Eurographics/ACM SIGGRAPH Symposium on Computer Animation (2005)
K. Anjyo, P. Faloutsos (Editors)

Transferable Videorealistic Speech Animation

Yao-Jen Chang[1] and Tony Ezzat[2]

[1]Advanced Technology Center, Computer and Communications Laboratories, ITRI, Taiwan
[2]Center for Biological and Computational Learning, MIT, USA

Abstract

Image-based videorealistic speech animation achieves significant visual realism at the cost of the collection of a large 5- to 10-minute video corpus from the specific person to be animated. This requirement hinders its use in broad applications, since a large video corpus for a specific person under a controlled recording setup may not be easily obtained. In this paper, we propose a model transfer and adaptation algorithm which allows for a novel person to be animated using only a small video corpus. The algorithm starts with a multidimensional morphable model (MMM) previously trained from a different speaker with a large corpus, and transfers it to the novel speaker with a much smaller corpus. The algorithm consists of 1) a novel matching-by-synthesis algorithm which semi-automatically selects new MMM prototype images from the new video corpus and 2) a novel gradient descent linear regression algorithm which adapts the MMM phoneme models to the data in the novel video corpus. Encouraging experimental results are presented in which a morphable model trained from a performer with a 10-minute corpus is transferred to a novel person using a 15-second movie clip of him as the adaptation video corpus.

Categories and Subject Descriptors (according to ACM CCS): I.3.7 [Computer Graphics]: Animation

1. Introduction

Image-based videorealistic speech animation [BCS97, CG00, EGP02] has drawn wide attention due to its supreme visual realism. Different from 3D graphics-based speech animation, image-based videorealistic speech animation can provide better realism such that it potentially can be used for creating virtual teachers for language learning and digital characters in movies. Even a visual "Turing test" was conducted in [EGP02], which showed ordinary observers can hardly distinguish synthetic speech animation from real speech video with the same utterances.

However, to realize image-based videorealistic speech animation, a large video corpus is required as the database for creating novel speech animation by re-arrangement [BCS97], concatenation [CG00], or as the training data for analyzing audio-visual dynamics to train morphable models [EGP02]. While user adaptation can be easily achieved for 3D graphics-based speech animation by deforming a 3D head model and reusing animation parameters [CC02, NN01], it is quite difficult to create image-based speech animation for a novel person without recollection of a large video corpus. This requirement limits the use of image-based videorealistic speech animation, since a large video corpus for a specific person uttering specified transcripts under a controlled environment setting may not be easily obtained. Furthermore, a large video corpus would take several days of pre-processing. It would be much better to be able to create personal speech animation with only small amount of video data.

Based on the work of trainable videorealistic speech animation proposed by Ezzat et al. [EGP02], we propose new approaches to resolve this problem. With a matching-by-synthesis approach, we are able to transfer an original morphable model (MMM) trained from a large corpus to a novel person with a small video corpus. Furthermore, we propose a model adaptation algorithm to refine the MMM phoneme model by incorporating a linear regression adaptation concept that is similar to one that is widely adopted in speech recognition literature. Thereby, the speaking style of the synthesized animation can be more similar to that of the novel person.

In the next section, a review of previous work is presented. The trainable videorealistic speech animation method [EGP02] as the foundation of this work is also briefly described in Section 3. The proposed model transfer and adaptation algorithms are detailed in Sections 4 and

5. Some preliminary experimental results are presented in Section 6. Finally, Section 7 concludes this paper and states some possible future directions.

2. Previous Work

Image-based videorealistic speech animation was first presented in the work of Video Rewrite proposed by Bregler et al. [BCS97]. By recording a large video corpus and performing triphone-based segmentation, speech synthesis is achieved by concatenating sequences in the video corpus that best match the desired novel utterance. Subsequently, Cosatto and Graf [CG00] proposed a similar audio-visual unit selection method by using the Viterbi search from a video corpus to find best matches that minimize the target cost of the viseme dissimilarity between the selected units and target phonemes and the concatenation cost between two selected consecutive images. These two approaches directly reuse the images in the pre-recorded video corpus without using any generative models for speech animation synthesis, which hinder themselves from transferring the speaker to another person without recollection of a large video corpus. Unlike these approaches, the trainable videorealistic speech animation proposed by Ezzat et al. [EGP02] adopted multidimensional morphable models for analysis and synthesis of speech animation, and is more amenable to transfer between speakers. This method forms the foundation of our work and is further described in the next section.

With the construction of 3D morphable models, Blanz et al. [BBPV03] proposed an approach that can generate synthetic speech animation from just one photo or portrait of a novel person. By decomposition of dynamic visual information as a linear combination of shape, motion, and texture components, life-like animation can be created for a novel person. However, the region inside the mouth cannot be well-modeled using this 3D representation; the use of artificial teeth or tongue is not as natural as image-based approaches. In [CFKP04], Cao et al. proposed real-time speech moition synthesis for 3D facial models constructed with photogrammetric technique of [PHL*98]. With collection of a large speech motion corpus, high-fidelity facial motions can be real-time synthesized for novel utterances based on a graph-based approach. In our work, we propose to learn the mouth appearances and dynamics from a small video corpus. Visual realism is thus retained for speech animation synthesis.

Also close to our work is the research on motion transfer, or retargetting [Gle98], with which the motion of one performer can be transferred to another character. For retargetting of facial animation, most researches focus on the work of expression mapping for face images [LSZ01, ZLGS03] or 3D facial models [NN01, WHL*04, NJ04]. However, these motion transfer techniques do not address how to retarget image-based speech animation. Beyond transferring motions of speech animation from one person to another, which can also be done with the proposed approach, we transfer and adapt the generative model from one person to another with a small adaptation video corpus. Thereby, videorealistic speech animation can be directly synthesized with the speaking style of the novel person.

