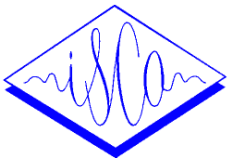

The Third ESCA/COCOSDA Workshop
(ETRW) on Speech Synthesis
Jenolan Caves House, Blue Mountains,
NSW, Australia, November 26-29, 1998

CLONING SYNTHETIC TALKING HEADS

Jialin Zhong *Joseph Olive*

Language Modeling Department
Multimedia Communication Research Lab.
Bell Labs, Lucent Technologies
Murray Hill, NJ 07974
{jlzhong, jpo} @research.bell-labs.com

ABSTRACT

The quality of Text-to-Visual-Speech synthesis is judged by how well it matches the visual perception of speech articulators with acoustic speech perception. Concurrently, different viewers often prefer different head models for subjective reasons. Traditional facial animation approach tied the parameterization of animation directly to the model. Switching the head model is difficult because a lengthy training process is required. In this paper, we present a method that creates a new talking head from an existing one without repeating the training process. It is assumed in this work that the visible motion of speech articulators can be described by a small set of feature points. By mapping the 3D trajectories of the feature points from the existing model to the new one, we can transfer the motion of articulators. A morphing algorithm is then used to animate a new talking head from these trajectories of feature points on the new model. The new talking head, though looking different, preserves the perceptual quality of the original one.

1. INTRODUCTION

Visual speech synthesis, or creating a synthetic talking head, was traditionally achieved through a direct parameterized model [1], where the underlying 3D wireframe model is animated through a set of parameters that correspond to the facial gestures correlated with speech production. Typical parameters include jaw rotation, mouth opening, tongue movement, etc.. Each animation parameter explicitly controls the positioning of a subset of vertices on the head model. Visual speech synthesis is then obtained by changing these control parameters according to rules that have been generated through a training process. The advantage of this direct parameterized approach is that it can animate a wireframe model with precise motion for speech articulators. However, the parameterization of visual speech synthesis is tied directly to the model. Since this direct parameterized approach requires a tedious training process to tune the parameters to the model, most 3D talking heads are based on a single wireframe model, the famous Parke's model. It is clearly desirable to have the capability of animating a generic wireframe model of human head while preserving the quality of visual speech of Parke's model.

In this paper, we describe an approach that creates a synthetic talking head from a previously existing one, such as the one using Parke's direct parameterized model. The main idea is to extract a description of the motion of speech articulators from the known talking head and then use this information to animate a generic head model. The information extracted in this work is the trajectories of a set of feature points on speech articulators. The facial expressions that are related to human speech production are the motion of lips and the parts that surround the lips. The feature points used for animation are a set of inner and outer lip contour points plus the tip of the chin and limit points, such as the tip of nose and the sides of face. It is assumed that these feature points contain sufficient information to recover the dynamics of synthetic visual speech from the original talking head. We call the process of creating a new talking head "cloning" since correct motion related to speech production is transferred from the original to the new model.

The cloning process includes the following steps. First, we identify the positions of corresponding feature points on the direct parameterized head model, such as the Parke's model and the generic one, at neutral facial expression. These parameters, called Facial Definition Parameters (FDP), define the initial positions of those feature points. Secondly, the trajectories of feature points on the trained model are computed while synthesizing visual speeches. The trajectories are measured in terms of the feature point positions relative to the FDPs, called Facial Animation Parameters (FAP). The values of the FAPs are normalized with respect to the mouth width in x axis and to the distance between mid mouth point and the nose tip in the y axis. Using the trajectories defined by the normalized FAPs from the known talking head as media, we can then map out the trajectories of feature points on the generic head model we want to animate. Thirdly, a 3D image morphing algorithm is used to compute the positions of the vertices of the generic head model from feature points during visual speech synthesis. The morphing algorithm is applied to a group of subregions within which each vertex is moved according to the feature points that define that region. Through this process, we can create a different talking head with the quality of visual speech preserved.

2. PERCEPTION AND REPRESENTATION

2.1. Perceptions of visual speech

There are two dimensions to the quality of a talking head: the aesthetic quality of its graphics and the perceptual quality of visual speech. The aesthetic quality is subject to the preference of viewers. Different viewers may prefer different head models. The perceptual quality of visual speech is related to how well human observers match the perception of the synthetic visual speech with that of acoustic speech. This perceptual quality depends on the synthesis of visemes and their coarticulations, and it directly affects the lip-readability of a talking head.

