Chapter 1

TOWARDS VIDEOREALISTIC SYNTHETIC VISUAL SPEECH

Barry Theobald, J. Andrew Bangham, Silko Kruse, Gavin Cawley
University of East Anglia, Norwich, UK, NR4 7TJ
b.theobald@uea.ac.uk, ab@sys.uea.ac.uk, smk@sys.uea.ac.uk, gcc@sys.uea.ac.uk

Iain Matthews
Robotics Institute, Carnegie Mellon, Pittsburgh, PA 15123
iainm@cs.cmu.edu

Abstract

In this paper we present preliminary results of work towards a videorealistic visual speech synthesiser. A generative model is used to track the face of a talker uttering a series of training sentences and an inventory of synthesis units is built by representing the trajectory of the model parameters with spline curves. A set of model parameters corresponding to a new utterance is formed by concatenating spline segments corresponding to synthesis units in the inventory and sampling at the original frame rate. The new parameters are applied to the model to create a sequence of images corresponding to the talking face.

Keywords: Shape and appearance models, principal component analysis, visual speech synthesis, facial animation

1. Introduction

Research in computer facial animation began in the early seventies with the pioneering work of Parke [Parke, 1974] and has received increasing interest since [Parke and Waters, 1996]. Applications for facial animation systems are as widespread as surgical planning, low bandwidth video conferencing, media/film production, human-computer in-

teraction, virtual humans in computer games and psychological experiments in the perception and understanding of speech.

Much of the early work focussed on *computer graphics* based techniques. A popular approach is to represent points on the surface of the face as vertices in a 3D space and approximate the surface itself by connecting the vertices to form a polygonal mesh. Parameters to animate the mesh are either derived empirically through observation [Cohen and Massaro, 1994, Le Goff and Benoit, 1996, Parke, 1974] or are based on some anatomical model [Lee et al., 1993, Platt and Badler, 1981, Waters, 1987]. To make the skin appear more life-like a facial image can be used as a texture map. However, as the model is animated the static nature of the skin texture is revealed and even models with very complex animation control parameters do not convince the viewer that they are seeing a human face.

More recent systems have focussed on *image* based techniques in an attempt to achieve videorealism. Ezzat and Poggio [Ezzat and Poggio, 1997] used a variation of traditional *key-frame* animation, where new sequences are created by interpolating between key-poses in the image domain. Images were selected with key lip shapes and optical flow used to morph between key-frames. Cosatto and Graf [Cosatto and Graf, 1998] segmented the face and head into several distinct parts and collected a library of triphone samples. New sequences are created by stitching together the appropriate mouth shapes and corresponding facial features based on acoustic triphones and a desired facial expression. Bregler's Video Rewrite system [Bregler et al., 1997] automatically segmented video sequences of a person talking and could reanimate a sequence with different speech by blending images from the training sequence in the desired order.

Image based animation is generally less flexible than model based animation. A model based approach allows new facial movements to be created, or existing expressions to be exaggerated through manipulation of the model parameters. This is difficult to achieve with image based approaches which usually replay existing footage in a new order.

In the following section we describe the construction of a face model capable of producing near photorealistic images of the face. Subsequent sections then discuss how this model is applied to creating near videorealistic synthetic visual speech, where a new visual sequence is synchronised with a previously unseen audio signal.

2. Modelling the Face

Following the notation of Cootes [Cootes et al., 1998], a statistical model of shape, the *point distribution model* (PDM), is trained by hand labeling a set of images and performing a principal component analysis (PCA) on the coordinates of the located landmarks (aligned to remove any pose variation). Any training shape can be approximated using $\mathbf{x} \approx \overline{\mathbf{x}} + \mathbf{P}_s \mathbf{b}_s$. Where $\mathbf{P}_s$ is the matrix of the first t_s eigenvectors of the covariance matrix, chosen to describe some percentage, say 95%, of the total variation, and $\mathbf{b}_s$ is a vector of t_s shape parameters.