3. Background: Trainable Videorealistic Speech Animation

In 2002, Ezzat et al. [EGP02] proposed an image-based videorealistic speech animation approach with machine learning techniques. Totally 46 image prototypes are selected from a recorded corpus, and its texture and motion flows with respect to a reference image with neutral face are utilized for modeling the space of all possible mouth appearances with a multidimensional morphable model (MMM) [JP98]. A trajectory synthesis procedure based on regularization technique is then employed to map an input phoneme stream to a trajectory of parameters in the trained MMM space. Therefore, a videorealistic speech animation can be synthesized from input audio that is phonetically transcribed and aligned. In the following subsections, the multidimensional morphable model and trajectory analysis/synthesis are briefly described. Interested readers are referred to [EGP02] for the details.

3.1. Multidimensional Morphable Model

The multidimensional morphable model (MMM) [JP98] was originally proposed by Jones and Poggio for image recognition, in which the visual information of an image is represented by shape and texture parameters. Instead of using MMM for image analysis, Ezzat et al. [EGP02] utilized an MMM for image synthesis of mouth texture and movements during speech. Firstly, a set of prototype images are automatically selected from the video corpus using the k-means clustering algorithm [Bis95]. Then, each prototype is decomposed into a motion component (represented by optical flow vectors) and a texture component. Each synthesized image can then be modeled as a linear combination of the motion and texture components of the selected prototype images.

More formally, given a set of M prototype images $\{I_{P_i}\}$ and prototype flows $\{C_{P_i}\}$, the motion component C^{syn} and texture component I^{syn} of each novel synthetic image can be modeled as:

$$C^{syn} = \sum_{i=1}^{M} \alpha_i C_{P_i}, \tag{1}$$

$$I^{syn} = \sum_{i=1}^{M} \beta_i I_{P_i}^{warped} = \sum_{i=1}^{M} \beta_i \mathbf{W_F}\left(I_{P_i}, \mathbf{W_F}\left(C^{syn} - C_{P_i}, C_{P_i}\right)\right), \tag{2}$$

where $\mathbf{W_F}(\boldsymbol{p}, \boldsymbol{q})$ is a forward warp operation that warps vectors $\boldsymbol{p}$ according to flow vectors $\boldsymbol{q}$. Conversely, given a set

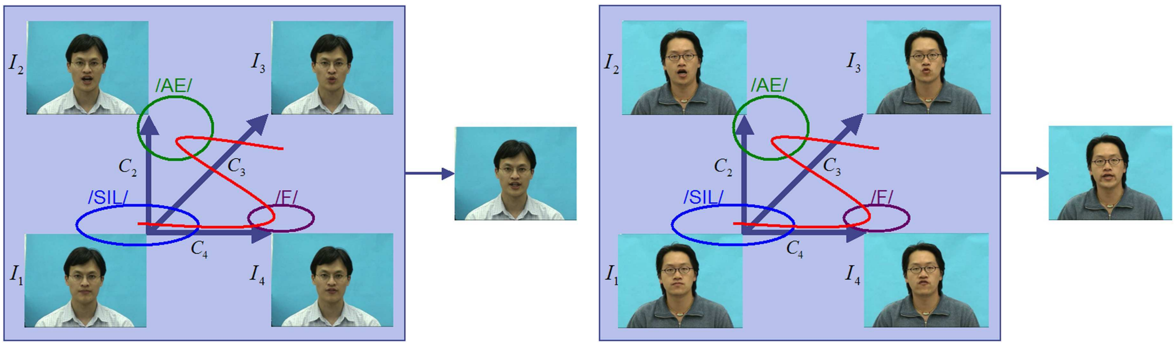


Figure 1: *The conceptual illustration of the idea of model transfer. The prototype images I_1 to I_4 and the optical flows C_2 to C_4 form the MMM parameter space. Phoneme models are approximated by Gaussian distributions as depicted in ellipses. The synthesized MMM parameters for a phoneme sequence {..., /SIL/, /F/, /AE/, ...} form a smooth trajectory (in red line) in the MMM space.* ***Left****: Image synthesis with the MMM of the original subject.* ***Right****: Image synthesis with the MMM of the new subject in the same MMM parameter space in which the phoneme model is directly transferred from the phoneme model of the original subject.*

of MMM parameters $\{\alpha_i, \beta_i\}_{i=1}^{M}$, a new mouth image I^{syn} can be synthesized by warping and blending the prototype images.

3.2. Trajectory Analysis and Synthesis

The goal of trajectory analysis and synthesis is to learn a phoneme model and use it to synthesize novel speech trajectories in the MMM parameter space. The characteristics of the MMM parameters for each phoneme are examined from corresponding image frames according to the audio alignment result. For simplicity, each phoneme p is modeled as a multidimensional Gaussian with mean vector μ_p and diagonal covariance matrix Σ_p. A trajectory of a novel speech sequence $\boldsymbol{y}$ is derived by minimizing the following objective function,

$$E_s = (\boldsymbol{y}-\boldsymbol{\mu})^T \mathbf{D}^T \boldsymbol{\Sigma}^{-1} \mathbf{D} (\boldsymbol{y}-\boldsymbol{\mu}) + \lambda \boldsymbol{y}^T \mathbf{W}_k^T \mathbf{W}_k \boldsymbol{y}, \tag{3}$$

where the synthetic MMM parameters $\boldsymbol{y}$ are obtained by minimizing the distance to the cascaded target mean vector $\boldsymbol{\mu}$ (weighted by the duration-normalization matrix $\mathbf{D}$, and the inverse of the covariance matrix $\boldsymbol{\Sigma}$), while also retaining smoothness controlled by the k-th order difference matrix $\mathbf{W}_k$.

