Animating Lip-Sync Speech Faces by Dominated Animeme Models

Fu-Chung Huang* Yu-Mei Chen† Tse-Hsien Wang† Bing-Yu Chen† Shuen-Huei Guan‡

*University of California at Berkeley †National Taiwan University ‡Digimax

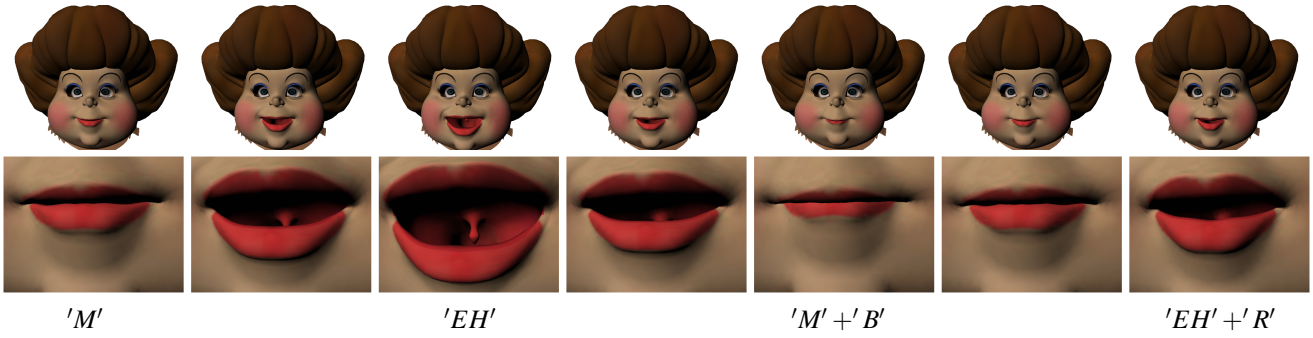

Figure 1: The result of speaking the 6 phonemes of the word - "member". Note that although the 1st and 3rd phonemes are the same, the lips shapes are different due to the *coarticulation effect*.

1 Introduction

Speech animation is traditionally considered as important but tedious work for most applications, because the muscles on the face are complex and dynamically interacting. In this paper, we introduce a framework for synthesizing a 3D lip-sync speech animation by a given speech sequence and its corresponding texts. We first identify the representative key-lip-shapes from a training video that are important for blend-shapes and guiding the artist to create corresponding 3D key-faces (lips). The training faces in the video are then cross-mapped to the crafted key-faces to construct the Dominated Animeme Models (DAM) for each kind of phoneme. Considering the coarticulation effects in animation control signals from the cross-mapped training faces, the DAM computes two functions: polynomial-fitted animeme shape functions and corresponding dominance weighting functions. Finally, given a novel speech sequence and its corresponding texts, a lip-sync speech animation can be synthesized in a short time with the DAM.

2 Overview

As shown in Figure 2, the framework can be divided into 2 subsystems. The first one learns the phoneme-animeme relationship with face cross-mapping functionality, since the training face is different from the target face. Both subsystems use SPHINX-II as phoneme alignment tool. The proposed DAM takes a time-aligned script and animation control signals as the input. In the end of training, DAM learns the animeme shape function of the animation control signals for phonemes, and the dominance weighting functions for interference. In the second subsystem, the DAM generates the animation control signals from an arbitrary speech, which can be used to generate the output animation. The major contribution of our framework is how the DAM learns the animation control signals for each phoneme and the interference (the coarticulation effect) among them.

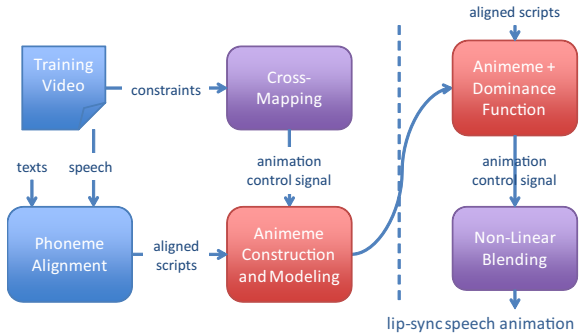

Figure 2: System flowchart. The phoneme-animeme relationship is learned in the training phase, and the speech animation is generated in the synthesis phase.

3 Dominated Animeme Model

To model the animation control signals across time, we use a mathematical approach to describe the animemes and interference among them. The animeme, approximated by an M-order polynomial, is the shape function for a sequence of animation control signals for a specific phoneme. The interference among animemes, called coarticulation effects, are modeled through weighted Gaussian dominance function. For a resulting animeme, the output signals are the convolution of its animeme and dominance functions. For a sentence composing several animemes, the signals are simply the summation of these convoluted signals. In mathematical expression, at time t^i the control signal w^i is the summation of J animemes in a sentence, given by:

$$w^i = \sum_{j=i}^{J} \mathbf{D_j}(t_j^i) \left[\sum_{m=0}^{M} a_j^m (t_j^i)^m \right],$$

where t_j^i is the local-frame transformed time t^i, and a_j^m is the M-order polynomial coefficient for j-th animeme. The dominance function $\mathbf{D_j}(t_j^i)$ is given by:

$$\mathbf{D_j}(t_j^i) = \exp\left\{ \frac{(t_j^i - \mu_j)}{(d_j \times \sigma_j)^2 + \varepsilon} \right\},$$

where μ_j is the center time of the occurring animeme, d_j is the duration, and σ_j is the animeme specific constant that controls the span of influence.

4 Result

We compare our result, shown in Figure 3, with other methods by Cohen-Massaro [1993] and Multi-dimensional Morphable Model (MMM) [Ezzat et al. 2002] along with the original animation control signals. The result by Cohen-Massaro has poorly modeled span and low synthesized control signals that lead to overly smoothed animation. MMM produces good span timing but still suffers from low control signals. Notice that our model not only synthesizes better signal values, but also appropriately preserves the span of each occurrence. In our accompanying video, our method preserves better coarticulation effects and characteristics from each animeme.

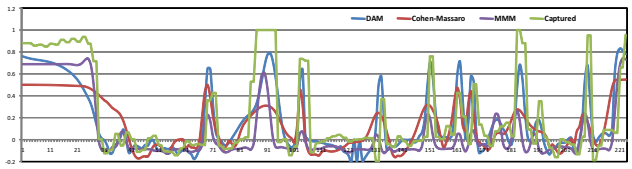

Figure 3: Comparison with different strategies. L^2 norms for DAM, Cohen-Massaro, and MMM are 0.0524, 0.0858, and 0.0793.

References

COHEN, M. M., AND MASSARO, D. W. 1993. Modeling coarticulation in synthetic visual speech. In *Computer Animation 1993 Conference Proceedings*, 139–156.

EZZAT, T., GEIGER, G., AND POGGIO, T. 2002. Trainable videorealistic speech animation. *ACM Transactions on Graphics 21*, 3, 388–398. (SIGGRAPH 2002 Conference Proceedings).