Rendering A Personalized Photo-Real Talking Head from Short Video Footage

Lijuan Wang[1], *Wei Han*[1,2], *Xiaojun Qian*[1,3], *Frank K. Soong*[1]

[1] Microsoft Research Asia, Beijing
[2] Department of Computer Science, Shanghai Jiao Tong University, Shanghai
[3] Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong, Hong Kong
{lijuanw, frankkps}@microsoft.com, weihan@live.com, xjqian@se.cuhk.edu.hk

Abstract

In this paper, we propose an HMM trajectory-guided, real image sample concatenation approach to photo-real talking head synthesis. An audio-visual database of a person is recorded first for training a statistical Hidden Markov Model (HMM) of Lips movement. The HMM is then used to generate the dynamic trajectory of lips movement for given speech signals in the maximum probability sense. The generated trajectory is then used as a guide to select, from the original training database, an optimal sequence of lips images which are then stitched back to a background head video. The whole procedure is fully automatic and data driven. For as short as 20 minutes recording of audio/video footage, the proposed system can synthesize a highly photo-real talking head in sync with the given speech signals (natural or TTS synthesized). This system won the first place in the A/V consistency contest in LIPS Challenge(2009), perceptually evaluated by recruited human subjects.

Index Terms: visual speech synthesis, photo-real, talking head, trajectory-guided

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, a computer assisted language teacher or a virtual talk show host in a TV program, etc. A lively, lip sync talking head can attract the attention of a user, make the human/machine interface more engaging or adds entertainment ingredients to an application. The major drawbacks of existing systems are lack of realistic appearance and personalized audio/visual features. An ideal personalized talking head should be able to replace a person with a virtual agent while mimicking the person's voice, facial appearance and animations, and speak any sentences he/she had never spoken before.

Generating animated talking heads that look like real people is challenging. The existing approaches to talking heads use either image-based 2D models [1,2] or geometry-based 3D ones [32,33]. 3D models provide more power to construct the head in any given view, but they are hard to build and usually lack the realistic appearance even after texture mapping. Image-based approaches have their advantages that the photo realistic appearance is guaranteed. A talking head needs to be not just photo-realistic in a static appearance, but exhibit convincing plastic deformations of the lips synchronized with the corresponding speech, realistic head movements and natural facial expressions. In this paper, we introduce the whole system for constructing personalized photo-real talking head from video footage, and focus on the articulator movements (including lips, teeth, and tongue), which is the most eye-catching region on a talking face.

Various approaches have been proposed before for synthesizing realistic talking head from video training data, roughly in three categories: key-frame based interpolation, unit selection synthesis and HMM-based synthesis.

The key-frame-based interpolation method [2] is based upon morphing between 2-D key-frame images. The most frequently used key-frame set is visemes (visual phonemes), which form a set of images spanning a large range of mouth shapes. Using morphing techniques, the transitions from one viseme to other viseme can be computed and interpolated automatically.

The unit selection, or sample-based method starts with collecting representative samples. The samples are then parameterized by its contextual label information so that they can be recalled according to the target context information in synthesis. Typically, minimal signal processing is performed to avoid introducing artifacts or distortions unnecessarily. Video snippets of tri-phone have been used as basic concatenation units [3-5]. Since these video snippets are parameterized with phonetic contextual information, the resulting database can become too large. Smaller units like image samples have shown their effectiveness in improving the coverage of candidate units. In LIPS2008 Challenge, Liu demonstrated a photo-real talking head [6] in a sample-based approach, which is an improved version of the original work of Cosatto and Graf [1].

The Hidden Markov Model (HMM) based speech synthesis has made a steady but significant progress in the last decade [7]. The approach was also tried for visual speech synthesis [8,9]. 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 [8]. Convincing mouth video can be rendered from the predicted visual parameter trajectories. One drawback of the HMM-based visual speech synthesis method is its blurring due to feature dimension reduction in PCA and the maximum likelihood-based statistical modeling. Therefore, further improvement is still needed to make a high quality, photo-real talking head.

Inspired by the newly proposed HMM-guided unit selection method in speech synthesis [10,11], we propose 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, 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.

ISBN 978-1-4244-6245-2/10/$26.00 ©2010 IEEE

This paper is organized as follows. Section 2 gives an overview of whole procedure of creating personalized talking head for a new speaker. Section 3 proposes the HMM trajectory-guided sample selection method. Section 4 discusses the experimental results, and section 5 draws the conclusions.