It is our objective to preserve the perceptual quality of the synthetic visual speech when cloning a generic talking head from an existing one. To achieve this objective, we conjecture in this work that there exists an inherent representation, in human perception of synthetic visual speech, that contains sufficient information for human to match the visual speech perception with acoustic one, and that other peripheral information can be extrapolated from this essential representation without adversely affecting the quality of visual speech perception. It is clear that this representation should be independent of speakers. For visual speech synthesis, this means that the representation is invariant to the underlying wireframe models. In this work, we use a set of feature points on the face and their trajectories as an description to the inherent representation for visual speech synthesis.

2.2. Feature points and their representations

Visual speech is perceived through the motion of visible speech articulators, including the lips, the tongue, the teeth and the chin. The feature points in this work are selected at strategic locations on these articulators. For this work, we use Facial Definition Parameters (FDP), defined in the MPEG-4 Synthetic and Natural Hybrid Coding document (MPEG4/SNHC) [4], as the feature points. For our application, we use the points in Group 2 and 8 of MPEG-4/SNHC as shown in Figure 1. FDP points in Group 2 consist of eight feature points on the inner lip contours and the tip of the chin; FDP points in Group 8 consist of ten feature points on the outer lip contours.

The trajectories of these feature points are characterized through a set of scalar quantities, called Facial Animation Parameters (FAP), which measure the displacement of FDP points from their corresponding positions on the model with a neutral facial expression, i.e., when the mouth is closed without facial expression. The arrows in Figure 1 show the possible directions of FAP values. To make the measurement roughly invariant to individual wireframe model, the FAPs are measured in the x direction in terms of the mouth width at neutral expression and in the y direction in terms of the distance from the nose tip to the middle of the mouth. These two quantities are called Facial Animation Parameter Units (FAPU). The trajectories of these feature points are then quantified relative to their neutral positions through these normalized units.

3. PARKE'S MODEL

Parke in [2] pioneered facial animation using a direct parameterized method. Cohen and Massaro [3] incorporated a coarticulation model into Parke's model for real-time 3D Text-to-Visual-Speech synthesis. Figure 3 shows the Parke's model in the form of a wireframe mesh. The head model is animated through a set of control parameters, each of which, through a mapping function, specifies a definitive behavior for the face. For example, the parameter for jaw rotation represents the rotational angles of those vertices for the jaw with respect to a fixed axis. Changing the jaw rotational angle changes the extent of mouth opening. Figure 2 shows the Parke's model at different jaw rotational angles. Vectors of target values for those control parameters are then used to define visemes, – visually distinctive 3D positions of speech articulators that correspond to phonemes. Visual speech synthesis is then achieved through interpolating these visemes using a set of mixing functions [3]. The parameterization of interpolation functions is obtained through observing human speak and adjusting parameters manually. The advantage of this approach is that the visual speech synthesis has high perceptual quality because it has been tuned to speech perception experts. However, it is clear that the parameterization is tied to the Parke's model. To create a talking head on a different wireframe model, one needs to go through the training process that requires a lot of expertise and work.

4. CLONING TALKING HEADS

In this section, we use a feature-based approach to clone talking heads from the Parke's model. The advantage of this approach is that we can create a talking head without going through Cohen's parameterization process [3], while still preserving the high perceptual quality of Parke's model. We first extract the FDP feature points and their trajectories measured in terms of FAPUs from the Parke's model and then map this information to the new 3D wireframe model. Then we use a 3D feature-based morphing algorithm to animate the new model [5]. Figure 4 shows the schematic diagram of the cloning process. Since the tongue is not fully visible to a viewer, we concentrate our effort on the lips and their surrounding regions. In the following sections, we will discuss each step in the cloning process.

