SYNTHESIZING VISUAL SPEECH TRAJECTORY WITH MINIMUM GENERATION ERROR

Lijuan Wang[1], Yi-Jian Wu[1], Xiaodan Zhuang[12], Frank K. Soong[1]

[1]Microsoft Research Asia, Microsoft Corporation, Beijing, China
[2]Beckman Institute, ECE Department, University of Illinois at Urbana-Champaign, U.S.A.
{lijuanw, yijiwu, frankkps}@microsoft.com, xzhuang2@uiuc.edu

ABSTRACT

In this paper, we propose a minimum generation error (MGE) training method to refine the audio-visual HMM to improve visual speech trajectory synthesis. Compared with the traditional maximum likelihood (ML) estimation, the proposed MGE training explicitly optimizes the quality of generated visual speech trajectory, where the audio-visual HMM modeling is jointly refined by using a heuristic method to find the optimal state alignment and a probabilistic descent algorithm to optimize the model parameters under the MGE criterion. In objective evaluation, compared with the ML-based method, the proposed MGE-based method achieves consistent improvement in the mean square error reduction, correlation increase, and recovery of global variance. It also improves the naturalness and audio-visual consistency perceptually in the subjective test.

Index Terms— visual speech synthesis, photo-real, talking head, trajectory-guided, minimum generation error

1. INTRODUCTION

Talking heads are useful in applications of human-machine interaction, e.g. reading emails, news or eBooks, acting as an intelligent voice agent or a computer assisted language teacher, etc. A lively, lip sync talking head can attract the attention of a user, make the human/machine interface more engaging or add entertainment ingredients to an application. Our motivation is to build a photo-real talking head where the facial animation is video realistic: that is, we desire our talking head to look as much as possible as if it were a video camera recording of a human subject, and not that of a cartoon-like character. In this work, we choose to focus our efforts on the issues related to the synthesis of the visual speech stream (including lips, teeth, and tongue), which is the most eye-catching region on a talking face.

To synthesize visual speech animations from audio-video parallel data, various approaches have been proposed before, like: key-frame based interpolation [1], unit selection synthesis [2], HMM-based synthesis [3,4], and the recently proposed hybrid approach using the HMM predicted trajectory to guide the unit selection. In [11], we proposed the trajectory-guided real sample concatenating method for generating lip-synced articulator movements for a photo-real talking head. In particular, in training stage, an audio/visual database is recorded and used to train a statistical Hidden Markov Model (HMM). In synthesis, the trained HMM is used to generate visual parameter trajectory in maximum likelihood sense first. Guided by the HMM predicted trajectory, a succinct and smooth lips sample sequence is searched from the image sample library optimally and the lips sequence is then stitched back to a background head video.

For both HMM-based parametric and HMM-guided hybrid approaches, the statistically trained HMM is crucial since the HMM predicted visual trajectories to a large extent determine how good the visual lips can be rendered. In our previous work [11], we use the maximum likelihood (ML) based estimation for the audio-visual joint HMM training. One noticeable observation is the constrained mouth movement in a much smaller dynamic range compared with the recorded one, which is mainly due to the over-smoothed visual trajectory generated from the ML-trained HMM. Since the ML-based training does not explicitly optimize the quality of generated trajectory, an audio-visual HMM with maximum likelihood on the training data does not necessarily result in generated visual trajectories that have minimized error in human perception.

In order to address the above issue, we propose to use the minimum generated visual trajectory error approach to improve visual speech synthesis. Inspired by minimum generation error (MGE) training in HMM-based speech synthesis [9,10], we propose further refining the model parameters by minimizing the mean square errors between the generated visual trajectories and the real ones using a probabilistic descent (PD) algorithm.

We incorporated the MGE training into our HMM trajectory-guided photo-real talking head rendering system. Evaluated on the LIPS 2009 Visual Speech Synthesis Challenge task [8], the MGE approach results in improved visual speech synthesis in both objective metric and subjective perception.

The rest of the paper is organized as follows. Section 2 gives an overview of the ML-based synthesis framework. Section 3 proposes the MGE-based model refinement for visual speech trajectory synthesis. Section 4 discusses the experimental results, and section 5 draws the conclusions.

2. ML-BASED VISUAL SPEECH TRAJECTORY SYNTHESIS

The HMM-based speech synthesis has made a steady but significant progress in the last decade [5]. The approach was also been tried for visual speech synthesis [3,4]. In HMM-based visual speech synthesis, audio and video are jointly modeled in HMMs and the visual parameters are generated from HMMs by using the dynamic ("delta") constraints of the features [3]. We propose using the same method in ML-based framework.