2. Personalized Photo-Real Talking Head

2.1. System framework

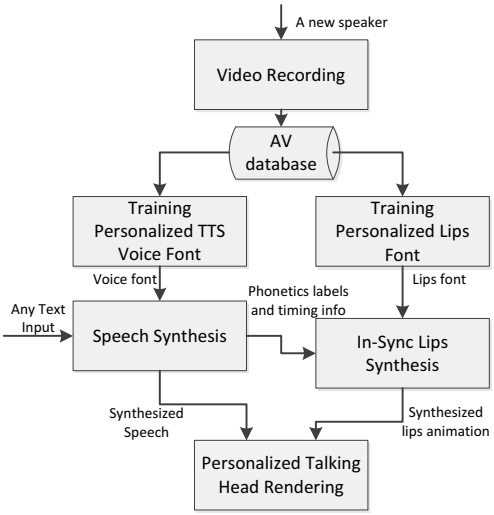

Fig. 1. The whole procedure of creating personalized talking head for a new speaker.

The system synthesizes a personalized photo-real talking head for a new speaker by using a small amount of recorded video footage. As Fig.1 shows, to enroll a new speaker, some frontal view video of this person is firstly required as the training data. In our proposed method, about 20-minute video footage is enough to build a high-quality talking head. The recorded audio-visual database is used to train the lips animation model and to build the speaker's personalized lips font. Meanwhile, the audio stream is used to train the text-to-speech (TTS) voice from scratch or to adapt a well-trained TTS voice to this specific speaker. At the synthesis stage, for any given text, the audio is firstly synthesized by TTS. Then, the video is generated synchronously with the lip movement. The final synthesized talking head (video + audio) can mimic the specific speaker with high fidelity by speaking any sentences he/she had never spoken before. Using this framework, we've built a few talking heads for English and Chinese speakers. The demonstration videos of our synthesis results can be found at: http://research.microsoft.com/en-us/projects/photo-real_talking_head/.

2.2. Video recording and pre-processing

Training the talking head of a speaker requires about 20-minute audio-visual recording of the speaker in frontal view reciting some prompted sentences. The sentences chosen for recording should have good phonetic coverage and contextual diversity, and be spoken in a neutral style. The lighting should ensure the visible articulators clear in the video and avoid any shadow on the face. The speaker can naturally move or rotate his/her head when speaking.

Since the speaker naturally moves/rotates his/her head during the recording, we need to do head pose normalization among all the raw images. By using the 3D model-based head pose tracking technology [9], as shown in Fig. 2&3, the Euler angels of the head on three dimensions as well as the translation are obtained. Given the 3D rotations and translations of head poses, every frame can be normalized to a full frontal view and facial aligned by applying an affine transform. The lips images are cropped out by using a fixed rectangle window on the normalized frames and a mouth image sample library is formed.

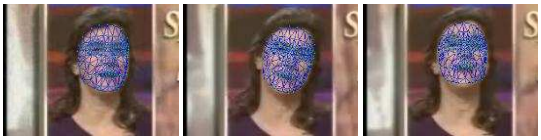

Fig. 2. Capturing 3D head motions from a video clip.

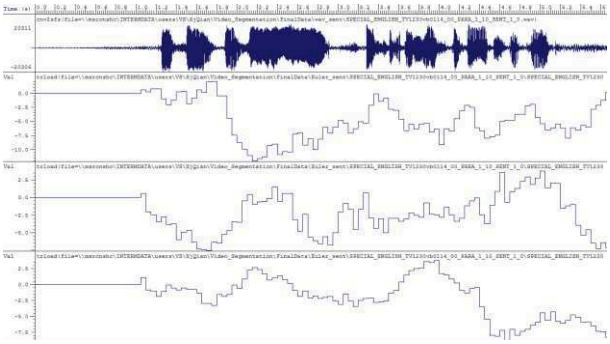

Fig. 3. 3D Head rotations (in Euler angles) during speaking ("This is the VOA special English development report.").

2.3. Lips animation synthesis

We propose the HMM trajectory guided sample selection method for lips animation synthesis, which consists of two, training and synthesis, stages.

In the training stage, audio/visual footage of a speaker is used to train the statistical audio-visual Hidden Markov Model (AV-HMM). The input of the HMM contains both the acoustic features and the visual features. The acoustic features consist of Mel-frequency cepstral coefficients (MFCCs), their delta and delta-delta coefficients. The visual features include the PCA coefficients and their dynamic features., The contextual dependent HMM is used to capture the variations caused by different contextual features. Also, the tree-based clustering technique is applied to the acoustic and visual features respectively to improve the robustness of the HMM.

In the synthesis stage, the input phoneme labels and alignments are firstly converted to a context-dependent label sequence. Meanwhile, the decision trees generated in the training stage are used to choose the appropriate clustered state HMMs for each label. Then parameter generation algorithm is used to generate the visual parameter trajectory in maximum likelihood sense. The HMM predicted trajectory is used as guidance for selecting a succinct mouth sample sequence from the image library. Finally the mouth image sequence is stitched onto the background head video.