However, the synthetic MMM parameters tend to be under-articulated when the mean and covariance for each phoneme are directly estimated from the pooled MMM parameters for each phoneme. To resolve the problem, gradient descent learning is employed to refine the phoneme model by iteratively minimizing the difference E_a between the synthetic MMM trajectories $\boldsymbol{y}$ obtained from Equation 3 and real MMM trajectories $\boldsymbol{z}$ derived from Equations 1 and 2.

More formally, the error between real and synthetic trajectories is defined by

$$E_a = (\boldsymbol{z}-\boldsymbol{y})^T (\boldsymbol{z}-\boldsymbol{y}) \tag{4}$$

and the phoneme model is refined by

$$\mu_p^{new} = \mu_p^{old} - \eta \frac{\partial E_a}{\partial \mu_p} \quad \text{and} \quad \Sigma_p^{new} = \Sigma_p^{old} - \eta \frac{\partial E_a}{\partial \Sigma_p} \tag{5}$$

with a small learning rate parameter η.

To summarize, the trainable videorealistic speech animation framework proposed by Ezzat et al. [EGP02] required two sets of parameters: a set of M prototype images I_{P_i} and prototype flows C_{P_i} to represent the flow and texture respectively of the subject's mouth; and a set of phoneme models $N(\mu_p, \Sigma_p)$ which model each phoneme p in the MMM space using a Gaussian distribution for trajectory analysis and synthesis.

4. Model Transfer

With a small video corpus from a novel person, there would not be enough data to retrain an entire MMM phoneme model. Therefore, one simple solution to model transfer is to choose a new set of prototype images from the new video corpus, and then directly transfer the original phoneme model to the novel person. As illustrated in Figure 1, the prototype images of the original user are replaced by corresponding images of the novel subject from the smaller corpus, and trajectory synthesis is performed with the original phoneme models.

Since each dimension of the MMM parameters is associated with a specific prototype image obtained from the original video corpus, it is required that the newly selected prototype images should exhibit similar mouth appearance to the corresponding prototype images of the original user. However, manual comparison for each image in the new video corpus with the original prototype images would take hours of work even for a short video corpus with about 300 frames. Inspired by the work presented by Beymer and Poggio [BP95] that generates synthetic images under various poses and expressions for robust face recognition, we propose a matching-by-synthesis approach to semi-automatically select new set of prototype images from the new video corpus that resemble the original prototype images.

Our matching-by-synthesis approach works in three steps: Firstly, an *initialization* step is performed in which manual correspondence is established between two reference frames in the original and novel corpus. The correspondence is defined intitially by a set of manually placed points, and then interpolated using an RBF method to establish dense point correspondence between the two reference images. Secondly, a *flow matching* step is performed in which a set of prototype image candidates are chosen from the novel corpus based on how much their optical flow shape matches the optical flow shape of the prototypes from the original corpus. Finally, a *texture matching* step is performed to refine the prototype candidates based on how much their "texture coordinates" match the texture coordinates of the prototypes from the original corpus.

In the following sections, we describe the steps of our matching-by-synthesis algorithm in detail.

4.1. Initialization

To simplify the calculation of dense point correspondence between images from different persons, an RBF-based interpolation method [Bis95] is utilized. A set of N feature points as illustrated in Figure 2(a) are manually marked on both of the reference image of the original user (denoted as Ref_A) and the reference of the novel person (denoted as Ref_B). With the RBF-based interpolation method, the dense correspondence between each point $\boldsymbol{p} = (p_x, p_y)^T$ in Ref_A and the corresponding point $S(\boldsymbol{p})$ in Ref_B is formulated by a linear combination of radial basis function augmented with a low-order polynomial function, i.e.,

$$S(\boldsymbol{p}) = \sum_{k=1}^{N} \lambda_k \phi\left(\left\|\boldsymbol{p} - \boldsymbol{p}_k^a\right\|\right) + Q(\boldsymbol{p}), \tag{6}$$

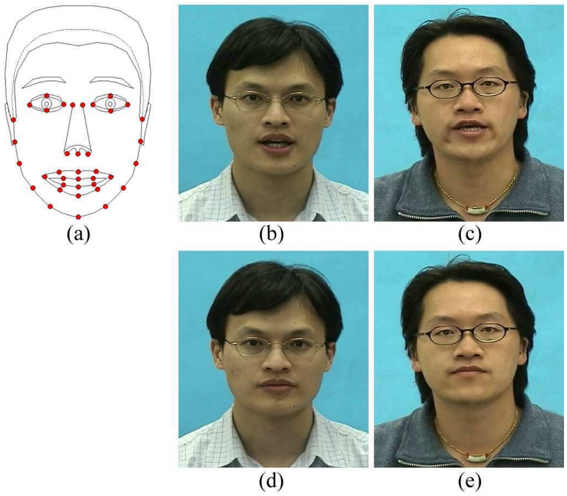
with

$$Q(\boldsymbol{p}) = (c_{00} + c_{01}p_x + c_{02}p_y, c_{10} + c_{11}p_x + c_{12}p_y)^T,$$
$$\boldsymbol{\lambda}_k = (\lambda_{k,x}, \lambda_{k,y})^T,$$

subject to $S(\boldsymbol{p}_k^a) = \boldsymbol{p}_k^b$, and side conditions imposed on the coefficients $\{\boldsymbol{\lambda}_k\}$:

$$\sum_{k=1}^{N} \boldsymbol{\lambda}_k \begin{bmatrix} 1 & p_{k,x}^a & p_{k,y}^a \end{bmatrix} = \boldsymbol{0}$$

where $\boldsymbol{p}_k^a$ and $\boldsymbol{p}_k^b$ are the k-th feature point in Ref_A and Ref_B, respectively, and $\phi(r) = \exp(-cr^2)$ is utilized as the radial basis function.