We use acoustic vectors $A_t = [a_t^T, \Delta a_t^T, \Delta\Delta a_t^T]^T$ and visual vectors $V_t = [v_t^T, \Delta v_t^T, \Delta\Delta v_t^T]^T$ which is formed by augmenting the static features and their dynamic counterparts to represent the audio and video data. Audio-visual HMMs, λ, are trained by

maximizing the joint probability $p(A,V|\lambda)$ over the stereo data of acoustic and visual feature training vectors. In order to capture the contextual effects, context dependent HMMs are trained and tree-based clustering is applied to acoustic and visual feature streams separately to improve the corresponding model robustness. For each AV HMM state, a single Gaussian mixture model (GMM) is used to characterize the state output. The state q has a mean vectors $\mu_q^{(A)}$ and $\mu_q^{(V)}$. In this paper, we use the diagonal covariance matrices for $\sum_q^{(AA)}$ and $\sum_q^{(VV)}$, null covariance matrices for $\sum_q^{(AV)}$and$\sum_q^{(VA)}$, by assuming the independence between audio and visual streams and between different components.

Given a continuous audio-visual HMM λ, and acoustic feature vectors $A = [A_1^\top, A_2^\top, \cdots, A_T^\top]^\top$, we use the following algorithm to determine the best visual parameter vector sequence $V = [V_1^\top, V_2^\top, \cdots, V_T^\top]^\top$ by maximizing the following likelihood function.

$$p(V|A,\lambda) = \sum_{\text{all } Q} p(Q|A,\lambda) \cdot p(V|A,Q,\lambda), \quad (1)$$

is maximized with respect to V, where $Q = [q_1, q_2, \ldots, q_T]$ is the state sequence.

At frame t, $p(V_t|A_t, q_t, \lambda)$ are given by

$$p(V_t|A_t, q_t, \lambda) = N\left(V_t; \hat{\mu}_{q_t}^{(V)}; \hat{\sum}_{q_t}^{(VV)}\right), \quad (2)$$

$$\hat{\mu}_{q_t}^{(V)} = \mu_{q_t}^{(V)} + \sum_{q_t}^{(VA)} \sum_{q_t}^{(AA)^{-1}} \left(A_t - \mu_{q_t}^{(A)}\right), \quad (3)$$

$$\hat{\sum}_{q_t}^{(VV)} = \sum_{q_t}^{(VV)} - \sum_{q_t}^{(VA)} \sum_{q_t}^{(AA)^{-1}} \sum_{q_t}^{(AV)}. \quad (4)$$

The complexity of solving Eq. (1) can be significantly reduced by the following two reasonable approximations.

First, the summation over all state and mixture components in Eq. (1) can be approximated with a single state sequence,

$$p(V|A,\lambda) \approx p(\hat{Q}|A,\lambda) \cdot p(V|A,\hat{Q},\lambda) \quad (5)$$

where $\hat{Q}$ is the optimal aligned state sequence by maximizing the likelihood function $\hat{Q} = \text{argmax}_Q p(Q|A,\lambda)$. With Eq. (5), the optimal visual trajectory $\hat{v} = argmax_v p(Wv|A,\hat{Q},\lambda)$ can then be solved in a closed least square solution [5] by setting $\frac{\partial}{\partial c} logp(V|A,\hat{Q},\lambda) = 0$.

$$\hat{v} = \left(W^\top \hat{U}^{(VV)^{-1}} W\right)^{-1} W^\top \hat{U}^{(VV)^{-1}} \hat{\mu}^{(V)}, \quad (6)$$

where W is a transformation matrix and

$$\hat{\mu}^{(V)} = \left[\hat{\mu}_{q_1}^{(V)}, \hat{\mu}_{q_2}^{(V)}, \cdots, \hat{\mu}_{q_T}^{(V)}\right]^\top, \quad (7)$$

$$\hat{U}^{(VV)^{-1}} = \text{diag}\left[\hat{\sum}_{q_1}^{(VV)^{-1}}, \hat{\sum}_{q_2}^{(VV)^{-1}}, \cdots, \hat{\sum}_{q_T}^{(VV)^{-1}}\right]^\top. \quad (8)$$

Second, in calculating Eq. (3) and Eq. (4), we may further simplify the problem by assuming $\sum_{q_t}^{(VA)} = 0$.