2.4. Post-processing for full face video generation

The remaining task is to stitch the lips image sequence into a full face background sequence [1]. Local deformations are required to stitch the shape of the mouth and jaw line correctly and also to avoid the unsmooth problem when the stitches are across the jaw line. After local deformation around the jaw line, the final stitching process is done by Poisson image editing. Poisson image editing [25] permits the seamless editing and cloning of a selected region from source image into a destination. We use a mouth replacement mask to

specify which region of the final video come from the selected lips image sequence and which come from the background video. Fig. 4 shows an example mouth replacement mask, applied to lips images and background images. In a similar way, we replace the upperface region like the eyes and eye brow by using an eye mask for rendering eye blink. To restore the natural head movement in the background sequence, these images are been transformed to its original head pose. The final rendered video is photo-, video-realistic, lip sync with speech, also with natural head motion.

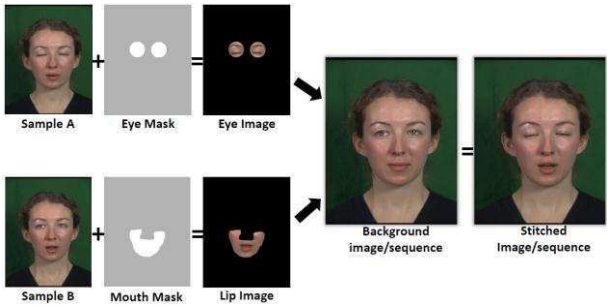

Fig. 4. Illustration of image stitching process.

3. HMM Trajectory-Guided Sample Selection

This section introduces the proposed HMM trajectory-guided sample selection approach. In training, first the original lip image samples $\boldsymbol{S}$ are encoded in low-dimensional visual feature vector $\boldsymbol{V}$. Then the visual features $\boldsymbol{V}$ along with the acoustic features $\boldsymbol{A}$ are used to train statistical HMM model $\boldsymbol{\lambda}$. In synthesis, for any arbitrary natural or Text-to-Speech (TTS) synthesized speech input $\boldsymbol{A}$, the trained model $\boldsymbol{\lambda}$ generates the optimal feature trajectory $\widehat{\boldsymbol{V}}$ in the maximum likelihood sense. The last step is to reconstruct $\widehat{\boldsymbol{V}}$ back to $\widehat{\boldsymbol{S}}$ in the original sample space by the proposed HMM trajectory-guided real sample selection method, so that the synthesis results can be seen. In particular, guided by the HMM predicted trajectory $\widehat{\boldsymbol{V}}$, a succinct and smooth image sample sequence $\widehat{\boldsymbol{S}}$ is searched optimally from the sample library and the mouth sequence is then stitched back to a background head video. To put it briefly, there are four main modules: $\Rightarrow \boldsymbol{V}$; $(\mathbf{A}, \boldsymbol{V}) \Rightarrow \boldsymbol{\lambda}$; $(\boldsymbol{\lambda}, \boldsymbol{A}) \Rightarrow \widehat{\boldsymbol{V}}$; and $\widehat{\boldsymbol{V}} \Rightarrow \widehat{\boldsymbol{S}}$.

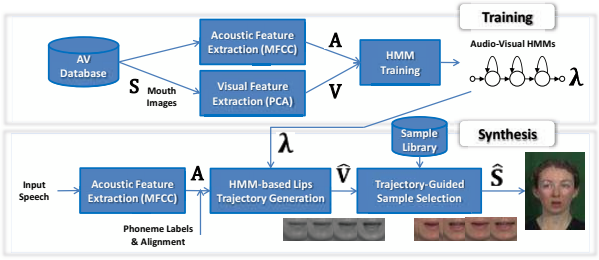

Fig. 5. HMM trajectory-guided sample selection.

3.1. Visual parameter extraction ($\boldsymbol{S} \Rightarrow \boldsymbol{V}$)

We obtain eigen-lips (eigenvectors of the lips images) by applying PCA to all the lips images. The top 20 eigen-lips contained about 90% of the accumulated variance. The visual feature of each lips image is formed by its PCA vector,

$$V^{\top} = S^{\top} W \tag{1}$$

where W is the projection matrix made by the top 20 eigen-lips.

3.2. Audio-Visual HMM modeling ($\mathbf{A}, \boldsymbol{V} \Rightarrow \boldsymbol{\lambda}$)

We use acoustic vectors $A_t = [a_t^{\top}, \Delta a_t^{\top}, \Delta\Delta a_t^{\top}]^{\top}$ and visual vectors $V_t = [v_t^{\top}, \Delta v_t^{\top}, \Delta\Delta v_t^{\top}]^{\top}$ 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 MFCC(acoustic) and PCA(visual) 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.

3.3. Visual trajectory generation ($\boldsymbol{\lambda}, \boldsymbol{A} \Rightarrow \widehat{\boldsymbol{V}}$)

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.

$$p(V|A, \lambda) = \sum_{all\ Q}\ p(Q|A, \lambda) \cdot p(V|A, Q, \lambda), \tag{2}$$

is maximized with respect to m, where Q is the state sequence.

