HIGH QUALITY LIP-SYNC ANIMATION FOR 3D PHOTO-REALISTIC TALKING HEAD

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

[1]Microsoft Research Asia, Beijing, China
[2]Department of Computer Science, Shanghai Jiao Tong University, China
lijuanw@microsoft.com, weihan_cs@sjtu.edu.cn, frankkps@microsoft.com

ABSTRACT

We propose a new 3D photo-realistic talking head with high quality, lip-sync animation. It extends our prior high-quality 2D photo-realistic talking head to 3D. An a/v recording of a person speaking a set of prompted sentences with good phonetic coverage for ~20-minutes is first made. We then use a 2D-to-3D reconstruction algorithm to automatically adapt a general 3D head mesh model to the person. In training, super feature vectors consisting of 3D geometry, texture and speech are augmented together to train a statistical, multi-streamed, Hidden Markov Model (HMM). The HMM is then used to synthesize both the trajectories of head motion animation and the corresponding dynamics of texture. The resultant 3D talking head animation can be controlled by the model predicted geometric trajectory while the articulator movements, e.g., lips, are rendered with dynamic 2D texture image sequences. Head motions and facial expression can also be separately controlled by manipulating corresponding parameters. In a real-time demonstration, the life-like 3D talking head can take any input text, convert it into speech and render lip-synced speech animation photo-realistically.

***Index Terms*—**audio/visual synthesis, lip-sync, 3D, photo-realistic, talking head

1. INTRODUCTION

Avatar is a hot research topic in both academia and industry because it has a wide range of potential applications. Avatars can be roughly divided into two categories, depending upon how avatars interact with the outside world: the first one plays in human-to-human communications, like tele-presence; the other one acts as an intelligent agent in human-computer interactions. However, it is very challenging to create a natural and realistic avatar. Cartoon avatars are relatively easier to build. The more human-like, realistic avatars, which can be seen in some games or movies, are much harder to build. Traditionally, expensive motion capture systems are required to track the real person's motion or, in an even more expensive way, have some artists to manually hand touch every frame. Some desirable features of the next generation avatar are: it should be a 3D avatar to be integrated easily into a versatile 3D virtual world; it should be photo-realistic; it can be customized to any user; last but not least, an avatar should be automatically created with a small amount of recorded data. The next generation avatar should be 3D, photo-realistic, personalized or customized, and easy to create with little bootstrapping data. These are the ultimate goal of this 3D photo-realistic talking head project.

In facial animation world, a great variety of different animation techniques based on 3D models exist [1]. In general, these techniques first generate a 3D face model consisting of a 3D mesh, which defines the geometry shape of a face. For that a lot of different hardware systems are available, which range from 3D laser scanners to multi camera systems. In a second step, either a human-like or cartoon-like texture may be mapped onto the 3D mesh. Besides generating a 3D model, animation parameters have to be determined for the later animation. A traditional 3D avatar requires a highly accurate geometric model to render soft tissues like lips, tongue, wrinkles, etc. It is both computationally intensive and mathematically challenging to make or run such a model. Moreover, any unnatural deformation will make the resultant output fall into the "uncanny valley" of human rejection. That is, it will be rejected as un-natural.

Image-based facial animation techniques achieve great realism in synthesized videos by combining different facial parts of recorded 2D images [2-6]. In general, image-based facial animations consist of two main steps: audiovisual analysis of a recorded human subject and synthesis of facial animation. In the analysis step a database with images of deformable facial parts of the human subject is collected, while the time-aligned audio file is segmented into phonemes. In the second step a face is synthesized by first generating the audio from the text using a Text-to-Speech (TTS) synthesizer. The TTS synthesizer sends phonemes and their timing to the face animation engine, which overlays facial parts corresponding to the generated speech over a background video sequence. [2,3,4,6] show some image-based speech animation that cannot be distinguished from recorded video. However, it is challenging to change head pose freely or to render different facial expressions. Also, it is hard to blend it into 3D scenes seamlessly.