Given a state mixture component q, the full covariance matrix in the joint space of V and A can be partitioned into $\sum_q^{(VV)}$, $\sum_q^{(VA)}$, $\sum_q^{(AA)}$ and $\sum_q^{(AV)}$. In many cases where training data is not abundant, it is not easy to obtain robust estimation of all the elements in these matrices. When the two signals are in the same feature space, the full covariance matrices are usually approximated with diagonal matrices. In audio-visual modeling, however, A and V are in different spaces with no strong correlation between the corresponding dimensions. Therefore, we only estimate $\sum_q^{(VV)}$ and $\sum_q^{(AA)}$, yielding the simplified Eq. (9) and Eq. (10).

$$\hat{\mu}_{q_t}^{(V)} \approx \mu_{q_t}^{(V)}, \quad \hat{\sum}_{q_t}^{(VV)} \approx \sum_{q_t}^{(VV)} \quad (9)$$

$$\hat{v} = \left(W^\top U^{(VV)^{-1}} W\right)^{-1} W^\top U^{(VV)^{-1}} \mu^{(V)}. \quad (10)$$

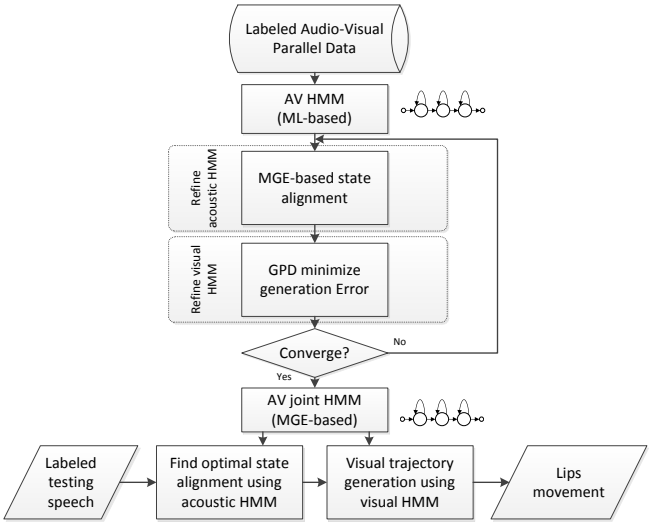

Fig.1: MGE-based visual speech synthesis system.

3. REFINING AV HMM MODELING WITH MGE

The ML-based approach is effective and outperforms previous methods. However, maximum likelihood training does not optimize directly towards visual generation error. In particular, an audio-visual HMM with maximum likelihood for the training data does not lead to generated visual trajectories with minimized error.

Similar problems exist for ML-based speech synthesis. To compensate this deficiency, a minimum generation error (MGE) [9] criterion had been proposed for HMM training. In particular, an appropriate generation error is defined, which is minimized by using a probabilistic descent (PD) algorithm to update the parameters of the HMMs.

We propose the Minimum Generated Trajectory Error (MGE) method to further refine the audio-visual joint modeling by minimizing the error between the generation result and the real target trajectories in the training set.

3.1. System framework with MGE

The proposed visual speech synthesis system with MGE is illustrated in Fig. 1. With the approximation using the optimal aligned state sequence adopted in Eq. (5), the visual generation problem becomes the following two steps. First, given the sequence of audio features, the optimal aligned state sequence $\hat{Q}$ is determined by forced alignment with acoustic HMMs. Second, given the optimal aligned state sequence, the visual trajectories are estimated in a maximum probability sense using the visual HMMs. Under MGE criterion, AV HMMs are jointly refined to minimize the visual stream generation error. In particular, acoustic HMM parameters are updated towards generating MGE preferred state alignment. Meanwhile, visual HMM parameters are refined explicitly towards minimizing generation error by using a PD algorithm. The two refinement steps are conducted iteratively until the generation error reduction converges on training data.