At frame t, $\mathrm{p}(m_t|\mathrm{A_t}, \mathrm{q_t}, \lambda)$ are given by

$$p(V_t|A_t, q_t, \lambda) = N\left(V_t; \hat{\mu}_{q_t}^{(V)}; \hat{\sum}_{q_t}^{(VV)}\right), \tag{3}$$

where

$$\hat{\mu}_{q_t}^{(V)} = \mu_{q_t}^{(V)} + \sum_{q_t}^{(VA)} \sum_{q_t}^{(AA)^{-1}} \left(A_t - \mu_{q_t}^{(A)}\right), \tag{4}$$

$$\hat{\sum}_{q_t}^{(VV)} = \sum_{q_t}^{(VV)} - \sum_{q_t}^{(VA)} \sum_{q_t}^{(AA)^{-1}} \sum_{q_t}^{(AV)}. \tag{5}$$

We only consider the optimal state sequence Q by maximizing the likelihood function $p(Q|A, \lambda)$ with respect to the given acoustic feature vectors A and model λ. Then, the logarithm of the likelihood function is written as

$$logp(V|A, Q, \lambda) = logp\left(V|\hat{\mu}^{(V)}, \hat{U}^{(VV)}\right)$$
$$= -\frac{1}{2} V^{\top} \hat{U}^{(VV)^{-1}} V + V^{\top} \hat{U}^{(VV)^{-1}} \hat{\mu}^{(V)} + K, \tag{6}$$

where

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

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

The constant K is independent of m. The relationship between a sequence of the static feature vectors $C = [v_1^{\top}, v_2^{\top}, \cdots, v_T^{\top}]^{\top}$ and a sequence of the static and dynamic feature vectors $m$can be represented as a linear conversion,

$$V = W_c C, \tag{9}$$

where W_c is a transformation matrix described in [7]. By setting$\frac{\partial}{\partial \mathrm{C}} \mathrm{logp}(m|\mathrm{A}, \mathrm{Q}, \lambda) = 0$, we obtain $\widehat{m}_{\mathrm{pt}}$ that maximizes the logarithmic likelihood function, as given by

$$\widehat{V}_{opt} = W_c C_{opt}$$
$$= W_c \left(W_c^{\top} \hat{U}^{(VV)^{-1}} W_c\right)^{-1} W_c^{\top} \hat{U}^{(VV)^{-1}} \hat{\mu}^{(V)}. \tag{10}$$

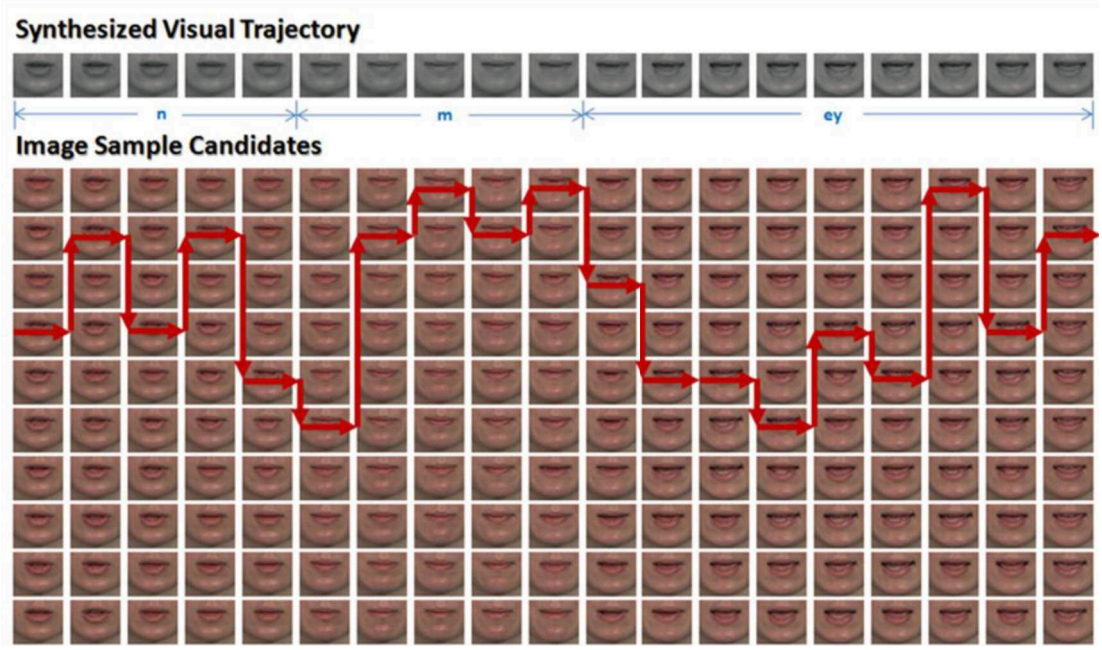

Fig. 6. Illustration for the HMM trajectory-guided sample selection approach. The top-line lips images (gray) are the HMM predicted visual trajectory. The bottom images (colored) are real samples lips candidates where the best lips sequence (red arrow path) is selected by Viterbi decoding.

