to the framework. The results of two models are evaluated by objective tests, and subjective opinions are also collected.

The evaluation states that both this two methods are acceptable for synthesizing video, while GMM-based method has a better result of separated images and VQ-based method produces more stable video.

For the future research, we prepare to focus on the following aspects:

- Detailed evaluations;
- Stabilize the video of GMM-based method;
- Improve the presentation of ultrasound images;
- Improve the framework; instead of synthesizing the images directly, pick up images from an image database and concatenated to a video might have a better result.

6. ACKNOWLODGEMENT

This work was supported by the National Natural Science Foundation of China (No.61471259, No.61304250 and No.61175016).

7. REFERENCES

[1] W H Sumby, "Visual Contribution to Speech Intelligibility in Noise," *The Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212, 1954.

[2] Keith K Neely, "Effect of Visual Factors on the Intelligibility of Speech," *The Journal of the Acoustical Society of America*, vol. 28, no. 6, pp. 1275–1277, Nov. 1956.

[3] Shigeo Morishima, K Aizawa, and H Harashima, "An intelligent facial image coding driven by speech and phoneme," in *International Conference on Acoustics, Speech, and Signal Processing, (ICASSP-89)*, 1989, pp. 1795–1798.

[4] S Morishima and H Harashima, "Speech-to-image media conversion based on VQ and neural network," in *International Conference on Acoustics, Speech, and Signal Processing, (ICASSP-91)*, 1991, pp. 2865–2868.

[5] Ram Rao and Tsuhan Chen, "Cross-modal prediction in audio-visual communication.," *International Conference on Acoustics, Speech, and Signal Processing, (ICASSP-96)*, vol. 4, pp. 2056–2059, 1996.

[6] Tomoki Toda, Alan W Black, and Keiichi Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, Mar. 2008.

[7] E Yamamoto, S Nakamura, and K Shikano, "Lip movement synthesis from speech based on Hidden Markov Models," *Speech Communication*, vol. 26, no. 1-2, pp. 105–115, Oct. 1998.

[8] Lijuan Wang, Yao Qian, M R Scott, Gang Chen, and F K Soong, "Computer-Assisted Audiovisual Language Learning," *Computer*, vol. 45, no. 6, pp. 38–47, 2012.

[9] B Denby, T Schultz, K Honda, T Hueber, J M Gilbert, and J S Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, Apr. 2010.

[10] Karen M Hiiemae and Jeffrey B Palmer, "Tongue Movements in Feeding and Speech," *Critical Reviews in Oral Biology & Medicine*, vol. 14, no. 6, pp. 413–429, Nov. 2003.

[11] Jintao Jiang, Abeer Alwan, Patricia A Keating, Edward T Auer, and Lynne E Bernstein, "On the relationship between face movements, tongue movements, and speech acoustics," *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1174–1188, 2002.

[12] M Celce-Murcia, D M Brinton, and J M Goodwin, "Teaching pronunciation: A reference for teachers of English to speakers of other languages," 1996.

[13] Pierre Badin, Yuliya Tarabalka, Frédéric Elisei, and Gérard Bailly, "Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding," *Speech Communication*, vol. 52, no. 6, pp. 493–503, June 2010.

[14] Xinyuan Zheng, Jianguo Wei, Wenhuan Lu, Qiang Fang, and Jianwu Dang, "Mapping between ultrasound and vowel speech using DNN framework," in *2014 9th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. 2014, pp. 372–376, IEEE.

[15] T Hueber, G Aversano, G Chollet, B Denby, G Dreyfus, Y Oussar, P Roussel, and M Stone, "Eigentongue Feature Extraction for an Ultrasound-Based Silent Speech Interface," in *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP 2007)*, 2007.

[16] Thomas Hueber, Elie-Laurent Benaroya, Gérard Chollet, Bruce Denby, Gérard Dreyfus, and Maureen Stone, "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips," *Speech Communication*, vol. 52, no. 4, pp. 288–300, Apr. 2010.