3.2. State alignment under MGE criterion

Under the MGE criterion, the optimal state sequence determined by audio feature input should be in the sense of minimizing visual generation error, which is:

$$\hat{Q} = argmin_Q e(\lambda) = argmin_Q \frac{1}{N} \sum_{i=1}^{N} D\left(y^i, \hat{y}^i(Q,\lambda)\right) \quad (11)$$

However, the parameter generation process depends on the whole state sequence, which makes it intractable to search for the optimal state sequence directly. Therefore, we adopt the same heuristic method in [10] to search for the optimal state sequence under MGE criterion and then re-estimate the acoustic model. The new state alignment with the refined acoustic model gets closer to the MGE preferred state boundary and yields reduced generation error. The process is as follows:

1) Initialize the state alignment for the input utterance by forced aligning the audio feature with the acoustic HMM using Viterbi search algorithm;
2) For each state boundary in the state alignment, try to perturb it to the left (or right), and calculate the visual trajectory generation errors before and after shifting the state boundary;
3) If the generation errors decrease, keep the new state boundary and go back to step 2); otherwise go to step 4);
4) Update the acoustic HMM parameters with the new state alignment;

3.3. Refined visual modeling

We define the visual generation error as the Euclidean distance between PCA vectors of the synthesis result and the real visual trajectory in training set,

$$D(y,\hat{y}) = \sum_{t=1}^{T} \|y_t - \hat{y}_t\| \tag{12}$$

Therefore, we can improve the generation performance by minimizing the empirical generation error, measured using a cost function $e(\lambda)$ similar to MGE in speech synthesis [9].

$$e(\lambda) = \frac{1}{N}\sum_{i=1}^{N} D\big(y^i, \hat{y}^i(\hat{Q},\lambda)\big), \tag{13}$$

where N is the number of training utterances.

Once the optimal aligned state sequence $\hat{Q}$ is determined, the visual HMM parameters can be refined by using the PD algorithm. Given the n^{th} training utterance, the updating rule for the parameters of the visual state sequence is:

$$\lambda(n+1) = \lambda(n) - \varepsilon_n \frac{\partial}{\partial\lambda} D\left(y^n, \hat{y}^n(\hat{Q},\lambda)\right)|_{\lambda=\lambda(n)} \tag{14}$$

where ε_n is a learning rate that decrease when the utterance index n increases.

$$\frac{\partial}{\partial\lambda} D\left(y^n, \hat{y}^n(\hat{Q},\lambda)\right) = 2(\hat{y}^n(\hat{Q},\lambda) - y^n)^T \frac{\partial}{\partial\lambda}\hat{y}^n(\hat{Q},\lambda). \tag{15}$$

In particular, with Eq. (10), the updating rule for the mean vector is

$$\mu_{i,j}(n+1) = \mu_{i,j}(n) - 2\varepsilon_n\big(\hat{y}^n(\hat{Q},\lambda) - y^n\big)^T \left(W^\top U^{(VV)^{-1}} W\right)^{-1} W^T U^{(VV)^{-1}} Z_\mu \tag{16}$$

where $Z_\mu = [0,\dots,0,1_{i\times D_Y+j},0,0,\dots,0]^T$.

Similarly, we denote $\upsilon_{i,j} = 1/\sigma^2_{i,j}$ and $Z_\upsilon = Z_\mu Z_\mu{}^T$, the updating rule for the covariance parameter is

$$\upsilon_{i,j}(n+1) = \upsilon_{i,j}(n) - 2\varepsilon_n\big(\hat{y}^n(\hat{Q},\lambda) - y^n\big)^T \left(W^\top U^{(VV)^{-1}} W\right)^{-1} W^T Z_\upsilon \left(\mu^{(V)} - W\hat{y}^n(\hat{Q},\lambda)\right). \tag{17}$$

4. EXPERIMENTAL RESULTS

4.1. Experimental setup

We employ the LIPS 2008/2009 Visual Speech Synthesis Challenge data [8] to evaluate the proposed MGE-based method. This dataset has 278 video files with corresponding audio track, each being one English sentence spoken by a single native speaker with neutral emotion. The video frame rate is 50 frames per sec. For each image, Principle Component Analysis (PCA) projection is performed on automatically detected and aligned mouth image, resulting in a 60-dimensional visual parameter vector. Mel-Frequency Cepstral Coefficient (MFCC) vectors are extracted with a 20ms time window shifted every 5ms. The visual parameter vectors are interpolated up to the same frame rate as the MFCCs. The A-V feature vectors are used to train the ML-based HMM models using HTS 2.1[5].