3.4. Trajectory-Guided Sample Selection ($\hat{\boldsymbol{V}} \Rightarrow \hat{\boldsymbol{S}}$)

The HMM predicted visual parameter trajectory is a compact description of articulator movements, in the lower rank eigen-lips space. However, the lips image sequence shown at the top of Fig. 6 is blurred due to: (1) dimensionality reduction in PCA; (2) ML-based model parameter estimation and trajectory generation. To solve this blurring, we propose the trajectory-guided real sample concatenation approach to constructing $\hat{\boldsymbol{S}}$ from $\hat{\boldsymbol{V}}$. It searches for the closest real image sample sequence in the library to the predicted trajectory as the optimal solution. Thus, the articulator movement in the visual trajectory is reproduced and photo-real rendering is guaranteed by using real image sample.

3.4.1. Cost function

Like the unit selection in concatenative speech synthesis [29], the total cost for a sequence of T selected samples is the weighted sum of the target and concatenation costs:

$$C\left(\hat{V}_1^T, \hat{S}_1^T\right) = \sum_{i=1}^{T} \omega^t C^t(\hat{V}_i, \hat{S}_i)\ S\ \sum_{i=2}^{T} \omega^c C^c(\hat{S}_{i-1}, \hat{S}_i) \quad (11)$$

The target cost of an image sample $\hat{S}_i$ is measured by the Euclidean distance between their PCA vectors.

$$C^t(\hat{V}_i, \hat{S}_i) = \left\| \hat{V}_i - \hat{S}_i{}^{\top} W \right\| \quad (12)$$

The concatenation cost is measured by the normalized 2-D cross correlation (NCC) between two image samples $\hat{S}_i$ and $\hat{S}_j$, as Eq. 13 shows. Since the correlation coefficient ranges in value from -1.0 to 1.0, NCC is in nature a normalized similarity score, which is an advantage superior to other similarity metrics.

$$NCC(I,J) = \frac{\sum_{x,y}[I(x,y) - \bar{I}][J(x,y) - \bar{J}]}{\left\{\sum_{x,y}[I(x,y) - \bar{I}]^2 \sum_{x,y}[J(x,y) - \bar{J}]^2\right\}^{0.5}} \quad (13)$$

Assume that the corresponding samples of $\hat{S}_i$ and $\hat{S}_j$ in the sample library are S_p and S_q, i.e., $\hat{S}_i = S_p$, and $\hat{S}_j = S_q$, where, p and q are the sample indexes in video recording. And hence S_p and S_{p+1}, $S_{q-1}$and S_q are consecutive frames in the original recording. As defined in Eq. 14, the concatenation cost between $\hat{S}_i$ and $\hat{S}_j$is measured by the NCC of the S_p and the S_{q-1} and the NCC of the S_{p+1} and S_q.

$$C^c(\hat{S}_i, \hat{S}_j) = C^c(S_p, S_q) = 1 - \frac{1}{2}\left[NCC(S_p, S_{q-1})\ S\ NCC(S_{p+1}, S_q)\right] \quad (14)$$

Since $NCC(S_p, S_p) = NCC(S_q, S_q) = 1$, we can easily derive,

$$C^c(S_p, S_{p+1}) = C^c(S_{q-1}, S_q) = 0$$

So that it would encourage the selection of consecutive frames in original recording.

3.4.2. Optimal sample sequence

The sample selection procedure is the task of determining the set of image sample $\hat{S}_1^T$ so that the total cost defined by Eq. 11 is minimized:

$$\hat{S}_1^T = \underset{\hat{S}_1, \hat{S}_2, \cdots, \hat{S}_T}{argmin}\ C\left(\hat{V}_1^T, \hat{S}_1^T\right) \quad (15)$$

Optimal sample selection can be performed with a Viterbi search. However, to obtain near real-time synthesis on large dataset, containing tens of thousands of samples, the search space must be pruned. This has been implemented by two pruning steps. Initially, for every target frame in the trajectory, K-nearest samples are identified according to the target cost. The beam width K is 40 in our experiments. The remaining samples are pruned with the concatenation cost.

4. Experimental Results

4.1. Experimental setup

We employ the LIPS 2008/2009 Visual Speech Synthesis Challenge data [12] to evaluate the proposed trajectory-guided sample selection methods. 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 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 HMM models using HTS 2.1 [7].

