A Minimum Converted Trajectory Error (MCTE) Approach to High Quality Speech-to-Lips Conversion

Xiaodan Zhuang[12], *Lijuan Wang*[1],
Frank Soong[1], *and Mark Hasegawa-Johnson*[2]

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

Abstract

High quality speech-to-lips conversion, investigated in this work, renders realistic lips movement (video) consistent with input speech (audio) without knowing its linguistic content. Instead of memoryless frame-based conversion, we adopt maximum likelihood estimation of the visual parameter trajectories using an audio-visual joint Gaussian Mixture Model (GMM). We propose a minimum converted trajectory error approach (MCTE) to further refine the converted visual parameters. First, we reduce the conversion error by training the joint audio-visual GMM with weighted audio and visual likelihood. Then MCTE uses the generalized probabilistic descent algorithm to minimize a conversion error of the visual parameter trajectories defined on the optimal Gaussian kernel sequence according to the input speech. We demonstrate the effectiveness of the proposed methods using the LIPS 2009 Visual Speech Synthesis Challenge dataset, without knowing the linguistic (phonetic) content of the input speech.

Index Terms: visual speech synthesis, speech-to-lips conversion, minimum conversion error, minimum generation error

1. Introduction

Speech-to-lips conversion aims to render realistic face video, particularly the lips, that is consistent with the input speech audio. This has various applications in multimedia communication. For example, the intelligibility of speech can be increased with a synthesized talking head. We can also reduce the network load for video conferencing by converting speech to face video at the receiving end. Speech-to-lips conversion may also find use in other scenarios when direct video capturing is inappropriate, e.g., when video conferencing from a private environment.

Various approaches have been proposed for speech-to-lips conversion, under different names, such as audio-visual mapping [1], synthesis [2] and lip synchronization [3]. In particular, *phone-based methods* model the audio-visual data using different phone models, mostly artificial neural network [4] and hidden Markov models (HMM) [5]. These models usually synthesize the visual parameters from a phone sequence that is either provided by human labelers or by an automatic speech recognizer (ASR). While the former is expensive and subject to inconsistency resulting from human disagreement in phone labeling, the latter requires a well trained speech recognizer that is usually complex and in need of handmade labels for training.

Direct audio-visual conversion, without using phones, has also been shown effective. For example, a comparison of several single HMM based conversion approaches is available at [1]. Terissi and Gomez [6] inverted an ergodic HMM instead of a set of left-to-right phone HMMs. Recently, Takacs et al.[7] reports that ASR-based speech-to-lips conversion has inferior performance compared with direct conversion using a neural network in their experiments. Some missing feature recovery literature [8] also argues that phone-based models are prone to segmentation and phone identification errors, though the problem can be alleviated by an audio-visual HMM inversion approach (HMMI) [1] that uses a set of phone HMMs but doesn't operate on a phone sequence obtained by Viterbi decoding.

Another class of direct audio-visual conversion methods uses Gaussian Mixture Models (GMM). In [3], while a set of HMMs are used for audio-visual conversion of spoken digits (small vocabulary), large vocabulary audio-visual conversion is performed using frame-by-frame MMSE visual parameter estimation based on a single joint audio-visual GMM. In this work, we focus on GMM-based speech-to-lips conversion that does not use phones as an intermediate representation.

GMMs are also extensively used in another closely related application, voice conversion. Besides frame-by-frame MMSE estimation [9], the GMM has been used for maximum likelihood estimation of complete parameter trajectories. In particular, [10] uses both static and dynamic feature statistics, as investigated in phone-based visual speech synthesis [2], to significantly improve the voice conversion quality. We adopt the same method in GMM-based speech-to-lips conversion to model the constraints between the static and dynamic visual parameters in the framework of maximum likelihood estimation.

Speech-to-lips conversion aims to convert input speech acoustics into lips video as similar to what would have been presented by a talking human as possible. The maximum likelihood estimation criterion provides a effective way to train the GMM and perform the conversion. However, maximum likelihood training does not explicitly optimize the quality of audio-visual conversion. First, the criterion weights all feature dimensions equally and does not take into consideration that they consist of two parts, i.e., the audio part and the video part. Second, an audio-visual GMM with maximum likelihood on the training data does not necessarily result in converted visual trajectories that have minimized error in human perception.