4.1. Feature mapping

It is clear that a given wireframe model will have quite different geometrical shapes and topological connections for its vertices from those of the Parke's model. To take advantage of existing knowledge in the Parke's model, we calibrate the new wireframe against the Parke's model when both are at neutral facial expressions. The FDP points on the new model correspond to those on the Parke's

model. The positions of these FDP points are obtained using an interactive GUI tool. For our application, only 25 FDP points need to be picked once at the preprocessing step. The trajectories of these feature points during visual speech synthesis are then computed as the FAPs, i.e., the displacement of FDP points relative to their initial positions. These FAPs are measured in terms of FAPUs to normalize the factor of physical scales. In this way, we can easy extract the trajectories of feature points on the Parke's model and map them to the new wireframe model.

4.2. Feature-based morphing

In this section, we summarize the procedures to animate a wireframe model from the trajectories of feature points. The trajectories on the new face model are obtained from mapping those on the Parke's model.

When people speak, the lip motions are generated by the contraction of facial muscles. Different muscles affect different parts of the face. For example, in lip rounding situations, both the left and right side of the mouth move toward the middle, and hence they have opposite motion directions. It makes sense to synthesize visual speech in local regions, so that within each region, the facial motions are conforming. In this work, we divide the mouth area into eight connected regions as shown in Figure 5. The four inner regions are for the lips, and the four outer ones are for the regions surrounding the lips. The borders of these regions are obtained by lines that connect FDP feature points on the inner and outer lip contours as well as the nose tip, the sides of the face and the chin tip. Using these borders as landmarks, we compute the 3D positions of vertices on the head model with a line feature-based morphing algorithms as described in [5]. It should be noted that for this line feature-based morphing algorithm, the motion is continuous at the common borders. Vertices that are out of the mouth region are assumed unrelated to the speech production and are not affected by the trajectories of feature points.

5. EXPERIMENT

In experiment, we used Bell Labs English Text-to-Speech (TTS) to drive the Parke's model as parameterized in [3]. The new model was an off-the-shelve"Child" head model. Following the schematic diagram shown in 4, in the pre-processing step, the FDP points were computed both for the Parke's model and the "Child" model at neutral expression. For each video frame, we extracted the FAPs from the Parke's model and then passed these values to the "Child" head model. The 3D feature-based morphing algorithm was then used to animate the "Child" model from the trajectories of its FDP points. Figure 6 shows a sequence of results. It can be seen that the mouth shapes of cloned "Child" model match those of Parke's model very well.

6. SUMMARY

To make a synthetic talking head look natural, the visual perception for speech articulators has to match acoustic perception of speech. Also subjecting to personal aesthetic preferences, different observers often prefer different head models. These two constraints represent the perceptual quality and graphic quality of a talking head. Unfortunately, in traditional direct parameterized method, creating a talking head on a different model requires a lengthy training process with human experts involved in the process. This training process makes it uneconomical to create new talking heads according to the preferences of viewers. In this paper, we presented a cloning approach that animated a generic head model using information existing in a synthetic talking head, such as the Parke's model. It is assumed that the visible motion of speech articulators can be characterized by a small set of parameters. Based on this assumption, we extracted a small set of normalized feature points on the face as our representation. Using these feature points as calibrated media, we developed a method to clone the Parke's model without losing its perceptual quality of synthetic visual speech. The advantage of our approach is that a talking head can be made from a model of viewer's choice without additional training cost, while the perceptual quality of visual speech in the Parke's model is preserved. In experiment, we demonstrated that this method can be used to create new talking heads out of generic wireframe models.

7. REFERENCES

1. F. Parke, "Computer Facial Animation", A. K. Peters, Wellesley, Mass., 1996.
2. F. Parke, "A Parametric Model for Human Faces", *Ph.D. thesis*, University of Utah, Salt Lake City, UT, Dec. 1997. UTEC-CSc-75-047
3. M. Cohen and D. Massaro, "Modeling Coarticulations in Synthetic Visual Speech", In N. M. Thalmann and D. Thalmann (Eds.) *Models and Techniques in Computer Animation*, Tokyo Springer-Verlag, 1993, pp. 139-156.
4. C. Horne (Eds.) *MPEG-4 SNHC working draft*, Sept. 1997.
5. J. Zhong, "Flexible Face Animation Using MPEG-4/SNHC Parameter Streams", *Proc. Internal Conference on Image Processing*, Chicago, Oct. 4-7, 1998.