(a) (b) (c)

(d) (e)

Figure 2: *(a) Feature point configuration for establishing the dense point correspondence across different subjects; The reference images of (b) the original user, and (c) the novel subject with mouth open; The reference images of (d) the original user, and (e) the novel subject with mouth closed.*

4.2. Flow Matching

Having established correspondence between reference frames from the two corpora, the goal of the flow matching step is to choose a set of initial prototype candidates I_{B,P_i} from the novel corpus which correspond in *shape* to the prototypes I_{A,P_i} from the original corpus.

With the RBF-based interpolation function in Equation 6, given a flow vector in Ref_A started from position $\boldsymbol{p}$ and moved to $\boldsymbol{p}' = \boldsymbol{p} + C_A(\boldsymbol{p})$, the corresponding flow vector in Ref_B will be started from position $S(\boldsymbol{p})$ and moved to $S(\boldsymbol{p}')$. Hence, the synthetic flow vector at $S(\boldsymbol{p})$ of Ref_B will be $C_B^{syn}(S(\boldsymbol{p})) = S(\boldsymbol{p}') - S(\boldsymbol{p})$. Thereby, the synthetic flow of each prototype images can be generated for the novel person. By calculating the differences between the synthetic flow in the mouth region with flow vectors of each image in the new video corpus, the best candidate can be found with the minimal flow differences. That is,

$$P_k^* = \arg\min_i \sum_{\boldsymbol{p}} w_f(\boldsymbol{p}) \left\| C_{B,P_k}^{syn}(S(\boldsymbol{p})) - C_{B,i}(S(\boldsymbol{p})) \right\|, \tag{7}$$

where $w_f(.)$ is a weighting mask emphasizing the lip region, C_{B,P_k}^{syn} is the synthetic flow for the k-th prototype image, and $C_{B,i}$ is the flow of the i-th image of the small video corpus of the novel person. The best candidates obtained from flow matching form the initial candidates for the following texture matching step.

4.3. Texture Matching

While the initial prototype candidates from the flow matching step are good, there is a need to further refine the candidates in an effort to improve the alignment of original and novel speaker MMM's. The texture matching step performs this alignment by trying to match the "texture coordinates" between original and novel speakers.

Since the prototype images in the MMM of the original user are selected by the k-means clustering algorithm [Bis95], they are not orthogonal both in flow and texture space. Hence, similar to Equation 2, we can formulate one prototype image of the original speaker as a linear combination of the other $(M-1)$ prototype images of the original speaker:

$$\begin{aligned} I^{syn}_{A,P_k} &= \sum_{i,i\neq k} \beta_{i,k} I^{warped}_{A,P_i\rightarrow P_k} \\ &= \sum_{i,i\neq k} \beta_{i,k}\mathbf{W_F}\left(I_{A,P_i}, \mathbf{W_F}\left(C_{A,P_k}-C_{A,P_i}, C_{A,P_i}\right)\right), \quad (8) \end{aligned}$$

subject to

$$\beta_{i,k} \geq 0 \; \forall i \quad \text{and} \quad \sum_{i,i\neq k} \beta_{i,k} = 1,$$

where the *"texture coordinates"* $\{\beta_{i,k}\}$ can be derived by minimizing the difference between the synthetic image I^{syn}_{A,P_k} and the k-th prototype image I_{A,P_k} of the original user with quadratic programming methods.

Our hypothesis is that the synthetic prototype image of a novel person should be generated with the same texture coordinates $\{\beta_{i,k}\}$ as the corresponding prototype of the original speaker:

$$\begin{aligned} I^{syn}_{B,P_k} &= \sum_{i,i\neq k} \beta_{i,k} I^{warped}_{B,P_i\rightarrow P_k} \\ &= \sum_{i,i\neq k} \beta_{i,k}\mathbf{W_F}\left(I_{B,P_i}, \mathbf{W_F}\left(C^{syn}_{B,P_k}-C_{B,P_i}, C_{B,P_i}\right)\right) \quad (9) \end{aligned}$$

where I_{B,P_i} is the texture of the i-th prototype image of the novel person selected by flow matching, and C^{syn}_{B,P_k} is the synthetic flow derived in Section 3.2.

Hence, similar to Equation 7, the texture matching can be performed by calculating the differences between the synthetic texture with texture of each image in the new video corpus,

$$P_k^{**} = \arg\min_i \sum_{\mathbf{p}} w_t(\mathbf{p}) \left\| I^{syn}_{B,P_k}(\mathbf{p}) - I_{B,i}(\mathbf{p}) \right\|, \quad (10)$$

where $w_t(.)$ is a weighting mask emphasizing the mouth region, and $I_{B,i}$ is the texture of the i-th image of the small video corpus of the novel person.

Note that the change of one candidate prototype image may affect the texture synthesis of other prototype textures; iterative updating for texture matching is this required.