In response to the above issues, we propose a minimum trajectory matching error approach, called Minimum Converted Trajectory Error (MCTE) method, for improved audio-visual conversion. First, we reduce the conversion error by weighting the audio and visual subspaces in training the joint audio-visual GMM. Inspired by Minimum Generation Error (MGE) in speech synthesis [11], we propose further refining the model parameters by minimizing the mean square error between the conversion result and the real visual trajectories using the generalized probabilistic descent (GPD) algorithm.

We develop a GMM-based direct speech-to-lips conversion system incoporating MCTE. Evaluated on the LIPS 2009 Visual Speech Synthesis Challenge task [12], the MTE approach results in improved audio-visual conversion. Although the linguistic content of the input speech is unknown to the presented system, we compare it with the top-rated LIPS2009 submission that has access to the aligned ground truth of phone transcription.

2. MLE-based Audio-visual Conversion

The GMM has been used for maximum likelihood estimation (MLE) of parameter trajectories in speech synthesis and voice conversion. In particular, [10] uses both static and dynamic features in MLE-based conversion to improve the voice conversion quality over frame-by-frame MMSE estimation [9]. A similar approach has been investigated in phone-based visual speech synthesis/rewriting [2]. We propose using the same method in GMM-based direct audio-visual conversion.

The audio-visual conversion leverages a mapping function $\hat{\mathbf{y}} = f(\mathbf{x})$, where $\mathbf{x} = [x_1, x_2, \cdots, x_T]$ is a time sequence of the source feature vectors and $\mathbf{y} = [y_1, y_2, \cdots, y_T]$ is the target feature sequence. Suppose x_t has D_x dimensions, and y_t has D_y dimensions, at each frame, these $static$ features are augmented with the $dynamic$ features and become $2D_x$ or $2D_y$ dimensions: $X_t = [x_t; \Delta x_t]$ and

[0]The first author performed the reported work mainly during a research internship at Microsoft Research Asia.

$Y_t = [y_t; \Delta y_t]$.

Similar to voice conversion [10], we formulate the audio-visual conversion problem as

$$Y = \underset{Y}{\operatorname{argmax}}\, P(Y|X) \approx \underset{Y}{\operatorname{argmax}}\, P(Y|X,\Theta), \tag{1}$$

where $Y = [Y_1; \ldots; Y_t]$, $X = [X_1; \ldots, X_t]$, and Θ is the GMM for the joint probability $P(X_t, Y_t)$.

Given that the GMM has M mixture components,

$$\begin{aligned} P(Y|X) &= \sum_{all\ m} P(m|X)P(Y|X,m) \\ &\approx \sum_{all\ m} P(m|X,\Theta)P(Y|X,m,\Theta) \\ &\approx \prod_{t=1}^{T}\sum_{m_t=1}^{M} P(m_t|X_t,\Theta)P(Y_t|X_t,m_t,\Theta) \end{aligned} \tag{2}$$

Note that

$$\begin{aligned} P(m_t|X_t,\Theta) &= \frac{\omega_{m_t} N(X_t; \mu_{m_t}^{(X)}, \Sigma_{m_t}^{(XX)})}{\sum_{n=1}^{M} \omega_n N(X_t; \mu_n^{(X)}, \Sigma_n^{(XX)})} \\ P(Y_t|X_t,m_t,\Theta) &= N(Y_t; E_{m_t,t}^{(Y)}, D_{m_t}^{(Y)}) \end{aligned} \tag{3}$$

$$\begin{aligned} E_{m_t,t}^{(Y)} &= \mu_{m_t}^{(Y)} + \Sigma_{m_t}^{(YX)}{\Sigma_{m_t}^{(XX)}}^{-1}(X_t - \mu_{m_t}^{(X)}) \\ D_{m_t}^{(Y)} &= \Sigma_{m_t}^{(YY)} - \Sigma_{m_t}^{(YX)}{\Sigma_{m_t}^{(XX)}}^{-1}\Sigma_{m_t}^{(XY)} \end{aligned} \tag{4}$$

As shown in the voice conversion literature [10], the sequence Y can be represented as a linear tranformation of the static vectors: $Y = Wy$, such that $\Delta y_t = \frac{1}{2}(y_{t+1} - y_{t-1})$. Similarly, $X = Wx$. Therefore,

$$\hat{y} = \operatorname{argmax} P(Wy|X,\Theta). \tag{5}$$

The complexity of solving Equation 5 can be significantly reduced by two reasonable approximations.