In this work, we propose a new 3D photo-realistic talking head which takes the advantages of both 2D image-based and 3D model-based facial animation techniques. It renders a 3D head by mapping or wrapping 2D video images around a 3D face model. The 2D video sequence can capture the natural movement of soft tissues (e.g. lips and tongue) and it helps the new talking head to bypass the problem caused by occluded articulators (e.g. tongue and teeth). While the 3D face model allows us to render any rigid body motion of a head and also can be deformed for various facial expressions. Therefore, this new 3D photo-realistic talking head combines the best of both the 2D image-based and 3D mesh model-based facial animation methods and overcomes the "uncanny valley" problem of unnatural rendering. The result shows that the 3D photo-realistic talking head looks natural and has realistic animations acceptable to human eyes.

The rest of the paper is organized as follows. Section 2 briefly reviews our previous work on 2D image sample-based lip-sync animation rendering. Section 3 proposes the new 3D photo-realistic

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

lips syncing method. Section 4 shows the experimental results. Section 5 draws the conclusions.

2. 2D IMAGE SAMPLE-BASED LIP-SYNC ANIMATION

2D sample-based synthesis methods use recorded image samples from which new video sequence can be generated. Our previous 2D photo-realistic talking head synthesis system consists of two, training and synthesis, stages [5-6].

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 Principal Component Analysis (PCA) coefficients of aligned mouth images and their dynamic features. The contextual phoneme dependent HMM is used to capture the variations caused by different contextual features. The audio/visual HMM modeling is firstly trained with the traditional maximum likelihood (ML) estimation, and then refined under the Minimum Generation Error (MGE) criterion [5], which can explicitly optimize the quality of generated visual speech trajectory.

In the synthesis stage, when a new audio comes, the input phoneme labels and alignments are firstly converted to a context-dependent label sequence. 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 and the mouth sequence is then stitched back to a background head video.

This approach can achieve high quality lip-sync for 2D facial animation, but it is hard to control head pose freely or add any facial expressions. Next, we propose to extend this work to a 3D photo-realistic talking head.

3. 3D PHOTO-REALISTIC LIP-SYNC ANIMATION

An ideal 3D talking head can mimic realistic motion of a real human face in 3D space. One challenge for rendering realistic 3D facial animation is on the mouth area. Our lip, teeth, and tongue are mixed with non-rigid tissues, and sometimes with occlusions. This means accurate geometric modeling is difficult and also it is hard to deform them properly. Moreover, they need to move together in sync with spoken audio, otherwise people can observe the asynchrony and think it unnatural.

Fig. 1 shows the flowchart of our proposed 3D photo-realistic lip-sync animation method. In training, super feature vectors consisting of 3D geometry, texture and speech are formed to train a statistical, multi-streamed HMM. The HMM is then used to synthesize both the trajectories of geometry animation and dynamic texture. The 3D talking head animation can be controlled by the rendered geometric trajectory while the articulator movements are rendered with the dynamic 2D texture image sequences. In this section, we introduce the details for each step.

3.1. Data acquisition

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. Since the speaker naturally moves/rotates his/her head during the recording, head pose normalization [7] is conducted to normalize every frame to a fully frontal view by applying an affine transform.

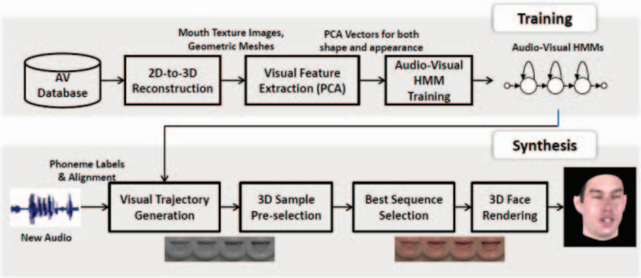
Training
Mouth Texture Images, Geometric Meshes
PCA Vectors for both shape and appearance
Audio-Visual HMMs
AV Database
2D-to-3D Reconstruction
Visual Feature Extraction (PCA)
Audio-Visual HMM Training
Synthesis
Phoneme Labels & Alignment
New Audio
Visual Trajectory Generation
3D Sample Pre-selection
Best Sequence Selection
3D Face Rendering

Fig. 1. Training and synthesis for 3D photo-realistic lip-sync animation.