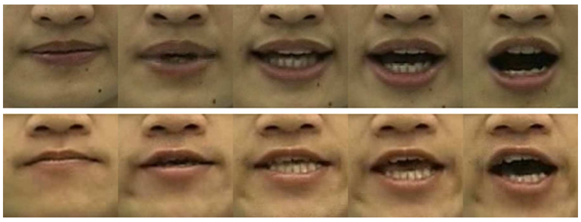


Figure 3: *Partial prototype image matching results: the prototype images of the original user with different degrees of mouth openness(top), and the corresponding new prototype images selected by the proposed algorithm(bottom).*

Moreover, the texture matching (Equation 10) and the flow matching (Equation 7) can also be linearly combined as

$$\begin{aligned} P_k^{***} = \arg\min_i \Big(& t\cdot \sum_p w_f(\mathbf{p}) \left\| C^{syn}_{B,P_k}(S(\mathbf{p})) - C_{B,i}(S(\mathbf{p})) \right\| \\ & +(1-t)\cdot \sum_p w_t(\mathbf{p}) \left\| I^{syn}_{B,P_k}(\mathbf{p}) - I_{B,i}(\mathbf{p}) \right\| \Big), \end{aligned}$$

with a scalar t to adjust the importance between flow matching and texture matching, such that both kinds of similarities can be taken into account for best candidate selection. With equal preference on flow and texture matching, t is set to be 0.5 in our experiments.

Furthermore, the use of multiple reference images is also possible. It potentially can achieve better synthesis quality, especially for texture synthesis. However, this would increase the effort for manual initialization. In our preliminary experiments, two sets of reference images as shown in Figures 2(b)-(e) are utilized to obtain good synthetic textures for matching.

Experimentally, combined texture and flow matching converges within 10 iterations, after which the best prototype candidates become stable. Partial prototype matching results are shown in Figure 3,revealing the proposed matching-by-synthesis algorithm can capture the subtle change of mouth dynamics. The flows and textures of these prototype images will form the MMM basis of the novel person. And phoneme models from the original person can be directly used for speech animation synthesis.

5. Model Adaptation

After steps of the model transfer performed in the previous section, the MMM prototypes are replaced by the images of the novel person while the phoneme model is directly transferred from the original user. Although the synthesized speech animation will be animated with the novel person's face, it actually behaves with the speaking style of the original user. Thus, there is a need to adapt the phoneme model to the speaking style of the new user.

In speech recognition applications, acoustic models and

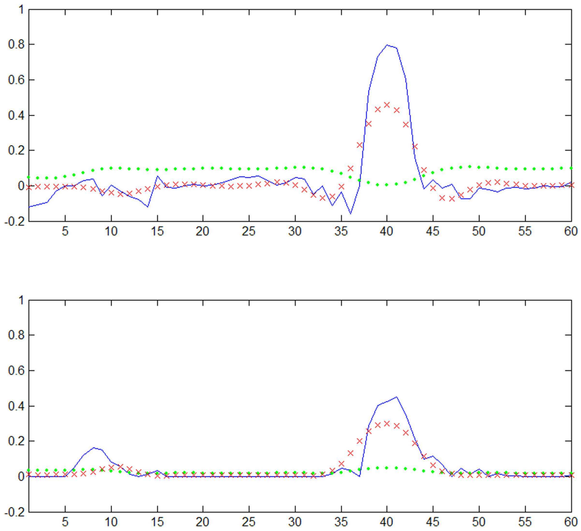


Figure 4: *Trajectory synthesis for partial MMM parameters.* ***Top****: The analyzed trajectory for the 39th flow coefficient* α_{39} *(in solid blue) compared with the synthesized* α_{39} *before adaptation (in green dots) and after adaptation (in red crosses).* ***Bottom****: Same as above, but the trajectory is for the 39th texture coefficient* β_{39}*. Both trajectories are from the two digits "nine" and "ten".*

language models are often adapted to reduce the mismatch in microphone, transmission channel, environment noise, speaker, style, and application contexts [HAH01]. For acoustic model adaptation, adaptive techniques such as MAP (maximum a posteriori) [GL94], MLLR (maximum likelihood linear regression) [LW95], and clustering adaptation [Gal98, KNJ*98] are utilized to modify model parameters with a small amount of adaptation data. Within these approaches, MLLR adaptation outperforms MAP when the amount adaptation data is small, and it does not require training data from a large variety of speakers as in clustering adaptation techniques.

Analogously, the phoneme model of the morphable model can be adapted from the small video corpus to make the synthesized animation more similar to the novel subject. However, while in speech recognition the acoustic model parameters are adapted to maximize the likelihood of the adaptation data, in speech animation synthesis, we would like to adapt the phoneme models such that MMM trajectories from the novel corpus are better reconstructed as depicted in Figure 4. Consequently, we propose a linear regression algorithm with gradient descent learning to adapt the MMM phoneme models.

Our algorithm consists of three parts: first, an *MMM reselection* step is performed to re-select a set of prototypes which best reconstruct the adaptation corpus; secondly, the phonemes are grouped together under a set of *regression classes* in order for learning to proceed even when no data is available for a particular phoneme in the adaptation corpus; finally, the *gradient descent linear regression* is performed as the last step which adapts the means of each regression class. In the following sections, we describe each step in detail.

5.1. MMM Re-Selection

The prototype images $\{I_{P_i}\}_{i=1}^{M}$ semi-automatically selected in Section 4 are chosen to be as similar in shape to the prototype images of the original person's morphable model. However, this criterion for choosing the prototypes may not be optimal in reconstructing the mouth appearance of the novel subject in the adaptation corpus. Consequently, the MMM Re-Selection step selects a new set of prototypes $\{I_{P'_i}\}_{i=1}^{M}$ which best reconstruct the adaptation corpus, and then rewrites the phoneme models in terms of this new basis.

