Synthesizing Photo-Real Talking Head via Trajectory-Guided Sample Selection

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

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

Abstract

In this paper, we propose an HMM trajectory-guided, real image sample concatenation approach to photo-real talking head synthesis. It renders a smooth and natural video of articulators in sync with given speech signals. An audio-visual database is used to train a statistical Hidden Markov Model (HMM) of lips movement first and the trained model is then used to generate a visual parameter trajectory of lips movement for given speech signals, all in the maximum likelihood sense. The HMM generated trajectory is then used as a guide to select, in the original training database, an optimal sequence of mouth images which are then stitched back to a background head video. The whole procedure is fully automatic and data driven. With an audio/video footage as short as 20 minutes from a speaker, the proposed system can synthesize a highly photo-real video in sync with the given speech signals. This system won the FIRST place in the Audio-Visual match contest in LIPS2009 Challenge, which was 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 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. Generating animated talking heads that look like real people is challenging. A photo-real 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 emotional facial expressions. In this paper, we focus on the articulator movements (including lips, teeth, and tongue), which is the most eye-catching region on a talking face.

To synthesize articulator movements from video training data, various approaches have been proposed before, 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.

This paper is organized as follows. Section 2 gives an overview of the synthesis framework. Section 3 introduces the HMM-based visual parameter trajectory generation. Section 4 proposes the trajectory-guided sample selection method. Section 5 discusses the experimental results, and section 6 draws the conclusions.

2. Overview of Synthesis Flow

Fig. 1 illustrates the synthesis framework of the proposed trajectory-guided sample selection approach. In training, first the original image samples $\boldsymbol{S}$ are encoded in low-dimensional visual feature vector $\boldsymbol{V}$. Then the features $\boldsymbol{V}$ 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 $\hat{\boldsymbol{V}}$

in the maximum likelihood sense. The last step is to reconstruct $\widehat{\boldsymbol{V}}$ back to $\widehat{\boldsymbol{S}}$ in the original high-dimensional sample space, so that the synthesis results can be seen/heard. To put it briefly, there are four main modules: $\boldsymbol{S} \Rightarrow \boldsymbol{V}$; $(\mathbf{A}, \boldsymbol{V}) \Rightarrow \boldsymbol{\lambda}$; $(\boldsymbol{\lambda}, \boldsymbol{A}) \Rightarrow \widehat{\boldsymbol{V}}$; and $\widehat{\boldsymbol{V}} \Rightarrow \widehat{\boldsymbol{S}}$. The main contribution of this paper is the last processing module, $\widehat{\boldsymbol{V}} \Rightarrow \widehat{\boldsymbol{S}}$, which is our proposed trajectory-guided real sample selection method for converting the low-dimensional visual parameter trajectory to samples in the original sample space. 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.

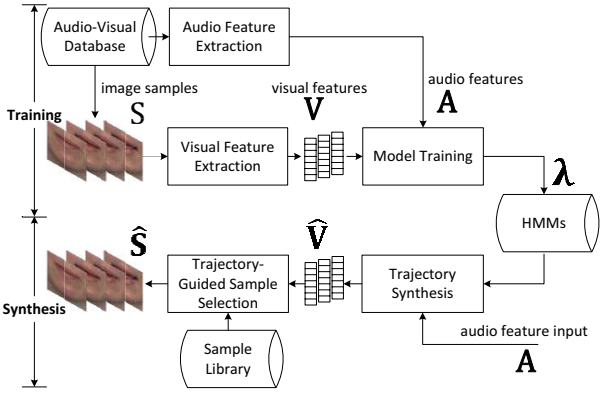

Fig. 1. Synthesis framework with trajectory-guided sample selection.

3. HMM-based Visual Parameter Trajectory Synthesis

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

Training the talking head requires a small, about 20-minutes of audio-visual data of a speaker recorded in reading prompted sentences. Since the speaker moves his/her head naturally during recording, head pose varies among the raw image frames. With the help of a 3D model-based head pose tracking algorithm, head poses of all frames are normalized and aligned to the full-frontal view. The lip images can then be cropped out with a fixed rectangle window and a library of lips sample is made. We obtain eigen-lips (eigenvectors of the lip images) by applying PCA to all the lip 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 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 V, where Q is the state sequence.

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

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

where

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

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

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

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

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

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,

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

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

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

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

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. 2 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 $\widehat{\boldsymbol{S}}$ from $\widehat{\boldsymbol{V}}$. 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.

4.1. Cost function

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:

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

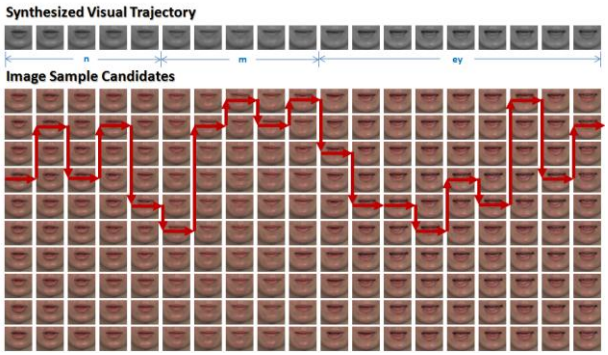

Fig. 2. Illustration for trajectory-guided sample selection approach. The top-line lips images (gray) are the HMM predicted visual trajectory. The bottom images (colored) are real samples lips candidates where the best lips sequence (red arrow path) is selected by Viterbi decoding.