A texture model is computed by warping the training images so the landmarks in each image lie in the position of the mean landmarks derived from all training images. This normalises the shape of the face, which allows the image to be re-sampled with the same number of pixels in every example and a PCA is computed on the re-sampled pixel values. With such a model any texture can be approximated using $\mathbf{a} \approx \overline{\mathbf{a}} + \mathbf{P}_a \mathbf{b}_a$. Where $\overline{\mathbf{a}}$ is the mean shape-free image, $\mathbf{P}_a$ is the matrix of the first t_a eigenvectors of the covariance matrix and $\mathbf{b}_a$ a vector of texture parameters.

Each image is now described by a set of shape parameters and a set of texture parameters, $\mathbf{b}_s$ and $\mathbf{b}_a$ respectively. An appearance model is computed by concatenating the PCA scores for the shape and texture models for each image and performing a third PCA (in $t_s + t_a$ dimensions). Where the number of parameters, t_s and t_a, are chosen so that typically 95% of the variance of their respective models is captured. The appearance model is given by Equation 1.1:

$$\mathbf{b} \approx \mathbf{Q}\mathbf{c}, \tag{1.1}$$

where $\mathbf{Q}$ is the matrix of eigenvectors of the covariance matrix and $\mathbf{c}$ a vector of parameters that reflect changes in shape and texture of the face. A large range of realistic images can be synthesised given a set of the first t parameters using Equation 1.2.

$$\mathbf{x} \approx \overline{\mathbf{x}} + \mathbf{P}_s \mathbf{W}_s \mathbf{Q}_s \mathbf{c}, \quad \mathbf{a} \approx \overline{\mathbf{a}} + \mathbf{P}_a \mathbf{Q}_a \mathbf{c}, \tag{1.2}$$

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_s \\ \mathbf{Q}_a \end{pmatrix},$$

where the matrix $\mathbf{W}_s$ takes into account the scaling mismatch between the parameters $\mathbf{b}_s$ (which model Euclidean distance) and $\mathbf{b}_a$ (which model pixel RGB intensity). This is computed as shown in [Cootes et al., 1998] and facial animations are generated by controlling the time trajectory of the vector $\mathbf{c}$. Example faces synthesised by the model are shown in Figure 1.1.

Figure 1.1. The top row shows the face extracted from selected frames of the video and the bottom the synthetic equivalent output by the face tracker.

This approach differs from other techniques that have applied PCA to the problem of synthesising visual speech in that Hallgren [Hallgren and Lyberg, 1998] tracked physical markers on the face and applied PCA to the coordinates of the tracked markers. Guiard-Marigny [Guiard-Marigny et al., 1996] animated a geometric model of the lips by representing the contours of the lips with mathematical equations and iteratively predicting one coefficient from others. Brooke and Scott [Brooke and Scott, 1998] performed a PCA on the pixel intensities of the general mouth region and used the principal modes to train a hidden Markov model (HMM) synthesiser. Here PCA is used to construct a compact model of both the shape and the texture of the face. The two models are projected into a combined shape and texture space and a concatenation scheme is used to create new synthesised sequences.

3. Data Capture

The training database for the face model was collected using a ELMO EM-02PAL camera and digitized at a frame rate of 25 frames per second using an IEEE 1394 capture card with a frame size of 720x576 (colour). The audio was captured at 11025 Hz stereo and used to phonetically segment the video using an HMM speech recogniser run in forced-alignment mode.

To eliminate unwanted sources of variation from the model the video was recorded using a head mounted camera, recording a single talker in one sitting; eliminating pose, identity and lighting variations. The speaker held their facial expression as neutral as possible (no emotion) so the variation of the facial features were due to speech.

The training data consisted of 279 sentences containing multiple occurrences of 6315 triphones. The database contained over 30,000 frontal

images of the face. A shape model and a texture model were trained on 50 hand-labelled images and the remainder were automatically labelled using the flexible appearance model algorithm [Baker et al., 2001]. The tracker output was manually checked and corrected by hand where necessary.

4. Synthesis

Given the labelled images and the segmented audio, an inventory of synthesis units is constructed. Firstly the shape and texture models used by the tracker are projected into a combined shape and appearance space, as described in section 2. Next, the appearance model parameters are computed for each of the training images. Since the landmark positions are known, the shape parameters can be found using $\mathbf{b}_s = \mathbf{P}_s^T(\mathbf{x} - \overline{\mathbf{x}})$ and the texture parameters found using $\mathbf{b}_a = \mathbf{P}_a^T(\mathbf{a} - \overline{\mathbf{a}})$. The shape and texture parameters are concatenated to form a vector $\mathbf{b}$ and the appearance parameters computed using $\mathbf{c} = \mathbf{Q}^T\mathbf{b}$, from Equation 1.1.