First, the summation over all mixture component sequences in Equation 2 can be approximated with a single component sequence,

$$P(Y|X,\Theta) \approx P(\hat{m}|X,\Theta)P(Y|X,\hat{m},\Theta), \tag{6}$$

where $\hat{m}$ is the Maximum A Posterior (MAP) mixture component sequence, $\hat{m} = \operatorname{argmax}_m P(m|X,\Theta)$.

With Equation 6, Equation 5 can then be solved in a closed form [10].

$$\hat{y} = (W^T {D_{\hat{m}}^{(Y)}}^{-1} W)^{-1} W^T {D_{\hat{m}}^{(Y)}}^{-1} E_{\hat{m}}^{(Y)}, \tag{7}$$

where

$$E_{\hat{m}}^{(Y)} = \left[E_{\hat{m}_1,1}^{(Y)}, \cdots; \cdots; \cdots, E_{\hat{m}_T,T}^{(Y)}\right] \tag{8}$$

$${D_{\hat{m}}^{(Y)}}^{-1} = diag\left[{D_{\hat{m}_1}^{(Y)}}^{-1}, \cdots; \cdots; \cdots, {D_{\hat{m}_T}^{(Y)}}^{-1}\right]. \tag{9}$$

It is observed in voice conversion that performance degradation owing to the above approximation is not significant [10]. Our preliminary results on audio-visual conversion also confirm that the full-fledged solution by EM algorithm performs no better than the approximated one.

Second, in calculating Equation 4 and Equation 4, we may further simplify the problem by assuming $\Sigma_m^{(XY)} = 0$.

Given a mixture component m_0, the full coviarance matrix in the joint space of X and Y can be partitioned into $\Sigma_{m_0}^{(XX)}$, $\Sigma_{m_0}^{(YY)}$, $\Sigma_{m_0}^{(XY)}$ and $\Sigma_{m_0}^{(YX)}$. In many cases where training data is not abundant, it is not easy to obtain robust estimation of all elements in these matrices. When X and Y are in the same feature space, such as in voice conversion, $\Sigma_{m_0}^{(XY)}$ and $\Sigma_{m_0}^{(YX)}$ are usually approximated using diagonal matrices. In audio-visual conversion, however, X and Y are in different spaces with no strong correlation between the corresponding dimensions. Therefore, we only estimate $\Sigma_{m_0}^{(XX)}$ and $\Sigma_{m_0}^{(YY)}$, yielding the simplified Equation 10.

$$E_{m_t,t}^{(Y)} \approx \mu_{m_t}^{(Y)}, D_{m_t}^{(Y)} = \Sigma_{m_t}^{(YY)} \tag{10}$$

3. AV conversion with MCTE

The MLE-based conversion algorithm is effective and outperforms previous methods. However, maximum likelihood training does not optimize directly towards audio-visual conversion error. In particular, an audio-visual GMM with maximum likelihood for the training data does not lead to converted visual trajectories with minimized error.

Similar problems exist for MLE-based speech synthesis. To compensate this deficiency, Minimum Generation Error (MGE) [11] has been proposed for HMM training. In particular, an appropriate generation error is defined, which is minimized by using a generalized probabilistic descent (GPD) algorithm to update the parameters of the HMMs.

We propose the Minimum Converted Trajectory Error (MCTE) method to further refine the audio-visual conversion result, or any conversion result in general, by minimizing the error between the conversion result and the real target trajectories in the training set.

In Figure 1, we illustrate the speech-to-lips conversion system.