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.

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

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.

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

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.

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:

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

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.

5. Experimental Results

5.1. 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/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) between $\hat{\boldsymbol{V}}$ and $\boldsymbol{V}$, $\hat{\boldsymbol{S}}$ and $\boldsymbol{S}$, as defined in Eq. 16 and 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.

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

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

5.2. Objective test

Fig. 3 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 movement trends in $\boldsymbol{V}$. Three mean square errors (MSE) are calculated for the open test.

1. $\left\| \hat{V} - V \right\|$: This measure shows how good the HMM predicted trajectory which will be used as a guide later is. The model parameters, like the numbers of tied states for the audio and visual streams, are optimized in closed test. The MSE distortion is 7.82×10^5 between the HMM-predicted trajectory and the ground truth in open test.
2. $\left\| \hat{S} - S \right\| \left(\hat{V} = V \right)$: This measure 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.,$\hat{V} = V$. 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 1.77×10^5.
3. $\left\| \hat{S} - S \right\|$: It is the total distortion in the synthesis, including both the trajectory prediction errors and sample selection errors. The total distortion 9.42×10^5 is slightly less than the summation ($7.82 \times 10^5 + 1.77 \times 10^5 = 9.59 \times 10^5$) of the first two distortions.

5.3. Pruning of sample library

Some samples in the sample library are rarely or never selected because they are too far away from the model predicted trajectory. We conducted a large scale synthesis test in order to estimate the frequency of selection for all the image samples in the library. The experiment is to synthesize 10,000 phonetic balanced sentences and compute the frequency of selection of all 61,244 images. As shown in Fig. 4, all the 61244 image samples in the library are rank ordered according to their occurrence count in the final best path (red curve) and k-nearest pre-selection (blue curve), respectively. It shows that there are less than 46% samples used in the pre-selection, while about 20% samples used in the final best path. The pruning is good because misaligned and outlying mouth images are discarded. Meanwhile, we achieve the same output quality but at a much faster speed (5 times) by keeping only a small subset (20%) of the original library.

5.4. 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. 5 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.

6. Conclusions

We propose a 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.

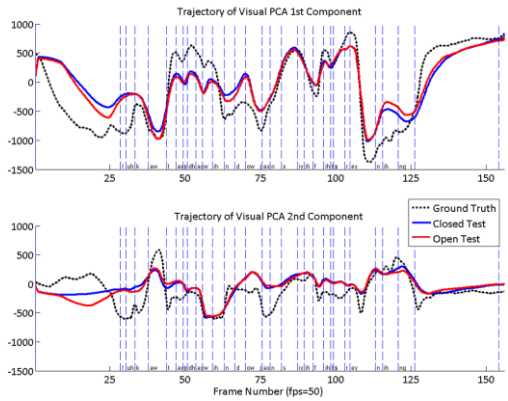

Fig. 3. Closed test predicted (blue curve), open test predicted (red curve) vs. actual (black curve) trajectories of the 1st (up) and 2nd (bottom) PCA coefficients for a testing utterance.

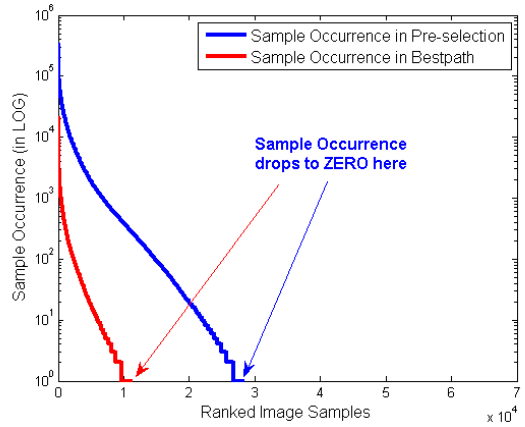

Fig. 4. Re-ranking by sample occurrence.

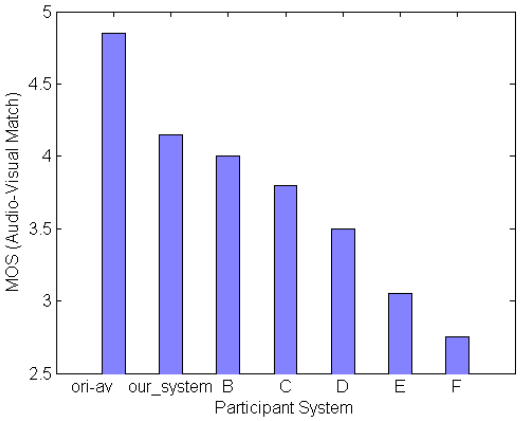

Fig. 5. MOS (Audio-Visual Match) of all the participant systems in LIPS Challenge 2009.

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.