4.3. Objective evaluation results

In objective evaluation, we measured the performance quantitatively of the ML-based synthesis method (baseline) and the proposed MGE-based method using mean square error (MSE), correlation coefficient, and global variance as defined in Eq. (18)-(20). In open test, leave-20-out cross validation is adopted to avoid data insufficiency problem.

$$MSE = \left\|\hat{V} - V\right\| = \frac{1}{T}\sum_{t=1}^{T}\left\|\hat{V}_t{}^\top - V_t{}^\top\right\| \tag{18}$$

$$ACC = \rho(V,\hat{V}) = \frac{1}{T\cdot d}\sum_{t=1}^{T}\sum_{i=1}^{d}\frac{\big(V_{t,i} - \mu_{V_i}\big)\big(\hat{V}_{t,i} - \mu_{\hat{V}_i}\big)}{\sigma_{V_i}\sigma_{\hat{V}_i}} \tag{19}$$

$$GV = \bar{\sigma}_V = \frac{1}{T.d}\sum_{t=1}^{T}\sum_{i=1}^{d}\big(V_{t,i} - \mu_{V_i}\big)^2 \tag{20}$$

Fig.2 shows the MSE results of the ML-based baseline system and the new MGE-based method. The left two bars represent the total MSE of all the PCA components, the rest bars show the MSE of the first four PCA components respectively. We can see that the proposed MGE-based method improves the synthesis results, especially for the first few dimensions of the PCA vector. We observe similar result in ACC as shown in Fig.3. The MGE-based method improves the overall correlation comparing with the baseline, especially for the first three dimensions of the visual PCA vector. Fig.4 shows the global variance of the real trajectories, the ML-based synthesized trajectories, and the MGE-based synthesized ones. Comparing with the results of real visual trajectory, the GV is dramatically reduced in ML-based training due to the statistical averaging. The MGE-based method can recover the GV and make it closer the natural one. In Fig.5, we illustrate the synthesized trajectory by the MGE method and by the ML-based method, respectively. The proposed MGE method is shown to result in trajectories more similar to the ground truth which a human speaker produces.

4.4. Subjective evaluation results

A subjective MOS "scoring" test is also carried out to compare the ML-based baseline, the proposed MGE-based approach and the original recording. We select twelve sentences from the LIPS 2009 test set, where each is constructed by a sequence of words but in a semantically meaningless order. These sentences are converted into video clips of the lower part of the face using each method. The original recordings cropped to the same area and the synthesis results are randomly assigned into six subjective test sessions, such that each session has two sentences from each method or the original recording. Each video clip also includes the ground truth input speech audio. The subjects are asked to score the perceived "audio-visual consistency" on a 1-5 basis for each sentence in each session. Each session is evaluated by three different subjects.

Fig. 6 shows the averaged subjective scores for "audio-visual consistency". The MGE-based method improves the perceptual

naturalness of the mouth movement. In particular, it increases the mouth dynamic range and makes the photo-real talking head speaks like a real human. More rendered full face videos can be found on http://dict.bing.com.cn/, where a talking head English teacher is built to read the text sample sentences on the website.

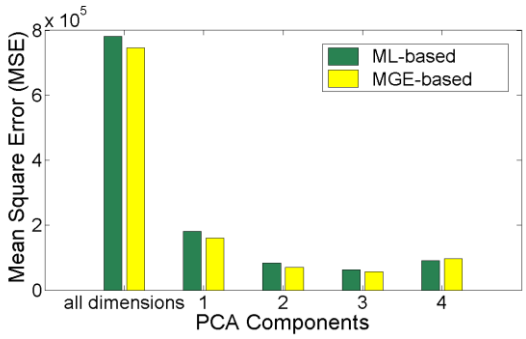

Fig. 2: Mean square error (MSE) of the synthesized PCA vectors.

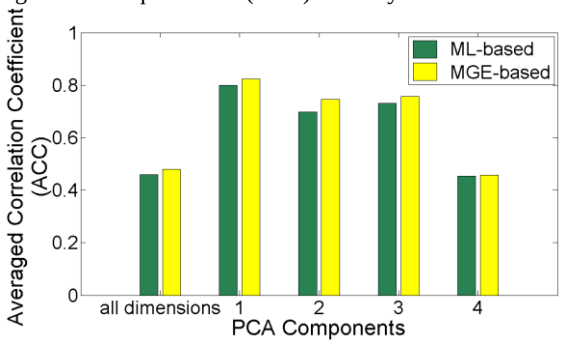

Fig. 3: Averaged correlation coefficient (ACC) of the synthesized PCA vectors.