In objective evaluation, we measured the performance quantitatively using mean square error (MSE), as defined in Eq. 16-17. In a closed test where all the data are used in training, the evaluation is done on all the training data. In open test, leave-20-out cross validation is adopted to avoid data insufficiency problem. In subjective evaluation, the performance of the proposed trajectory-guided approach was evaluated by 20 native language speaking subjects in the audio/visual consistency test in LIPS2009 challenge.

$$\varepsilon_1 = \left\|\hat{V} - V\right\| = \frac{1}{T}\sum_{t=1}^{T}\left\|\hat{V}_t{}^{\top} - V_t{}^{\top}\right\| \qquad (16)$$

$$\varepsilon_2 = \left\|\hat{S} - S\right\| = \frac{1}{T}\sum_{t=1}^{T}\left\|\hat{S}_t{}^{\top} - S_t{}^{\top}\right\| \qquad (17)$$

4.2. Objective test

Fig. 7 shows an example of the HMM predicted trajectory $\hat{\boldsymbol{V}}$ in both the closed and open tests. Comparing with the ground truth $\boldsymbol{V}$, the predicted visual trajectory $\hat{\boldsymbol{V}}$ closely follows the moving trends in $\boldsymbol{V}$. To objectively evaluate the predicted trajectory, we calculated the mean square error of $\hat{\boldsymbol{V}}$ with respect to the ground truth trajectory $\boldsymbol{V}$. Fig. 8 shows the total MSE between V and $\hat{\mathrm{V}}$, and also the respective MSE of the first four PCA components, both in the closed and open tests. The MSE distortion is $7.82\text{x}10^5$ between the HMM-predicted trajectory and the ground truth in open test.

Fig. 9 shows the performance of the HMM prediction with different amount of training data ranging from 4 minutes to 20 minutes long. The MSE is lowered almost in a linear trend when the data size increases. Due to the limited size of the database, the saturate point cannot be observed even with all the 20 minutes A/V data. We believe that more data can further improve the result. We also made some preliminary subjective evaluations on the synthesized mouth videos (restored image sequence from PCA) using the visual parameters generated by the trained HMM models. The results show that using 16 and 20 minutes training data can generate more convincing and natural mouth animations.

The measure, $\left\|\hat{\mathrm{S}} - \mathrm{S}\right\|\ (\hat{\mathrm{V}} = \mathrm{V})$, is to evalutate the performance of trajectory-guided sample selection by ignoring the trajectory prediction error, or ideally we can assume the predicted trajectory is perfect, i.e., $\hat{\mathrm{V}} = \mathrm{V}$. In this oracle experiment, we take the ground truth trajectory as the perfect guidance in order to test the sample selection performance alone. For each test sentence, we use the image samples from other sentences to do the selection and concatenation. The MSE distortion of the sample selection is $1.77\ \text{x}10^5$.

As summarized in Table I, $\left\|\hat{\mathrm{S}} - \mathrm{S}\right\|$ is the total distortion in the synthesis, including both the trajectory prediction errors and sample selection errors. The total distortion $9.42\text{x}10^5$ is slightly less than the summation ($7.82\text{x}10^5$ + $1.77\text{x}10^5$ = $9.59\text{x}10^5$) of the first two distortions.

4.3. Subjective Test

We participate in the LIPS2009 Challenge contest with the proposed photo-real talking head. The contest was conducted in the AVSP (Auditory-Visual Speech Processing) workshop and subjectively evaluated by 20 native British English speaking subjects with normal hearing and vision. All contending systems were evaluated in terms of their audio-visual consistency. When each rendered talking head video sequence was played together with the original speech, the viewer was asked to rate the naturalness of visual speech gestures (articulator movements in the lower face) in a five point MOS score. Fig. 10 shows the subjective results. Our system got the highest MOS score 4.15 among all other participants, which is only inferior to the 4.8 MOS score of the original AV recording.

TABLE I

pairs	$\left\|\hat{V} - V\right\|$	$\left\|\hat{S} - S\right\|$ $(\hat{V} = V)$	$\left\|\hat{S} - S\right\|$
MSE ($\text{x}10^5$)	7.82	1.77	9.42

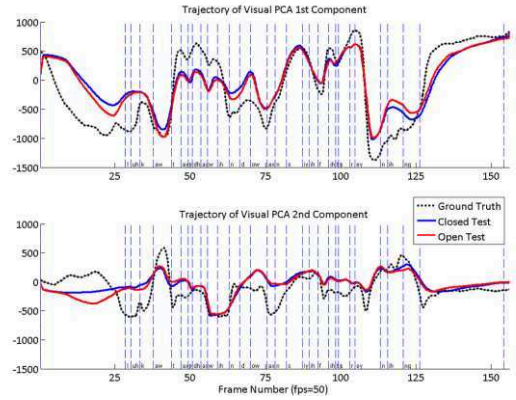

Fig. 7. Closed test predicted (blue curve), open test predicted (red curve) vs. actual (black curve) trajectories of the 1st (up) and 2nd (bottom) PCA coefficients for a testing utterance.