3.2. From 2D to 3D

In real world, when people talk, led by vocal organs and facial muscles, both the 3D geometry and texture appearance of the face are constantly changing. Ideally, we can capture both geometry change and texture change simultaneously. There is lots of on-going research for solving this problem. For example, with the help of Microsoft Kinect kinds of motion sensing device, people try to use the captured 3D depth information for better acquire the 3D geometry model. On the other hand, people try to recover the 3D face shape from single or multiple camera views [8-10]. In this work, as we don’t have any captured 3D geometry information, we adopt the work in [8] which reconstruct a 3D face model from a single frontal face image.

The only required input to the 2D-to-3D system is a frontal face image of a subject with normal illumination and neutral expression. A semi-supervised ranking prior likelihood models for accurate local search and a robust parameter estimation approach is used for face alignment. Based on this 2D alignment algorithm, 87 key feature points are automatically located, as shown in Fig.2. The feature points are accurate enough for face reconstruction in most cases. A general 3D face model is applied for personalized 3D face reconstruction. The 3D shapes have been compressed by the PCA. After the 2D face alignment, the key feature points are used to compute the 3D shape coefficients of the eigenvectors. Then, the coefficients are used to reconstruct the 3D face shape. Finally, the face texture is extracted from the input image. By mapping the texture onto the 3D face geometry, the 3D face model for the input 2D face image is reconstructed. We reconstruct a 3D face model for each 2D image sample in recordings, as examples shown in Fig.2. Thus a 3D sample library is formed, where each 3D sample has a 3D geometry mesh, a texture, and the corresponding UV mapping which defines how a texture is projected onto a 3D model.

3.3. Parameterization of 3D samples

After 2D-to-3D transformation, original 2D sample recordings turns into 3D sample sequences, which consists of three synchronous streams: geometry mesh sequences for depicting the dynamic shape, texture image sequences for the changing appearance, and the corresponding speech audio. Firstly, the speech audio data of all recorded sequences is segmented into phonemes by a speech recognition system using forced alignment. For 3D samples, PCA projections are performed on both the texture images and the 3D geometric shapes separately. Then the

PCA vectors for texture and 3D geometry are concatenated as a visual features vector with 20 principle components. Acoustic feature MFCCs 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.

3.4. HMM trajectory-guided 3D sample selection

3.4.1 Audio-Visual HMM modeling ($A, V \Rightarrow \lambda$)

We use acoustic vectors $A_t = [a_t^{\top}, \Delta a_t^{\top}, \Delta\Delta a_t^{\top}]^{\top}$ and visual vectors $V_t = [{v_t^1}^{\top}, \Delta {v_t^1}^{\top}, \Delta\Delta {v_t^1}^{\top}; {v_t^2}^{\top}, \Delta {v_t^2}^{\top}, \Delta\Delta {v_t^2}^{\top}]^{\top}$ which is formed by augmenting the static features and their dynamic counterparts to represent the audio, texture, and 3D geometry data, where v_t^1 is texture PCA feature and v_t^2 is for 3D geometry PCA. Multi-streamed audio-visual HMMs, λ, are trained by maximizing the joint probability $p(A, V|\lambda)$ over the training data formed by MFCC(acoustic), texture and geometry PCA vectors. In order to capture the contextual effects, context phoneme dependent HMMs are trained and tree-based clustering is applied to acoustic, texture, and geometry feature streams separately to improve the corresponding model robustness. The MGE training is adopted to further refine the model parameters for improved visual trajectory generation. For each AV HMM state, a single Gaussian mixture model (GMM) is used to characterize the state output. The state q has mean vectors $\mu_q^{(A)}$ and $\mu_q^{(V)}$. In this paper, we use the diagonal covariance matrices for $\Sigma_q^{(AA)}$and $\Sigma_q^{(VV)}$, null covariance matrices for $\Sigma_q^{(AV)}$and$\Sigma_q^{(VA)}$, by assuming the independence between audio and visual streams and between different components.