Similar to the normal procedures of the MMM construction as stated in Section 3.1, the k-means clustering algorithm [Bis95] is utilized to select representative images as new prototype images $\{P'_k\}_{k=1}^{M}$. Each of the old prototypes $\{P_k\}_{k=1}^{M}$ may thus be analyzed in terms of the new prototypes $\{P'_k\}_{k=1}^{M}$ as:

$$C_{P_k} = \sum_{i=1}^{M} \alpha_{i,k} C_{P'_i}, \tag{11}$$

and

$$I_{P_k} = \sum_{i=1}^{M} \beta_{i,k} I^{warped}_{P'_i}. \tag{12}$$

Hence, to substitute the MMM with the new prototype images, the mean and covariance of phoneme models should be updated. Given a phoneme p with mean μ_p and covariance matrix Σ_p , the updated mean and covariance matrix is:

$$\mu'_p = \mathbf{M_{AB}}\mu_p \quad \text{and} \quad \Sigma'_p = \mathbf{M_{AB}}\Sigma_p\mathbf{M}^T_{\mathbf{AB}}, \tag{13}$$

with the transform matrix formulated as

$$\mathbf{M_{AB}} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{bmatrix},$$

where elements of matrices **A** and **B** are $\alpha_{i,j}$ and $\beta_{i,j}$ obtained in Equations 11 and 12.

5.2. Regression Classes Clustering

The idea of linear regression is to use a common linear transform to adapt the mean vectors of multiple components of the model when only small amount of adaptation data is available. Since some components may have quite different characteristics from other components, regression classes are clustered such that the same regression matrix is shared with the components belonging to the same regression class. In our work, phoneme models are clustered based on their MMM parameters using the k-means clustering algorithm [Bis95]. Nine regression classes are derived from 40 phonemes including the silence (SIL) and breath (BR): {AO, W}, {AA, AE, AH, AX, EH, HH, BR}, {B, M, P}, {F, V}, {AXR, D, DH, K, S, T, TH, Z}, {DX, G, IH, IX, IY, N, NG, Y}, {ER, L, R, UH}, {CH, JH, SH, UW, ZH}, {SIL}. As expected, these classes roughly exhibit similar structures as viseme groups.

5.3. Gradient Descent Linear Regression

For each phoneme p in the same regression class g, the mean vector μ_p is adapted by a common regression matrix R_g such that $\mu_p^{adapt} = R_g \xi_p$, where the extended mean vector is formed by $\xi_p = [\ 1 \quad \mu_p\ ]^T$. Different from the conventional MLLR [LW95] where the optimal regression matrix is obtained by maximizing the likelihood of the adaptation data, we seek to find the regression matrix that can minimize the error between the synthesized MMM parameters and the real parameters of the adaptation data. Given the regression matrices, the synthesized MMM parameters $\boldsymbol{y}$ are obtained by minimizing the modified objective function:

$$E_s = (\boldsymbol{y} - \mathbf{R}\boldsymbol{\xi})^T \mathbf{D}^T \boldsymbol{\Sigma}^{-1} \mathbf{D} (\boldsymbol{y} - \mathbf{R}\boldsymbol{\xi}) + \lambda \boldsymbol{y}^T \mathbf{W}_k^T \mathbf{W}_k \boldsymbol{y}, \quad (14)$$

where $\boldsymbol{\xi}$ is the cascaded extended mean vector, and $\mathbf{R}$ is the sparsely cascaded regression matrix. By taking the derivative and minimizing yields the following equation for optimal synthesized MMM parameters:

$$\left(\mathbf{D}^T \boldsymbol{\Sigma}^{-1} \mathbf{D} + \lambda \mathbf{W}_k^T \mathbf{W}_k\right) \boldsymbol{y} = \mathbf{D}^T \boldsymbol{\Sigma}^{-1} \mathbf{D} \mathbf{R} \boldsymbol{\xi}. \quad (15)$$

Instead of adapting the mean and covariance of each phoneme model as in Equation 5, the regression matrix for each regression class g is adapted by gradient descent learning. With the objective function $E_a = (\boldsymbol{z} - \boldsymbol{y})^T (\boldsymbol{z} - \boldsymbol{y})$, the gradient between E_a and the regression matrix R_g can be derived by chain rule:

$$\frac{\partial E_a}{\partial R_g} = \left(\frac{\partial E_a}{\partial \boldsymbol{y}}\right)^T \left(\frac{\partial \boldsymbol{y}}{\partial R_g}\right),$$

where $\partial E_a / \partial \boldsymbol{y}$ can be easily obtained by

$$\frac{\partial E_a}{\partial \boldsymbol{y}} = -2(\boldsymbol{z} - \boldsymbol{y}),$$

and $\partial \boldsymbol{y} / \partial R_g$ can be derived from Equation 15:

$$\left(\mathbf{D}^T \boldsymbol{\Sigma}^{-1} \mathbf{D} + \lambda \mathbf{W}_k^T \mathbf{W}_k\right) \frac{\partial \boldsymbol{y}}{\partial R_g} = \mathbf{D}^T \boldsymbol{\Sigma}^{-1} \mathbf{D} \frac{\partial \mathbf{R}}{\partial R_g} \boldsymbol{\xi}. \quad (16)$$

Then, each regression matrix is updated with the computed gradient by

$$R_g^{new} = R_g^{old} - \eta \frac{\partial E_a}{\partial R_g}, \quad (17)$$

with a small learning rate parameter η, and the initial regression matrix R_g^{old} is set to be the identity transformation which consists of an $M \times M$ identity matrix augmented by a column of zeros on the left side. Afterwards, the mean vector of phonemes in the same regression class is adapted accordingly.

6. Experimental Results

Experiments are conducted to verify the performance of the proposed model transfer and adaptation approach. Two video corpora with different amount of data are collected as the experimental data.

