Lips-Sync 3D Speech Animation

Fu-Chung Huang* Bing-Yu Chen*‡ Yung-Yu Chuang* Shuen-Huei Guan†

*National Taiwan University ‡The University of Tokyo †Digimax

1 Introduction

Facial animation is traditionally considered as an important but tedious task for many applications.Recently the demand for lips-syncs animation is increasing, but there seems few fast and easy generation methods.In this talk, a system to synthesize lips-syncs speech animation given a novel utterance is presented. Our system uses a nonlinear blend-shape method and derives key-shapes using a novel automatic clustering algorithm. Finally a Gaussian-phoneme model is used to predict the proper motion dynamic that can be used for synthesizing a new speech animation.

2 Algorithm

The system is divided into three sub-systems. The first component identifies how many and what are the key-shapes used in the blend-shape process from a training video. The second component maps the face in the training video onto a newly created character. In addition, the parameters of the cross-mapped animation can be retrieved, and these values are analyzed for a Gaussian-phoneme model. Finally these phoneme-parameter relation can build a new speech animation by given a novel speech.

2.1 Key-Shape Identification

Given a training video, we want to determine what are the key-shapes that blend and span the original sequence. The problem can be explained as to approximate and reconstruct each frame f_j, we want to find certain number key-shapes $S_{1...K}$ using the blending parameters $w_{1j...Kj}$, and formulated as:

$$\min \sum_j || f_j - \sum_{i=1}^{K} S_i w_{ij} ||^2. \quad (1)$$

A dilemma exists in the problem, because the minimization objective contradicts to the choice of smaller K. In general, K is selected to be small enough as to be reasonable for future analysis, but not too small as to distort the reconstruction. Applicable algorithm such as k-means are widely used, but many tests for the number of K should be performed. Hence, an automatic clustering algorithm called Affinity Propagation by Frey and Dueck [2007] is used. After the identification process,we have certain amount of face images as key-shapes, and the artist can create corresponding 3D novel face models according to those key-shapes as Figure 1.

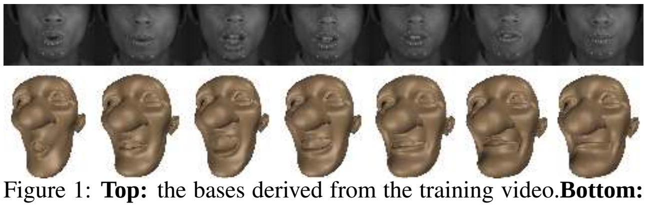

Figure 1: **Top:** the bases derived from the training video.**Bottom:** artist-sculpted models according to the bases.

2.2 Face Cross-Mapping

Face cross-mapping is a process that transfers the motion from one face to the other. To increase the data utilization for faithful result, we use MeshIK by Sumner *et al.* [2005], which takes the artist-sculpted examples as the prior. MeshIK estimates the position for the rest facial vertices using a non-linear (exponential-map) blending function in the gradient domain, and solves the exact values as Poisson equation does. Blending in the gradient domain and solving the position back preserve the local shape property and guarantee the control vertices condition, even there are very few examples available, which meets the first requirement for our system.

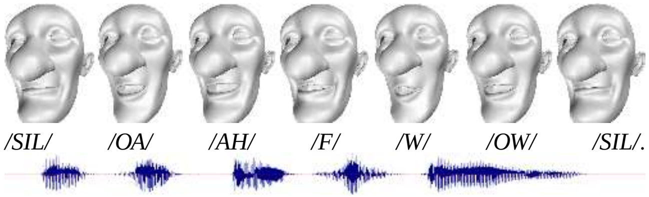

Figure 2: **Top:** the result of novel sequence ”O’er the fields we go” from ”Jingle Bells”. **Bottom:** the corresponding wave signal.

We first translate the training video into parameter space using the key-shapes found in Sec. 2.1 and solve with non-negative least square fitting for Eq. 1, thereafter each frame is associated with a set of blending coefficients, which are used to linearly combine the control vertices on the artist-sculpted models. Finally these control vertices and sculpted models are fed into MeshIK to estimate the rest vertices, and the result can be newly synthesized corresponding to the training faces and their coefficients for composition.

2.3 Motion Analysis and Synthesis

Having the cross-mapped novel speech animation (and the MeshIK coefficients), a Gaussian model of every pre-defined phoneme set can be analyzed. For each phoneme, we calculate the mean and diagonal covariance from the coefficients.

Given a new speech composed of a sequence of phoneme, a trajectory, as used in [Ezzat et al. 2002], going through each phoneme is obtained and a lips-sync speech animation is synthesized using such information. Finding such trajectory can be formulated as minimizing a regularization problem, as the following equation, with a data term describing the objective sticking with the coefficient of the current phoneme, and a smooth term requiring the animation having a natural transition in motion.

$$E = \sum_{t=0}^{T} (x_t - \mu_i)^T D_i^T \Sigma_i^{-1} D_i (x_t - \mu_i) + \lambda \mathbf{X}^T W^T W \mathbf{X}, \quad (2)$$

where the trajectory x_t at time t must passes through the i-th phoneme region influenced by its inverse phoneme duration D_i and normalized by its diagonal covariance Σ_i. We concatenate each frame x_i as $\mathbf{X}$, with W to calculate first order difference, so the frame-by-frame dissimilarity can be performed to measure the motion smoothness. Co-articulation effects are modeled through the covariance matrix Σ_i and self-multiplied difference matrix W.

3 Result

In the accompanying video, several novel spoken utterances are synthesized.In Figure 2 (also in the video), a singing sequence is synthesized with the faces corresponded to the certain phonemes.

References

Ezzat, T., Geiger, G., and Poggio, T. 2002. Trainable videorealistic speech animation. In *SIGGRAPH 2002*, 388–398.

Frey, B. J., and Dueck, D. 2007. Clustering by passing messages between data points. *Science 315*, 5814.

Sumner, R. W., Zwicker, M., Gotsman, C., and Popović, J. 2005. Mesh-based inverse kinematics. *ACM TOG 24*, 3, 488–495. (SIGGRAPH 2005).

*e-mail: {jonash,robin,cyy}@cmlab.csie.ntu.edu.tw

†e-mail: drake.guan@gmail.com