3.4.2. Visual trajectory generation ($\lambda, A \Rightarrow \hat{V}$)

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.

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

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

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

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

where

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

$$\hat{\Sigma}_{q_t}^{(VV)} = \Sigma_{q_t}^{(VV)} - \Sigma_{q_t}^{(VA)} {\Sigma_{q_t}^{(AA)}}^{-1} \Sigma_{q_t}^{(AV)}. \tag{4}$$

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

$$\begin{aligned} \log p(V|A, Q, \lambda) &= \log p\left(V|\hat{\mu}^{(V)}, \hat{U}^{(VV)}\right) \\ &= -\frac{1}{2} V^{\top} {\hat{U}^{(VV)}}^{-1} V + V^{\top} {\hat{U}^{(VV)}}^{-1} \hat{\mu}^{(V)} + K, \end{aligned} \tag{5}$$

where

$$\hat{\mu}^{(V)} = \left[\hat{\mu}_{q_1}^{(V)}, \hat{\mu}_{q_2}^{(V)}, \cdots, \hat{\mu}_{q_T}^{(V)}\right]^{\top}, \tag{6}$$

$${\hat{U}^{(VV)}}^{-1} = \operatorname{diag}\left[{\hat{\Sigma}_{q_1}^{(VV)}}^{-1}, {\hat{\Sigma}_{q_2}^{(VV)}}^{-1}, \cdots, {\hat{\Sigma}_{q_T}^{(VV)}}^{-1}\right]^{\top}. \tag{7}$$

The constant K is independent of V. 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 V can be represented as a linear conversion,

$$V = W_c C, \tag{8}$$

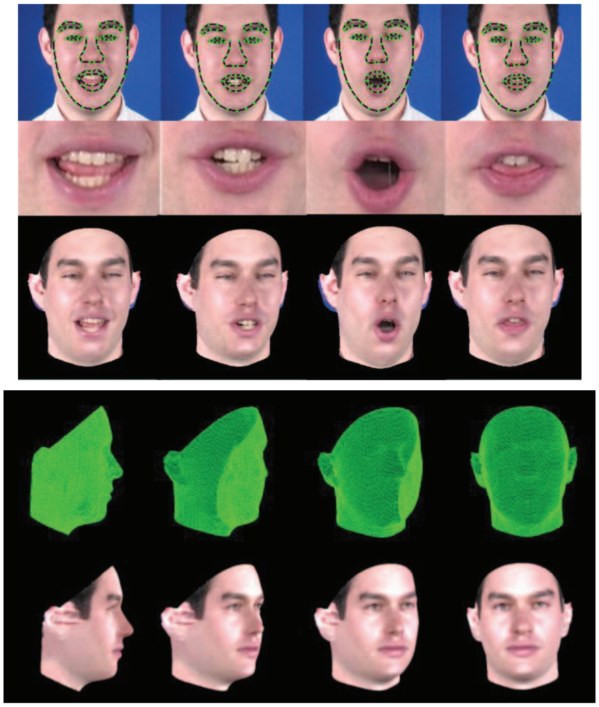

Fig. 2. Auto-reconstructed 3D face model in different mouth shapes and in different view angles (w/o and w/ texture).