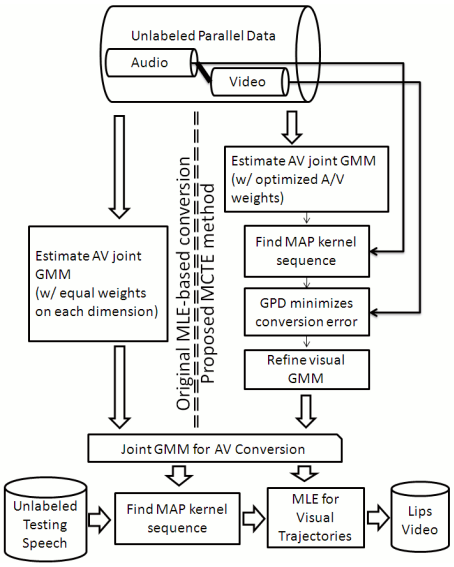

Figure 1: Speech-to-Lips Conversion

3.1. Refined audio modeling

When training the GMM in the joint audio-visual feature space using the EM algorithm, it is common to impose equal weight on all feature dimensions. For the conversion task, this criterion doesn't take into consideration that the training features consist of two parts. We propose weighting the audio and visual subspaces with parameters α_X and α_Y respectively:

$$\begin{aligned} &\log(\mathcal{N}([XY]; \mu_m, \Sigma_m)) \\ =\ & -\log((2\pi)^D |{\Sigma_m^{(XX)}}^{\alpha_X} {\Sigma_m^{(YY)}}^{\alpha_Y}|^{\frac{1}{2}}) \\ & -\frac{1}{2}\alpha_X (X - \mu_m^X)^T {\Sigma_m^{(XX)}}^{-1}(X - \mu_m^X) \\ & -\frac{1}{2}\alpha_Y (Y - \mu_m^Y)^T {\Sigma_m^{(YY)}}^{-1}(Y - \mu_m^Y). \end{aligned} \tag{11}$$

In our experiments, we observe consistently that weighting the audio spaces more than the visual space reduces the mean square error of the converted visual trajectories. According to Equation 2, the conversion quality is affected by $P(m|X,\Theta)$ and $P(Y|m,\Theta)$, which can be interpreted as choosing the right mixture component for mapping given the audio observation and estimating the visual distribution given the mixture component. Heavier weighting on the audio subspace in Equation 11 leads to more distinguishable mixture components in $P(m|X,\Theta)$ but increased perplexity of $P(Y|m,\Theta)$. The observation suggests that $P(m|X,\Theta)$ may be dominating the approximation quality of Equation 2. This may also depend on the particular parameterization of the features.

Note that though it is possible to fine tune the weighting parameters, we find the empirical choice of weighting exclusively on the audio subspace ($\alpha_X = 1, \alpha_Y = 0$) already result in significant performance improvement.

3.2. Refined visual modeling

Inspired by the MGE, we further improve the conversion result by refining the visual GMM model using the GPD algorithm.

We define the conversion error as the Euclidean distance between the conversion result and the real visual trajectory in the training set,

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

With the approximation using the MAP mixture component sequence adopted in Equation 6, the conversion problem, i.e., maximizing $P(Y|X, \Theta)$, becomes the following two steps. First, given the sequence of audio features X, a MAP mixture sequence is estimated: $\hat{m} = \text{argmax}\, P(m|X, \Theta)$. Second, given the MAP mixture sequence, the visual features are estimated by maximizing $P(Y|\hat{m}, \Theta)$. Note that the second step is the same as a parameter generation problem for a mixture component sequence $\hat{m}$. In other words, we tackle the conversion problem by generating features from a corresponding HMM, which has a sequence of states and Gaussian kernels $\hat{m}$ determined by the MAP process.

Therefore, we can improve the conversion performance by minimizing the empirical conversion error, measured using a cost function $L(\Theta)$ similar to MGE in synthesis [11].

$$L(\Theta) = \frac{1}{N} \sum_{i=1}^{N} D(y^i, \hat{y}^i(\hat{m}^i, \Theta)), \tag{13}$$

where N is the number of training utterances.