The appearance model required 77 parameters to account for 99% of the total variation in the training images and the particular variations captured by the first 3 modes are shown in Figure 1.2.

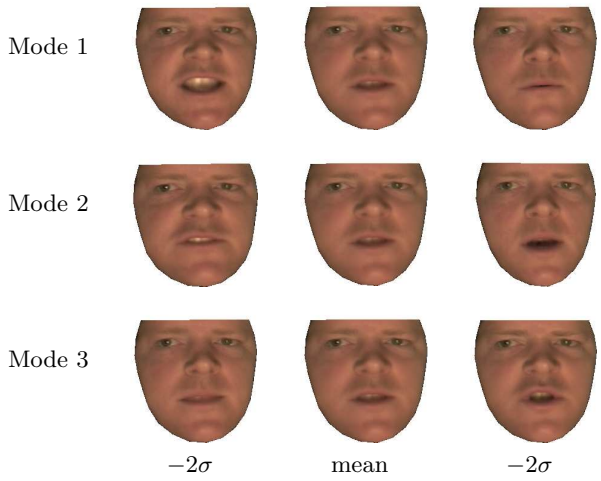

Figure 1.2. The first three modes of variation captured by the combined shape and appearance model at -2, 0 and +2 standard deviations from the mean.

Next, a continuous representation of the trajectories of the original model parameters in each sentence is obtained using Hermite interpolation. A look-up table is created using the speech recogniser output;

for each phone the start time in the sentence, the end time, the context and an index indicating in which sentence the phone occurred are stored. The context width is taken as one phone either side of centre phone. Triphones were selected as the synthesis unit as these were the units used by Bregler in the Video Rewrite system [Bregler et al., 1997] and Cosatto in the AT&T visual speech synthesiser [Cosatto and Graf, 1998].

To synthesise a new sentence a sequence of phoneme symbols and durations is required. Presently the system produces a synthetic visual sequence aligned to the audio of a previously unseen utterance. The speech recogniser is used to convert the audio signal to the constituent phoneme symbols and durations. For each phone to be synthesised, the inventory is searched for the closest context to the desired context (see next section). The corresponding portion of the original spline curves are extracted and temporally warped from their original length to the desired length. Note, at present only the portion of the curves corresponding to the centre phones are extracted, the phones either side are used only to determine the context.

The spline segments for each of the synthesis phones are concatenated to form a series of new trajectories, which are sampled at the desired frame rate to give the synthesis model parameters. Smoothing splines [de Boor, 2001] are fitted through the sampled model parameters to ensure a smooth transition from frame to frame and the smoothed parameters are applied to the model (Equation 1.2) to produce the synthetic image sequence of the talking face.

4.1 Synthesising Unseen Contexts

In order to account for unseen contexts a visual phoneme similarity matrix is used to find a context in the training data that is 'close' to the desired context, as described initially by Arslan [Arslan and Talkin, 1998]. The similarity matrix is automatically derived from the training data and therefore contains an objective measure of similarity between each phoneme and all others.

Arslan built the similarity matrix based on a pair-wise Euclidean distance between vectors representative of each phoneme in a compact shape space. Each occurrence of a phone is represented by an $N \times 5$ matrix, where the original principal component trajectories are sampled at five evenly spaced intervals over the duration of the phone and N is the number of model parameters. Each phoneme is then represented by first averaging the $N \times 5$ matrices, then averaging the columns of the resultant $N \times 5$ matrix, giving an $N \times 1$ vector. The similarity is

measured using:

$$\mathbf{S}_{ik} = e^{-\nu \mathbf{D}_{ik}} \tag{1.3}$$

where $\mathbf{D}_{ik}$ is the distance between phoneme i and k, and ν is a scalar allowing the dynamic range of the similarity to be controlled.