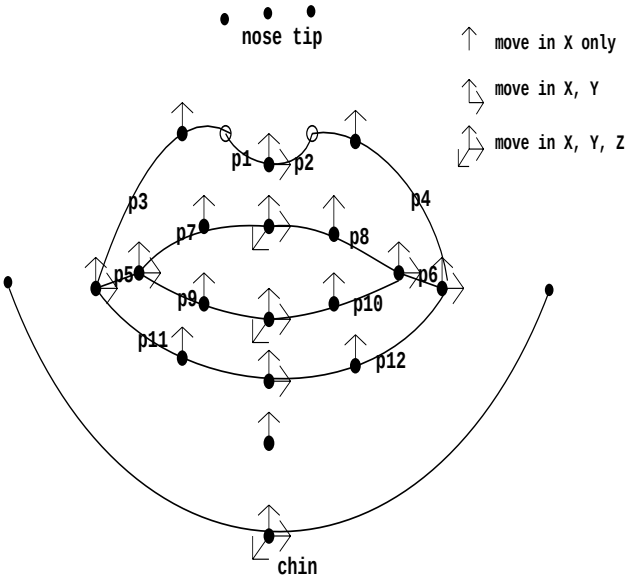

Figure 1: A set of feature points, FDP point, around the mouth region are used as an inherent representation. At each FDP points, the arrows represent the possible directions for FAP.

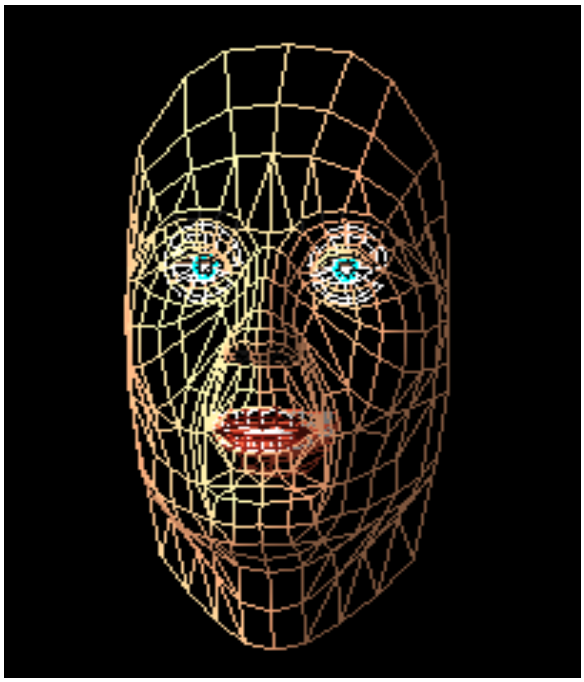

Figure 2: The underlying wireframe model for Parke's model.

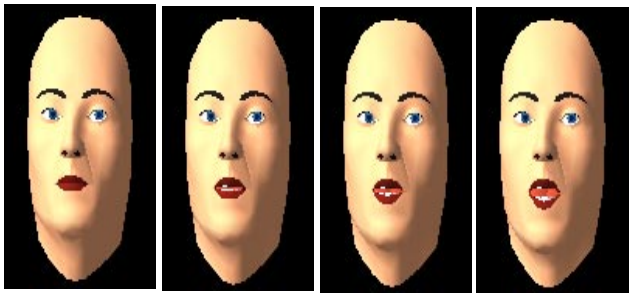

Figure 3: The jaw rotation parameter controls the angle of jaw rotation in Parke's model.

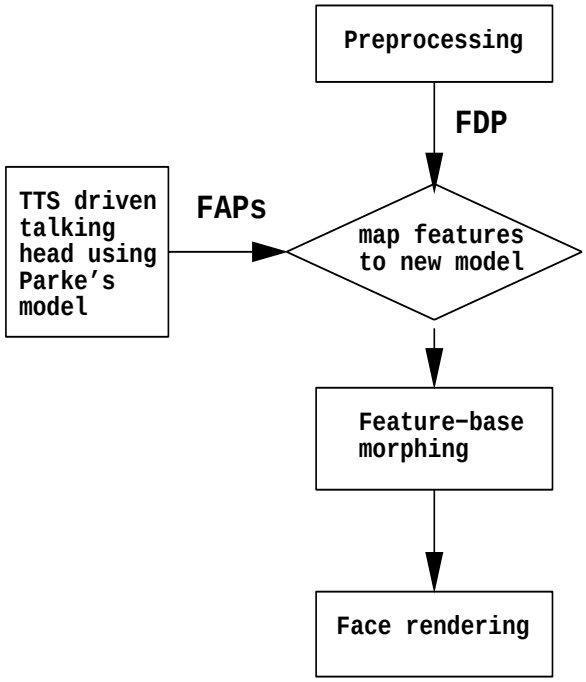

Figure 4: Schematic diagram for cloning a talking head form the Parke's model.