The first corpus is a 10-minute video recorded from a male sitting statically in front of the camcorder, uttering 304 English words. Based on the procedures of [EGP02], a set of 50 prototype images are selected by k-means clustering algorithm [Bis95] to construct the MMM. For phoneme models, each phoneme of the 40 phonemes stated in Section 5.2 is approximated by one Gaussian distribution with diagonal covariance matrix.

The second corpus is a short video clip recorded from another male under the same recording setup, uttering English digits from one to ten. The length of the small video is only 15 seconds (450 frames). Given such a short video clip, the average number of frames per phoneme is as few as 11.25 frames (3.87 frames if not take /SIL/ into account!) 21 of the 40 phonemes are not present. Except for the silence, other 39 phonemes all occupy less than 15 frames.

For initialization of model transfer, we setup 38 feature points for dense correspondence calculation between images from different persons. The configuration of these feature points is shown in Figure 2(a). No feature points are located on the eyebrows or forehead since the mouth region is the region of interest. To achieve better quality of synthesized flow and texture, two sets of reference images, one with mouth open and another with mouth closed, are used for prototype image matching as shown in Figures 2(b)-(e). Feature points are manually marked on these reference images for dense correspondence calculation with the RBF-based interpolation.

Partial prototype matching results from flow matching followed by six iterations of combined flow and texture matching are shown in Figure 3. It can be observed that the automatically selected new prototype images catch subtle changes of mouth dynamics quite well.

After prototype image matching, the MMM re-selection procedure is performed to replace the automatically selected prototype images with the images selected by k-means clustering algorithm. Afterwards, the phoneme models are grouped to nine regression classes for gradient descent linear regression, which converged around 30 iterations with the learning rate η equal to 0.005. Figure 4 depicts synthesized trajectories for some parameters before adaptation (in

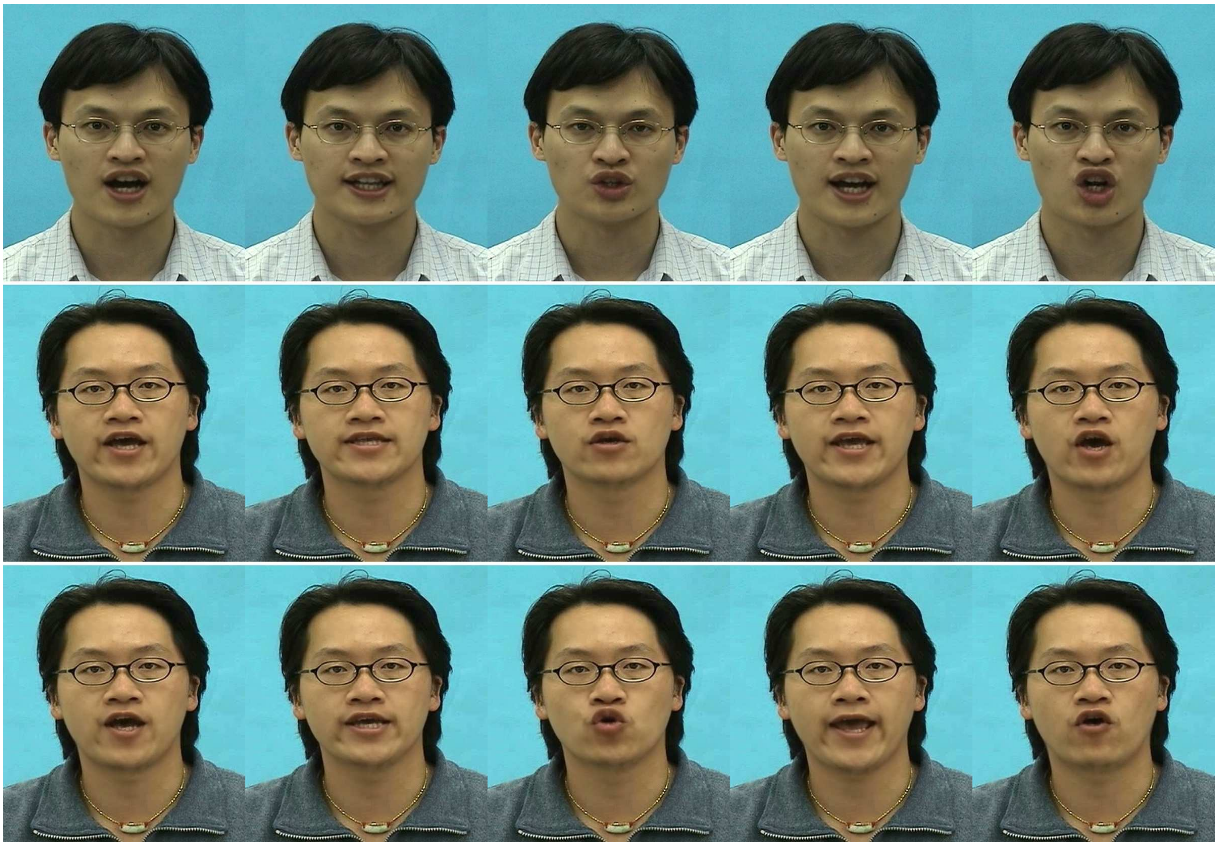


Figure 5: *Mouth image synthesis of five basic English vowels {/AA/, /IY/, /UH/, /EH/, /AO/} for the original user (top), and the novel subject before adaptation (middle) and after adaptation (bottom).*

green dots) and after adaptation (in red crosses) compared to the real trajectories (in solid blue). Figure 5 also shows the synthesized mouth images of five basic vowels in English before and after phoneme model adaptation. It is shown that the images synthesized from the adapted model exhibit more correct mouth shapes. By oberving the accompanied videos for sentence and song synthesis, the mouth dynamics synthesized with the adapted model apparently mimics the speaking style of the novel person.