The distance metric used in this work differed to that described above. Each instance of a phone is represented by an $N \times 5$ matrix, as above, and an average $N \times 5$ matrix is computed for each phoneme. The rows of these averaged matrices are weighted according to the significance they have in the appearance model, determined by the percentage of variation captured by the respective mode of variation. The distance between two phonemes is then found by computing the sum of the squared differences between the individual elements of the two matrices and the similarity measured using Equation 1.3.

To find the closest match for an unseen context a measure of similarity is obtained between the desired context and each of the contexts appearing in the inventory using Equation 1.4:

$$\mathbf{s}_j = \sum_{i=1}^{C} \frac{\mathbf{S}_{l_i j}}{i+1} + \sum_{i=1}^{C} \frac{\mathbf{S}_{r_i j}}{i+1} \tag{1.4}$$

where $\mathbf{s}_j$ is the similarity between the desired context and the j^{th} context in the inventory, C is the context width, $\mathbf{S}_{l_i j}$ is the similarity between the i^{th} left phoneme in the j^{th} inventory context and the corresponding phoneme in the desired context, $\mathbf{S}_{r_i j}$ is the similarity between the i^{th} right phoneme of the j^{th} inventory context and the corresponding phoneme in the desired context. Only the triphones in the inventory with the same centre phone as the desired triphone are searched.

Since triphones are used in this system, C is 1, however this measure allows the context width of the synthesiser to be easily extended later.

5. Results

The result of applying the synthesis strategy outlined in the previous sections is shown in Figures 1.3 and 1.4. Each of the figures shows the original and synthesised trajectory of the first and second model parameters over the course of a sentence. The first parameter accounted for 27.74% and the second 15.95% of the total variation in the original training image. As an indication of performance the correlation coefficient is used, which is shown in Table 1.1.

Table 1.1. Correlation coefficients for the parameter trajectories shown in Figures 1.3 and 1.4

	Parameter 1	*Parameter 2*
Figure 1.3	0.9146	0.8260
Figure 1.4	0.9316	0.7714

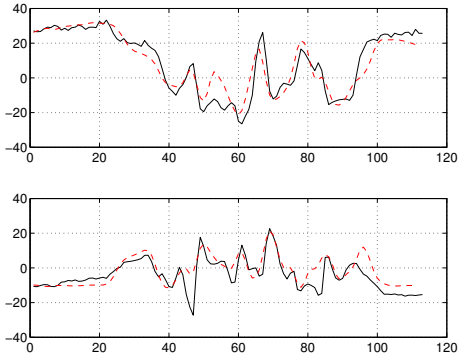

Figure 1.3. Trajectory of the first (top) and second principal component values. The solid black lines are the trajectory of the original parameters and the dashed red lines the trajectory of the synthesiser output. The original data was completely held out from the synthesiser.

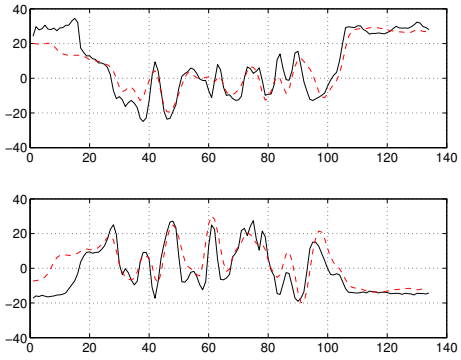

Figure 1.4. Trajectory of the first (top) and second principal component values. The solid black lines are the trajectory of the original parameters and the dashed red lines the trajectory of the synthesiser output. The original data was completely held out from the synthesiser.

6. Summary

In this paper we have described our first generation visual speech synthesiser that is able to produce near videorealistic synthetic visual

speech. The basis of the system is a statistical model of the face that produces near photorealistic facial images. The model is used to track the face of a talker in a video sequence and the trajectory of the tracked model parameters are interpolated using Hermite interpolation. A speech recogniser gives timing information for the triphones appearing in the training sentences and this information is used to construct a synthesis inventory. New sequences are created by locating the best match in the inventory for a desired synthesis context and the spline curves corresponding to the centre phone of the closest triphone are extracted and concatenated. These concatenated trajectories are re-sampled and the resultant parameters smoothed and applied to the model to produce the synthetic image sequence.