Using the GPD algorithm, given the n^{th} training utterance, the updating rule for the parameters of the mixtures on the MAP sequence is

$$\begin{aligned} & \Theta(n+1) \\ = \;& \Theta(n) - \epsilon_n \frac{\partial}{\partial \Theta} D(y^n, \hat{y}^n(\hat{m}^n, \Theta))|_{\Theta=\Theta(n)} \\ & \frac{\partial}{\partial \Theta} D(y^n, \hat{y}^n(\hat{m}^n, \Theta)) \\ = \;& 2\left(\hat{y}^n(\hat{m}^n, \Theta) - y^n\right)^T \frac{\partial}{\partial \Theta} \hat{y}^n(\hat{m}^n, \Theta). \end{aligned} \tag{14}$$

In particular, with Equation 7 and Equation 14,

$$\begin{aligned} & \frac{\partial \hat{y}^n(\hat{m}^n, \Theta)}{\partial E^{(Y)}_{\hat{m}_t, t, d}} \\ = \;& \left(W^T \left(D^{(Y)}_{\hat{m}}\right)^{-1} W\right)^{-1} W^T \left(D^{(Y)}_{\hat{m}}\right)^{-1} Z_E \end{aligned} \tag{15}$$

where $E^{(Y)}_{\hat{m}_t, t, d}$ is the d^{th} dimension of the mean vector of the t^{th} mixture in the MAP mixture sequence, and $Z_E = [0, \ldots, 0, 1_{t \times D_Y + d}, 0, 0, \ldots, 0]^T$.

For simplicity, we further assume that $\Sigma^{(YY)}_{m_0}$ has only diagonal non-zero elements, i.e., $\sigma^2_{t,d}$ is the variance corresponding to $E^{(Y)}_{\hat{m}_t, t, d}$. Denote $v_{t,d} = 1/\sigma^2_{t,d}$ and $Z_v = Z_E Z_E^T$,

$$\begin{aligned} & \frac{\partial \hat{y}^n(\hat{m}^n, \Theta)}{\partial v_{t,d}} \\ = \;& \left(W^T \left(D^{(Y)}_{\hat{m}}\right)^{-1} W\right)^{-1} W^T Z_v \left(E^{(Y)}_{\hat{m}} - W \hat{y}^n(\hat{m}^n, \Theta)\right) \end{aligned} \tag{16}$$

In contrast to the MGE, which directly estimates the parameters in the involved HMMs, MCTE uses the GPD algorithm to update the visual distribution parameters of the MAP mixture component sequence, which replace the corresponding parameters in the visual GMM.

4. Experiments and Results

4.1. Setup

We employ the dataset used in LIPS 2008/2009 Visual Speech Synthesis Challenge [12] to evaluate the proposed audio-visual conversion methods. This dataset has 278 video files with corresponding audio tracks, each being one English sentence spoken by a single native speaker with neutral emotion.

The video is sampled at every 20ms, or 50 frames per second. For each image, Principle Component Analysis (PCA) 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 from local windows of 20ms with a step size of 5ms. The visual parameter vectors are interpolated up to the same sampling frequency as the MFCCs. In audio-visual conversion, the input sequence of MFCCs are converted into a sequence of visual PCA vectors, which drives a lips movement image sequence before it is stitched to a facial background video [13].

We compare the performance of several alternative conversion modules.

1. *PhnRewriting* is a phone rewriting method leverages a set of tied triphone visual parameter HMMs, trained using the visual PCA sequences segmented by human labeled phone transcription. The visual speech synthesis has access to ground truth phone labels and boundaries, and is performed by using the visual HMMs to synthesize the visual PCAs. This was the MSRA submission to LIPS 2009 [13].
2. *Conv(equal)* is a direct audio-visual conversion method based on maximum likelihood estimation. Each audio and visual dimension is weighted equally when training the joint GMM. We empirically determined to use 1024 Gaussian mixtures.
3. *MCTE* is the proposed MCTE method, with the same number of Gaussian mixtures as the alternative direct conversion modules.
4. *Conv(a-weighted)* and *MCTE(a-weighted)* weight each audio dimension equally and the visual dimensions have weight zero when training the joint GMM. *Conv(v-weighted)* and *MCTE(v-weighted)* have weight zero for the audio dimensions.

Note that while Takacs et al. [7] uses an automatic speech recognizer to obtain the phone sequence for the phone-based video rewriting system, we use the LIPS 2009 ground truth phone labels and boundaries in the phone rewriting methods, eliminating the question about the quality of the speech recognition results. This should be helpful in understanding the performance of phone rewritting approach and the direct conversion approach.