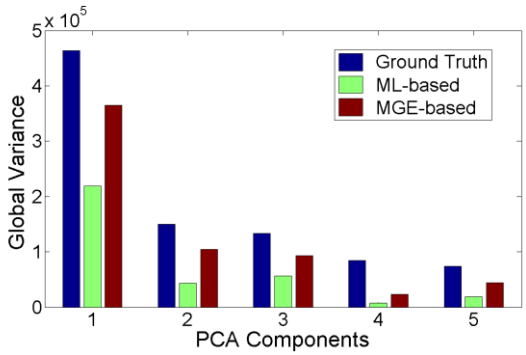

Fig. 4: Global variance of each PCA components.

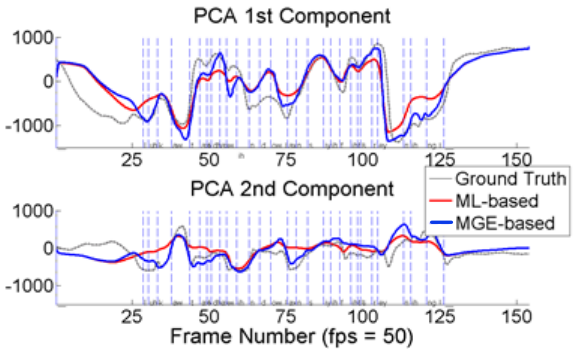

Fig. 5: Synthesized trajectories of the 1st, 2nd PCA components.

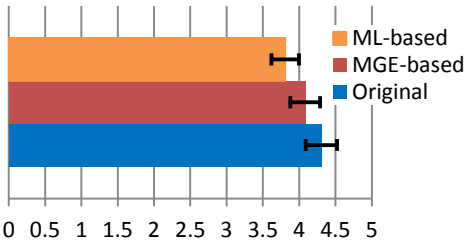

Fig. 6: Subjective scores for "audio-visual consistency" (with standard errors)

5. CONCLUSIONS

We propose the minimum generation error training method to refine the audio-visual HMM for improved visual speech trajectory synthesis. Under the MGE criterion, the joint audio-visual HMM modeling is refined with a heuristic method for finding the optimal state alignment and the generalized probabilistic descent algorithm. In objective evaluation, comparing with the ML-based method, we get consistent improvement in the mean square error reduction and increase of correlation, and recovery of global variance. Perceptually it increases the mouth dynamic range and makes the photo-real talking head speaks like a real human.

6. REFERENCES

[1] C. Bregler, M. Covell, M. Slaney, "Video Rewrite: Driving Visual Speech with Audio," In Proc. ACM SIGGRAPH 97, Los Angeles, CA, 1997, pp. 353-360.

[2] E. Cosatto and H.P. Graf, "Photo-realistic talking heads from image samples", IEEE Trans. Multimedia, 2000, vol. 2, no. 3, pp. 152-163.

[3] S. Sako, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "HMM-based text-to-audio-visual speech synthesis," Proc. ICSLP2000, vol. 3, 2000, pp. 25–28.

[4] L. Xie, Z.Q. Liu, "Speech Animation Using Coupled Hidden Markov Models," Proc. ICPR'06, August 2006, pp. 1128-1131.

[5] K. Tokuda, H. Zen, etc., "The HMM-based speech synthesis system (HTS)," http://hts.ics.nitech.ac.jp/.

[6] S. Fu, etc., "Audio/visual mapping with cross-modal hidden markov models," IEEE Transactions on Multimedia, vol. 7, no. 2, pp. 243-252, April 2005.

[7] T. Toda, A. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.

[8] B.-J. Theobald, S. Fagel, G. Bailly, and F. Elisei, "LIPS 2008: Visual speech synthesis challenge," Proc. Interspeech2008, pp. 2310–2313.

[9] Y.-J. Wu and R.-H. Wang, "Minimum generation error training for HMM-based speech synthesis," Proc. ICASSP 2006, Toulouse, France, May 2006, vol. I, pp. 89-92.

[10] Y.-J. Wu, L. Qin, K. Tokuda, "An Improved Minimum Generation Error based Model Adaptation for HMM-based Speech Synthesis," Proc. Interspeech2009, Brighton, UK, Sept. 2009, pp.1787-1790.

[11] L.-J. Wang, X.-J. Qian, W. Han, and F. Soong, "Synthesizing photo-real talking head via trajectory-guided sample selection," Interspeech2010, Chiba, Japan, Sept. 2010, pp.446-449.