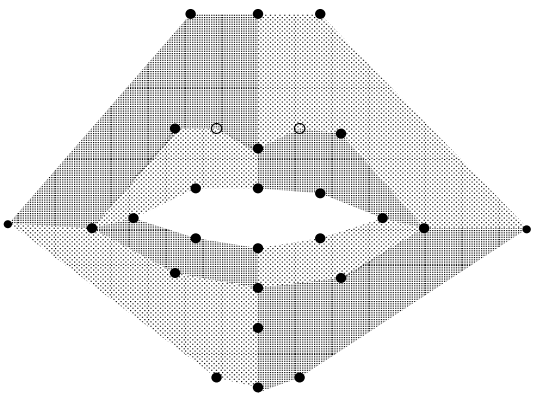

Figure 5: The mouth region is divided into eight sections, each of which is handled separately.

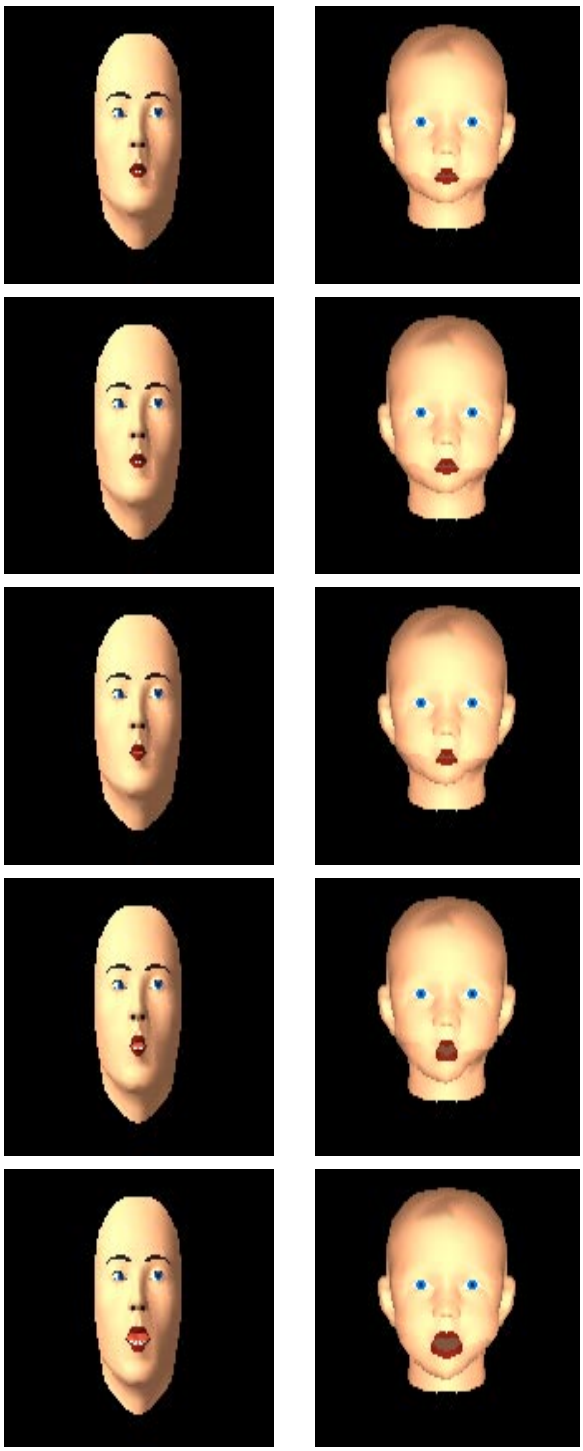

Figure 6: The left column is the visual speech synthesis based on Parke's direct parametric model; the right column is the results of 3D feature-based morphing algorithm using the inherent information extracted from Parke's model.