4.2. Objective evaluation results

The objective evaluations are performed using two metrics. First, we use all the data for both training and conversion, for the "Training" set performances. Second, we perform leave-20-out cross validation, and the measures from all the folds are averaged to form the "Testing" set performance.

The conversion performances are evaluated using Mean Square Error (MSE) and Average Correlation Coefficient (ACC), defined as follows,

$$MSE = \frac{1}{T} \sum_{t=1}^{T} \|\hat{y}_t - y_t\|, \tag{17}$$

$$ACC = \frac{1}{TD} \sum_{t=1}^{T} \sum_{d=1}^{D} \frac{(y_{t,d} - \mu_{y_d})(\hat{y}_{t,d} - \mu_{\hat{y}_d})}{\sigma_{y_d} \sigma_{\hat{y}_d}}. \tag{18}$$

In Figure 2, we can see that weighting only the audio dimensions in training the joint GMM consistently improves both the MLE-based conversion and the MCTE conversion. The proposed MCTE method consistently outperforms MLE-based direct conversion. Having access to human labeled phone transcription in both training and testing gives the phone rewriting methods an advantage. The gap however is largely reduced by adopting the proposed MCTE method. According to the objective measures, the best MCTE performance is comparable to the Phn-Rewriting method, which won the top audio-visual consistency ranking in LIPS 2009 [13] .

In Figure 3, we illustrate the conversion results by the MCTE method and by the MLE-based direct conversion method, respectively. The proposed MCTE method is shown to result in trajectories more similar to the ground truth which a human speaker produces.

4.3. Subjective evaluation results

A subjective "scoring" test is also carried out to compare Conv(a-weighted), MCTE(a-weighted), PhnRewriting and the original recording. We select twelve sentences from the LIPS 2009 test set, 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 conversion results are randomly assigned into six subjective

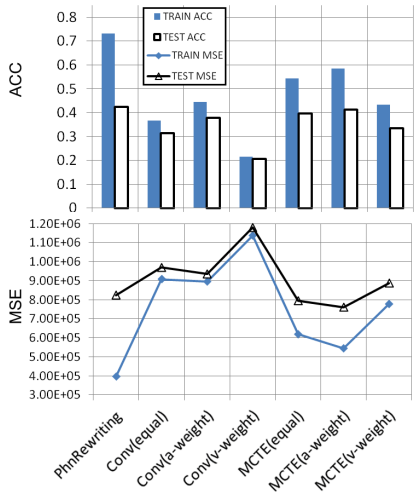

Figure 2: Objective Evaluations

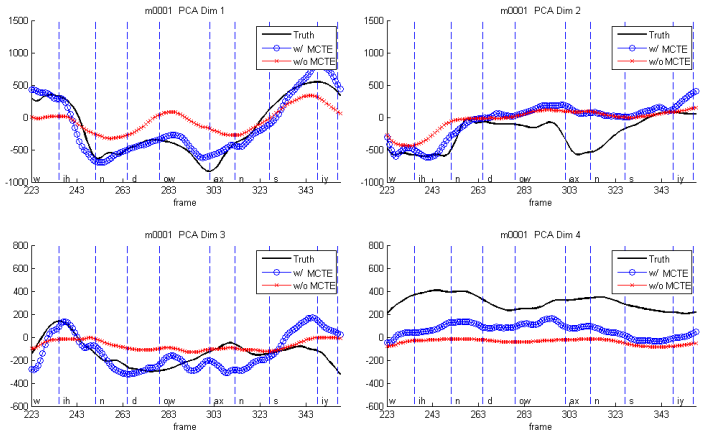

Figure 3: Top PCA dimensions w/ and w/o MCTE (coded as blue circles and red crosses respectively)

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.

Figure 4 shows the averaged subjective scores for "audio-visual consistency". Besides the similar observations as in the objective evaluations, we also point out the p-values in the unpaired two-tailed T-Test: MCTE(a-weighted) and Conv(a-weighted) 0.0002%; MCTE and PhnRewriting 3.9%.