7. Conclusions and Future Work

In this work, the framework of transferable videorealistic speech animation is proposed. Rather than transferring facial motions directly from one person to another, a generative morphable model is transferred to a novel person given a small video corpus. The contributions of this work are two-fold: First, a matching-by-synthesis method is utilized to choose a set of MMM prototypes from the novel speaker which match the MMM prototypes of the original speaker. Second, a model adaptation approach based on gradient descent linear regression is proposed to refine the phoneme models with the limited amount of video corpus as adaptation data. The synthesis from the adapted phoneme models learns the speaking style of the novel person. Good results are obtained from experiments performed successfully on an 15-sec. adaptation video, which justifies the applicability of the proposed method.

Currently, the synthesized mouth images are composited onto the same adaptation video corpus as the background sequence. Consequently, the head movement will seem unnatural with respect to the content of the speech. Furthermore, the length of the given small video may not be sufficient to synthesize a long speech utterance. The idea of combining visual prosody [GCSH02] and video texture [SSSE00] with the use of morphable models to generate background sequences with desired head movements is worthy of future investigation.

References

[BBPV03] Blanz V., Basso C., Poggio T., Vetter T.: Reanimating faces in images and video. In *Proc. Eurographics '03* (2003), vol. 22.

[BCS97] BREGLER C., COVELL M., SLANEY M.: Video rewrite: Driving visual speech with audio. In *Proc. SIGGRAPH '97* (1997), pp. 353–360.

[Bis95] BISHOP C. M.: *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[BP95] BEYMER D., POGGIO T.: Face recognition from one example view. In *Proc. IEEE 5th International Conference on Computer Vision* (1995), pp. 500–507.

[CC02] CHANG Y. J., CHEN Y. C.: Facial model adaptation from a monocular image sequence using a textured polygonal model. *Signal Processing: Image Communication 17*, 5 (May 2002), 373–392.

[CFKP04] CAO Y., FALOUTSOS P., KOHLER E., PIGHIN F.: Real-time speech motion synthesis from recorded motions. In *Proc. 2004 ACM SIGGRAPH/Eurographics Sympsium on Computer Animation* (2004), pp. 347–355.

[CG00] COSATTO E., GRAF H. P.: Photo-realistic talking-heads from image samples. *IEEE Trans. on Multimedia 2*, 3 (Sept. 2000), 152–163.

[EGP02] EZZAT T., GEIGER G., POGGIO T.: Trainable videorealistic speech animation. In *Proc. SIGGRAPH '02* (2002), vol. 21, pp. 388–397.

[Gal98] GALES M. J. F.: Cluster adaptive training for speech recognition. In *Proc. the 5th International Conference on Spoken Language Processing* (1998), pp. 1783–1786.

[GCSH02] GRAF H. P., COSATTO E., STROM V., HUANG F. J.: Visual prosody: facial movements accompanying speech. In *Proc. 5th IEEE International Conference on Automatic Face and Gesture Recognition* (2002), pp. 381–386.

[GL94] GAUVAIN J. L., LEE C. H.: Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Trans. on Speech and Audio Processing 2*, 2 (Apr. 1994), 291–298.

[Gle98] GLEICHER M.: Retargetting motion to new characters. In *Proc. SIGGRAPH '98* (1998), pp. 33–42.

[HAH01] HUANG X., ACERO A., HON H. W.: *Spoken language processing: a guide to theory, algorithm and system development.* Pearson Education, 2001.

[JP98] JONES M., POGGIO T.: Multidimensional morphable models: a framework for representing and matching object classes. *International Journal of Computer Vision 29*, 2 (Aug. 1998), 107–131.

[KNJ*98] KUHN R., NGUYEN P., JUNQUA J. C., GOLDWASSER L., NIEDZIELSKI N., FINCKE S., FIELD K., CONTOLINI M.: Eigenvoices for speaker adaptation. In *Proc. the 5th International Conference on Spoken Language Processing* (1998), pp. 1771–1774.

[LSZ01] LIU Z., SHAN Y., ZHANG Z.: Expressive expression mapping with ratio images. In *Proc. SIGGRAPH '01* (2001), pp. 271–276.

[LW95] LEGGETTER C. J., WOODLAND P. C.: Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language 9*, 2 (1995), 171–185.

[NJ04] NA K., JUNG M.: Hierarchical retargetting of fine facial motions. In *Proc. Eurographics '04* (2004).

[NN01] NOH J. Y., NEUMANN U.: Expression cloning. In *Proc. SIGGRAPH '01* (2001), pp. 277–288.

[PHL*98] PIGHIN F., HECKER J., LISCHINSKI D., SZELISKI R., SALESIN D.: Synthesizing realistic facial expressions from photographs. In *Proc. SIGGRAPH '98* (1998), pp. 75–84.

[SSSE00] SCHÖDL A., SZELISKI R., SALESIN D. H., ESSA I.: Video textures. In *Proc. SIGGRAPH '00* (2000), pp. 489–498.

[WHL*04] WANG Y., HUANG X., LEE C. S., ZHANG S., LI Z., SAMARAS D., METAXAS D., ELGAMMAL A., HUANG P.: High resolution acquisition, learning and transfer of dynamic 3-d facial expressions. In *Proc. Eurographics '04* (2004).

[ZLGS03] ZHANG Q., LIU Z., GUO B., SHUM H.: Geometry-driven photorealistic facial expression synthesis. In *Proc. 2003 ACM SIGGRAPH/Eurographics Sympsium on Computer Animation* (2003), pp. 177–186.