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

$$\hat{V}_{opt} = W_c \left({W_c}^{\top} {\hat{U}^{(VV)}}^{-1} W_c\right)^{-1} {W_c}^{\top} {\hat{U}^{(VV)}}^{-1} \hat{\mu}^{(V)}. \tag{9}$$

3.4.3. Trajectory-Guided 3D Sample Selection ($\hat{V} \Rightarrow \hat{\mathbf{S}}$)

The HMM is used to synthesize both the trajectories of geometry animation and dynamic texture. The HMM predicted visual parameter trajectory is a compact description of articulator movements, in the lower rank eigen-lips space. However, the predicted trajectory 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 a real 3D sample sequence $\hat{S}$ from the predicted visual trajectory $\hat{V}$. It searches for the closest real 3D 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 3D samples.

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

$$C\left(\hat{V}_1^T, \hat{S}_1^T\right) = \sum_{i=1}^{T} \omega^t C^t(\hat{V}_i, \hat{S}_i) + \sum_{i=2}^{T} \omega^c C^c(\hat{S}_{i-1}, \hat{S}_i) \tag{10}$$

The target cost of a 3D sample $\hat{S}_i$ is measured by the Euclidean distance between their PCA vectors.

$$C^t\left(\hat{V}_i, \hat{S}_i\right) = \left\|\hat{V}_i - {\hat{S}_i}^{\top} W\right\| \tag{11}$$

The concatenation cost is measured by the normalized 2-D cross correlation (NCC) between the 2D texture images of two 3D samples.

The sample selection procedure is to determine the sequence of 3D sample $\hat{S}_1^T$ so that the total cost defined by Eq. 10 is minimized:

$$\hat{S}_1^T = \underset{\hat{S}_1,\hat{S}_2,\cdots,\hat{S}_T}{\operatorname{argmin}} C(\hat{V}_1^T,\hat{S}_1^T) \tag{12}$$

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 remaining samples are then pruned by measuring the concatenation cost.

3.5. 3D facial animation rendering

The 3D facial animation is controlled by deforming the 3D mesh model according to the rendered geometric trajectory while the articulator movements are rendered with the dynamic 2D texture image sequences. After trajectory-guided 3D sample selection, we get a sequence of 3D samples, which means a sequence of texture images, and a sequence of 3D geometry mesh. Instead of deforming an accurate geometry model, facial animation is rendered by: 1) overlaying the dynamic, time varying texture sequence onto a 3D head model; 2) deforming the 3D face mesh model smoothly according to the 3D geometry sequence. Although the 3D face model is simple without geometric details for lips, tongue and teeth, the realistic articulation motion can be realized by the photo-realistic texture sequence. We project the image sequence of the mouth movement onto the 3D head. As the mouth opens and closes in the 2D image sequence, its projection on the 3D head also opens and closes. Also, the projection can be observed in different directions. Therefore, it can bypass the difficulties in rendering soft tissues like lips, tongue, wrinkles, and make the 3D facial animation look natural and photo-realistic.

3.6. Head movement and facial expressions

With the versatile 3D geometry model, head pose, illumination, and facial expressions of the 3D talking head can be freely controlled. In particular, head movement can be controlled by rotating and translating the head mesh model by viewing it as a rigid object. Different illumination can be realized by changing the lighting in 3D rendering. Various facial expressions like happy or sad can be controlled by deforming the 3D mesh model.

4. EXPERIMENTAL RESULTS

We build a 3D photo-realistic talking head for a male subject using 30-minutes of his 2D video recordings. In a real-time demonstration, the life-like 3D talking head can take any input text and convert it into speech animation photo-realistically. Fig. 3 shows the snapshot of the animation synthesized by our approach. The result shows that the 3D photo-realistic talking head can achieve great realism in lip-sync animation, which is the advantage inherited from image sample-based approach. Also, by taking the advantage of 3D face model, its head motion, illuminations, facial expressions can be flexibly controlled. Informally evaluated by many subjects, the talking head animation looks natural and acceptable to human eyes. Here, a demo video is enclosed to demonstrate the results of the 3D photo-realistic talking head, where both the speech and facial animation are synthesized. Please note that the speech in this demo is synthesized by a standard HMM-based TTS engine.
(video demo: http://research.microsoft.com/en-us/projects/photo-real_talking_head/3d_intro_2.avi)