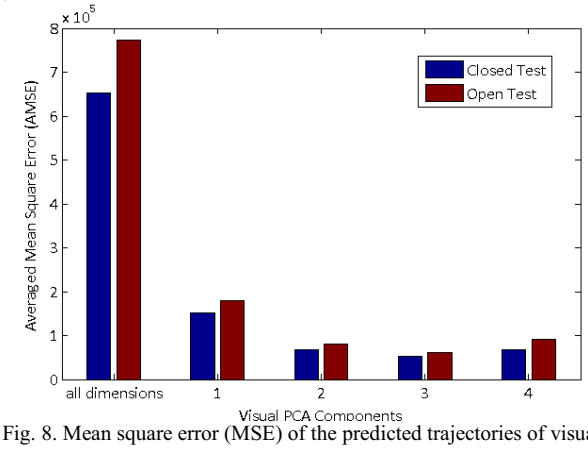

Fig. 8. Mean square error (MSE) of the predicted trajectories of visual PCA coefficients.

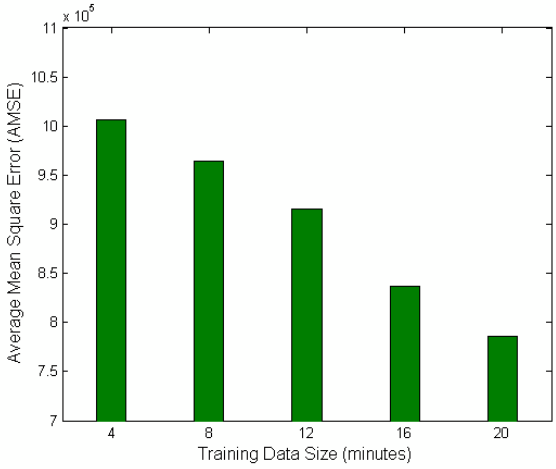

Fig. 9. Mean square error (MSE) vs. Training data size.

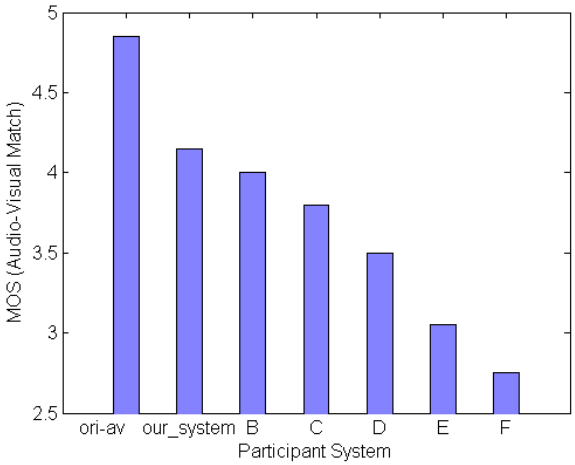

Fig. 10. MOS (Audio-Visual Match) of all the participant systems in LIPS Challenge 2009.

5. Conclusions

We propose a HMM trajectory-guided, real sample concatenating approach for synthesizing high-quality photo-real articulator animation. It renders a photo-real video of articulators in sync with given speech signals by searching for the closest real image sample sequence in the library to the HMM predicted trajectory. Objectively, we evaluated the performance of our system in terms of MSE and investigate the pruning strategies in terms of storage and processing speed. Our talking head took part in the LIPS2009 Challenge contest and won the FIRST place with a subjective MOS score of 4.15 in the Audio-Visual match evaluated by 20 human subjects.

6. Acknowledgements

The authors would like to thank Dr. Qiang Wang who used to be in Microsoft now in Samsung Electronic R&D China, and Dr. Lin Liang in Microsoft Research Asia, for their expertise on head pose tracking and normalization.

7. References

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

[2] 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.

[3] F. Huang, E. Cosatto, H.P. Graf, “Triphone based unit selection for concatenative visual speech synthesis,” Proc. ICASSP 2002. Vol. 2, 2002 pp.2037-2040.

[4] T. Ezzat, G. Geiger, and T. Poggio, “Trainable video realistic speech animation,” Proc. ACM SIGGRAPH2002, San Antonio, Texas, 2002, pp. 388-398.

[5] W. Mattheyses, L. Latacz, W. Verhelst, H. Sahii, “Multimodal Unit Selection for 2D Audiovisual Text-to-Speech Synthesis,” Proc. MLMI 2008, The Netherlands, 2008, pp. 125-136.

[6] K. Liu, J. Ostermann, “Realistic Facial Animation System for Interactive Services,” Proc. Interspeech2008, Brisbane, Australia, Sept. 2008, pp.2330-2333.

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

[8] S. Sako, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “HMM-based Text-To-Audio-Visual Speech Synthesis,” ICSLP 2000.

[9] L. Xie, Z.Q. Liu, “Speech Animation Using Coupled Hidden Markov Models,” Pro. ICPR’06, August 2006, pp. 1128-1131.