5. Conclusion & Discussion

This work investigates the problem of speech-to-lips conversion and aims to render photo-realistic lips movement that are consistent with the input speech signal without knowing the underlying linguistic content. Instead of frame-based conversion, it adopts the maximum likelihood based Gaussian Mixture Model (GMM) in estimating visual parameter trajectories. We propose Minimum Converted Trajectory Error (MCTE) training to refine the converted visual trajectories. The proposed method leverages a joint audio-visual GMM trained with audio-visual-weighted maximum likelihood criterion. MCTE uses the generalized probabilistic descent algorithm to minimize conversion error of the visual parameter trajectories defined on the optimal mixture component sequence ob-

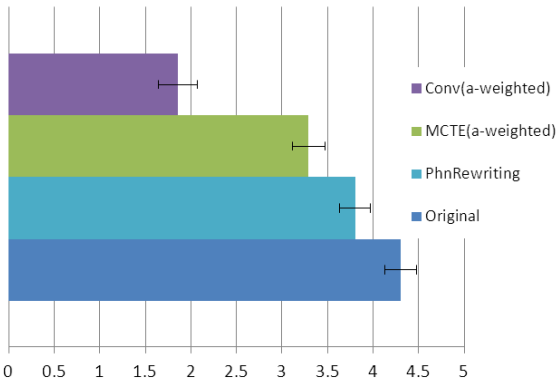

Figure 4: Subjective scores for "audio-visual consistency" (with standard errors)

tained using the input speech. On the LIPS 2008/2009 visual speech synthesis challenge dataset, we demonstrate the effectiveness of the proposed MCTE method. The best presented system, without knowing the linguistic content of the input speech, is compared with the top-rated LIPS 2009 submission that utilized the given ground truth phone sequence and their timing information. The proposed MCTE method can be applied to general conversion problems, not necessarily limited to the speech-to-lips conversion reported in this work.

6. Acknowledgements

This research is partially funded by NSF grant IIS-0703624.

7. References

[1] S. Fu, R. Gutierrez-Osuna, A. Esposito, P. K. Kakumanu, and O. N. Garcia, "Audio/visual mapping with cross-modal hidden markov models," *IEEE Transactions on Multimedia*, vol. 7, no. 2, pp. 243–252, April 2005.

[2] S. Sako, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "HMM-based text-to-audio-visual speech synthesis," in *International Conference on Speech and Language Processing (ICSLP)*, vol. 3, 2000, pp. 25–28.

[3] T. Chen, "Audiovisual speech processing," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 9–21, Jan 2001.

[4] P. Hong, Z. Wen, and T. Huang, "Real-time speech-driven face animation with expressions using neural networks," *Neural Networks, IEEE Transactions on*, vol. 13, no. 4, pp. 916–927, Jul 2002.

[5] L. Xie and Z.-Q. Liu, "A coupled HMM approach to video-realistic speech animation," *Pattern Recognition*, vol. 40, no. 8, pp. 2325 – 2340, 2007, part Special Issue on Visual Information Processing.

[6] L. D. Terissi and J. C. Gomez, "Audio-to-visual conversion via HMM inversion for speech-driven facial animation," *Lecture Notes in Computer Science: Advances in Artificial Intelligence - SBIA 2008*, vol. 5249, pp. 33–42, 2008.

[7] G. Takacs, "Direct, modular and hybrid audio to visual speech conversion methods - a comparative study," in *Proc. Interspeech*, Brighton, UK, 2009.

[8] B. Raj and R. Stern, "Missing-feature approaches in speech recognition," *Signal Processing Magazine, IEEE*, vol. 22, no. 5, pp. 101–116, Sept. 2005.

[9] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, no. 2, pp. 131–142, Mar 1998.

[10] 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.

[11] Y.-J. Wu and R.-H. Wang, "Minimum generation error training for HMM-based speech synthesis," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, May 2006, pp. I–I.

[12] B.-J. Theobald, S. Fagel, G. Bailly, and F. Elisei, "LIPS 2008: Visual speech synthesis challenge," in *Interspeech*, 2008, pp. 2310–2313.

[13] L.-J. Wang, X.-J. Qian, W. Han, and F. Soong, "Synthesizing photo-real talking head via trajectory-guided sample selection," 2010, accepted to Interspeech.