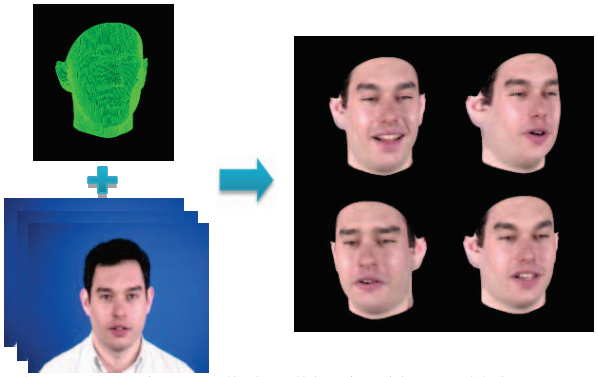

Fig. 3. A 3D photo-realistic talking head by combining 2D image samples with a 3D face model.

5. CONCLUSION

We propose a new 3D photo-realistic talking head with high quality lip-sync animation. Our method combines the best of both 2D image sample-based and 3D model-based facial animation technologies. It renders realistic articulator animation by wrapping 2D video images around a simple and smooth 3D face model. The 2D video sequence can capture the natural movement of soft tissues and it helps the new talking head to bypass the difficulties in rendering occluded articulators (e.g. tongue and teeth). Moreover, with the versatile 3D geometry model, different head poses and facial expressions can be freely controlled. The 3D talking head can be customized to any user by using the 2D video of the user. The new talking head has many useful applications such as voice-agent, tele-presence, gaming, social networking, etc.

6. REFERENCES

[1] N. Ersotelos and F. Dong, "Building Highly Realistic Facial Modeling and Animation: a Survey," *Visual Computer*, vol. 28, pp. 13-30, 2008.

[2] E. Cosatto and H. P. Graf, "Photo-Realistic Talking Heads from Image Samples," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 152-163, 2000.

[3] T. Ezzat, G. Geiger, and T. Poggio, "Trainable Video Realistic Speech Animation," in *Proc. ACM SIGGRAPH2002*, San Antonio, Texas, 2002, pp. 388-398.

[4] K. Liu, J. Ostermann, "Realistic Facial Animation System for Interactive Services," in *Proc. INTERSPEECH 2008*, Brisbane, Australia, Sept. 2008, pp. 2330-2333.

[5] L. Wang, Y. Wu, X. Zhuang, and F. Soong, "Synthesizing Visual Speech Trajectory with Minimum Generation Error," in *Proc. ICASSP 2011,* Prague, Czech Republic, May 2011, pp. 4580-4583.

[6] L. Wang, W. Han, X. Qian, and F. Soong, "Synthesizing Photo-Real Talking Head via Trajectory-Guided Sample Selection," in *Proc. INTERSPEECH 2010*, Chiba, Japan, Sept. 2010, pp. 446-449.

[7] K. Wu, L. Wang, F. Soong, Y. Yam, "A Sparse and Low-Rank Approach to Efficient Face Alignment for Photo-Real Talking Head Synthesis," in *Proc. ICASSP 2011,* Prague, Czech Republic, May 2011, pp. 1397-1400.

[8] Y. Hu, D. Jiang, S. Yan, L. Zhang, H. Zhang, "Automatic 3D Reconstruction for Face Recognition," in *Proc. of the Sixth IEEE international Conference on Automatic Face and Gesture Recognition (FGR'04),* 2004, pp. 843-848.

[9] L. Xin, Q. Wang, J. Tao, X. Tang, T. Tan, and H. Shum, "Automatic 3D Face Modeling from Video," in *Proc. ICCV 2005*, Oct. 2005, pp. 1193-1199.

[10] Z. Liu, Z. Zhang, D. Adler, E. Hanson, M. Cohen, "A Robust and Fast Face Modeling System," in Proc. IEEE Pacific Rim Conference on Multimedia 2001, pp. 269-276.