[10] Z.H. Ling and R.H. Wang, “HMM-based unit selection using frame sized speech segments,” Proc. Interspeech 2006, Sep. 2006, pp. 2034-2037.

[11] Z.J. Yan, Y. Qian, F. Soong, “Rich-Context Unit Selection (RUS) Approach to High Quality TTS,” Proc. ICASSP 2010, March 2010, pp.4798-4801.

[12] B. Theobald, S. Fagel, G. Bailly, and F. Elisei, “LIPS2008: Visual Speech Synthesis Challenge,” Proc. Interspeech2008, Brisbane, Australia, Sept. 2008, pp.2310-2313.

[13] T. Chen, “Audiovisual speech processing,” Signal Processing Magazine, IEEE Vol.18, Issue 1, Jan. 2001, pp.9-21.

[14] S.A. King, R.E. Parent,”Creating speech-synchronized animation,” Visualization and Computer Graphics, IEEE Transactions on Vol. 11, Issue 3, May-June 2005, pp.341-352.

[15] E. Cosatto and H.P. Graf, “Sample-based synthesis of photo-realistic talking heads,” Proc. IEEE Computer Animation, pp. 103-110, 1998.

[16] T. Ezzat, T. Poggio, “Miketalk: A talking facial display based on morphing visemes,” Proc. Computer Animation, June 1998, pp. 96-102.

[17] B.J. Theobald, J.A. Bangham, I.A. Matthews, G.C. Cawley, “Near videorealistic synthetic talking faces: implementation and evaluation,” Speech Communication 2004, Vol. 44, pp.127-140.

[18] K. Liu, A.Weissenfeld, J. Ostermann, “Parameterization of Mouth Images by LLE and PCA for Image-Based Facial Animation,” Proc. ICASSP 2006, Vol. V, May 2006, pp.461-464.

[19] Q. Wang, W. Zhang, X. Tang, H.Y. Shum, “Real-time Bayesian 3-d pose tracking,” IEEE Transactions on Circuits and Systems for Video Technology 16(12) (2006), pp.1533-1541.

[20] S. Nakamura, “Statistical Multimodal Integration for Audio-Visual Speech Processing,” IEEE Transactions on Neural Networks, Vol.13, No.4, July 2002, pp.854-866.

[21] S. Lucey, T. Chen, S. Sridharan, and V. Chandran, “Integration Strategies for Audio-Visual Speech Processing: Applied to Text-Dependent Speaker Recognition,” IEEE Transactions on Multimedia, Vol.7, No.3, June 2005, pp.495-506.

[22] K. Tokuda, T. Masuko, T. kobayashi and S. Imai, “Speech synthesis using HMMs with dynamic features,” Proc. ICASSP 1996, Vol. I, pp. 389-392.

[23] K. Tokuda, T. Masuko, N. Miyazaki, T. Kobayashi, “Hidden Markov models based on Multi-space probability distribution for pitch pattern modeling,” Proc. ICASSP 1999, Vol. I, pp.229-232.

[24] T. Toda, A. Black, K. Tokuda, “Spectral Conversion Based on Maximum Likelihood Estimation Considering Global Variance of Converted Parameter,” Proc. ICASSP 2005, Vol. I, pp. 9-12.

[25] P. Perez, M. Gangnet, A. Blake, “Poisson Image Editing,” Proc. ACM SIGGRAPH2003, pp.313-318.

[26] X. Huang, A. Acero, H. Hon, Y. Ju, J. Liu, S. Merdith, and M. Plumpe, “Recent improvements on Microsoft’s trainable text-to-speech system – Whistler,” Proc. ICASSP 1997, pp. 959-962.

[27] R.E. Donovan, and E.M. Eide, “The IBM trainable speech synthesis system,” Proc. ICSLP 1998, pp.1703-1706.

[28] T. Hirai, and S. Tenpaku, “Using 5 ms segments in concatenative speech synthesis,” Proc. of 5th ISCA Speech Synthesis Workshop, Pittsburgh, PA, USA, 2004, pp. 37-42.

[29] A. Hunt and A. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” Proc. ICASSP 1996, pp. 373-376.

[30] J.P. Lewis, "Fast Normalized Cross-Correlation," Industrial Light & Magic.

[31] P. Perez, M. Gangnet, A. Blake, “Poisson Image Editing,” in ACM Transactions on Graphics (SIGGRAPH'03), 22(3), pp.313-318.

[32] V. Blanz and T. Vetter, "A Morphable Model for the Synthesis Of 3D Faces," In Proc. ACM SIGGRAPH 99, Los Angeles, CA, 1999, pp.187-194.

[33] F. Pighin, et al, "Synthesizing Realistic Facial Expressions from Photographs," In Proc. ACM SIGGRAPH 98, Orlando, 1998, pp.75-84.

[34] Video demonstration of our synthesis results: http://research.microsoft.com/en-us/projects/photo-real_talking_head/