While the early results look promising (see Figures 1.3 and 1.4), the system has yet to undergo any formal evaluation. We will adopt the scheme Cohen and Massaro used to evaluate their visual speech synthesiser, Baldi [Massaro, 1998].

Further work will investigate how varying the context width affects the quality of the synthesised speech; whether extracting all the parameters for a triphone and blending overlapping regions gives smoother and more accurate synthetic visual speech; and investigating whether coarticulation models, such as that used by Baldi [Cohen and Massaro, 1994], perform as well as concatenative approaches using the appearance model.

References

[Arslan and Talkin, 1998] Arslan, L. and Talkin, D. (1998). Speech driven 3-d face point trajectory synthesis algorithm. In *Proceedings of the Internation Conference on Speech and Language Processing (ICSLP)*.

[Baker et al., 2001] Baker, S., Dellaert, F., and Matthews, I. (2001). Aligning images incrementally backwards. Technical Report CMU-RI-TR-01-03, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

[Bregler et al., 1997] Bregler, C., Covell, M., and Slaney, M. (1997). Video rewrite: driving visual speech with audio. In *Proceedings of SIGGRAPH*, pages 353–360.

[Brooke and Scott, 1998] Brooke, N. and Scott, S. (1998). Two- and three-dimensional audio-visual speech synthesis. In *Proceedings of Auditory-Visual Speech Processing*, pages 213–218.

[Cohen and Massaro, 1994] Cohen, M. and Massaro, D. (1994). Modeling coarticualtion in synthetic visual speech. In Thalmann, N. and D, T., editors, *Models and Techniques in Computer Animation*, pages 141–155. Springer-Verlag.

[Cootes et al., 1998] Cootes, T., Edwards, G., and Taylor, C. (1998). Active appearance models. In Burkhardt, H. and Neumann, B., editors, *Proceedings of the European Conference on Computer Vision*, volume 2, pages 484–498. Springer-Verlag.

[Cosatto and Graf, 1998] Cosatto, E. and Graf, H. (1998). Sample-based synthesis of photo-realistic talking heads. In *Proceedings of Computer Animation*, pages 103–110.

[de Boor, 2001] de Boor, C. (2001). Calculation of the smoothing spline with weighted roughness measure. *Mathematical Models and Methods in Applied Sciences*, 11(1):33–41.

[Ezzat and Poggio, 1997] Ezzat, T. and Poggio, T. (1997). Videorealistic talking faces: A morphing approach. In *Proceedings of the Audio-visual Speech Processing Workshop*, Rhodes, Greece.

[Guiard-Marigny et al., 1996] Guiard-Marigny, T., Tsingos, N., Adjoudani, A., Benoit, C., and Gascuel, M. (1996). 3d models of the lips for realistic speech animation. In *Computer Graphic 96*.

[Hallgren and Lyberg, 1998] Hallgren, A. and Lyberg, B. (1998). Visual speech synthesis with concatenative speech. In *Proceedings of Auditory-Visual Speech Processing*, pages 181–183.

[Le Goff and Benoit, 1996] Le Goff, B. and Benoit, C. (1996). A text-to-audiovisual-speech synthesizer for french. In *Proceedings of the International Conference on Speech and Language Processing (ICSLP)*, Philadelphia, USA.

[Lee et al., 1993] Lee, Y., Terzopoulos, D., and Waters, K. (1993). Constructing physics-based facial models of individuals. In *Proceedings of Graphics Interface*, pages 1–8.

[Massaro, 1998] Massaro, D. (1998). *Perceiving Talking Faces*. The MIT Press.

[Parke, 1974] Parke, F. (1974). *A Parametric Model for Human Faces*. PhD thesis, University of Utah, Saltlake City, Utah.

[Parke and Waters, 1996] Parke, F. and Waters, K. (1996). *Comptuer Facial Animation*. A K Peters.

[Platt and Badler, 1981] Platt, S. and Badler, N. (1981). Animating facial expression. *Computer Graphics*, 15(3):245–252.

[Waters, 1987] Waters, K. (1987). A muscle model for animating three-dimensional facial expressions. *Proceeding of ACM SIGGRAPH*, 21(4):17–24.