TOWARDS VIDEO REALISTIC SYNTHETIC VISUAL SPEECH

Barry J. Theobald,[*] *J. Andrew Bangham,*[*] *Iain A. Matthews*[†] *and Gavin C. Cawley*[*]

[*]School of Information Systems, University of East Anglia, Norwich, NR4 7TJ, UK
[†]Robotics Institute, Carnegie Mellon, Pittsburgh, PA 15123, USA
b.theobald@uea.ac.uk, {ab, gcc}@sys.uea.ac.uk, iainm@cs.cmu.edu

ABSTRACT

In this paper we present initial work towards a video-realistic visual speech synthesiser based on statistical models of shape and appearance. A synthesised image sequence corresponding to an utterance is formed by concatenation of synthesis units (in this case phonemes) from a pre-recorded corpus of training data. A smoothing spline is applied to the concatenated parameters to ensure smooth transitions between frames and the resultant parameters applied to the model — early results look promising.

1. INTRODUCTION

The field of facial animation and visual speech synthesis has received increasing interest in recent years, see [1] for an overview. Applications include desktop agents for personal computers, film/media production, low bandwidth video conferencing and surgical planning amongst others. The work described in this paper is motivated by the need to develop a low bandwidth virtual human (avatar) capable of delivering sign language at a quality comparable to high bandwidth video. Perceptual tests reported elsewhere [2] show a synthetic talking face based on the models presented here are considered comparable to original video even at an extremely low bandwidth (typically around 3.6 kbits/s).

The main approaches to modelling and animating the face are: *graphics* based systems [1, 3, 4], *image* based systems [5–9] and *hybrid* systems [10]. Image based systems extract a model and parameters to animate the model from video or image sequences of real faces using computer vision/image processing techniques. Such systems can offer a high degree of realism since real facial images form the basis of the model, although producing video realistic synthetic visual speech is still a difficult problem.

In previous work, Ezzat et al. [8] represented each discernible mouth shape with a static image and used an optical flow algorithm to morph between the images. A problem with this approach is it is difficult to modify lip shapes depending on context, where at any instance in time the lip shape is influenced by previous and future lip shapes. Cosatto et al. [7] segment the face and head into several parts using various computer vision techniques. Measurements on facial features allow a library of representative samples to be indexed. New sequences are created by stitching together the appropriate mouth shapes based on triphones to be synthesised and the corresponding features for a particular emotion. Bregler et al. [5] describe a system that is capable of taking existing video of a talker and modifying it to create new unspoken utterances. An audio-visual corpus of triphones is built from existing video of a talker by automatically labelling the lip and jaw positions in training footage. New video sequences are created by concatenating the appropriate triphone sequences from the training corpus.

2. THE FACE MODEL

In this work, parameters to animate a model are extracted from a video sequence of a face exhibiting the variations we wish to capture. Specifically facial gestures are represented using the principal component scores drawn from a combined model of shape and appearance [11]. Statistical models of the appearance of the face (or the mouth region) have been used as the basis for a visual speech synthesiser [6]. Here we describe a system based on the statistics of the combined shape and appearance of the face.

Following the notation of Cootes [11], a statistical model of shape, the *point distribution model* (PDM), is trained by hand labelling landmark points in a set of images and performing a principal component analysis (PCA) on the coordinates. Any shape can then be approximated by deforming the mean shape 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.

An appearance model is computed by shape normalising the training images, see Figure 1, so the landmarks in each image lie in the position of the landmarks of the mean shape. A PCA is computed on the resultant images such that any can be approximated using $\mathbf{a} \approx \overline{\mathbf{a}} + \mathbf{P}_a\mathbf{b}_a$. Where $\overline{\mathbf{a}}$ is the mean shape-normalised image, $\mathbf{P}_a$ is the matrix of the first t_a eigenvectors of the covariance matrix and $\mathbf{b}_a$ a vector of appearance parameters.

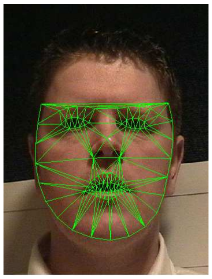

Fig. 1. The mesh used to shape normalise the training images. The vertices of the mesh are landmark points forming the shape model and the image is warped by perturbing the landmarks from their position in the image to the position of the mean shape.

Each image is described by a set of shape parameters and a set of appearance parameters, $\mathbf{b}_s$ and $\mathbf{b}_a$ respectively. A combined model of shape and appearance is computed by concatenating the PCA scores for the shape and appearance models for each image and performing a third PCA. The combined shape and appearance model is given by Equation 1,

$$\mathbf{b} \approx \mathbf{Q}\mathbf{c}, \tag{1}$$

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 2, see Figure 2.

$$\mathbf{x} \approx \overline{\mathbf{x}} + \mathbf{P}_s\mathbf{W}_s\mathbf{Q}_s\mathbf{c}, \quad \mathbf{a} \approx \overline{\mathbf{a}} + \mathbf{P}_a\mathbf{Q}_a\mathbf{c}, \tag{2}$$

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 [11] and facial animations are generated by controlling the time trajectory of the vector $\mathbf{c}$.

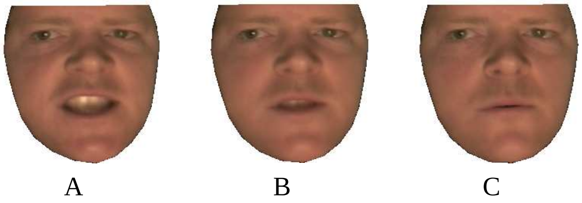

Fig. 2. Facial images synthesised using a combined model of shape and appearance. The first element of the vector $\mathbf{c}$ was perturbed by A) -2, B) 0 and C) +2 standard deviations from the mean.

3. DATA CAPTURE

When compiling the training data for the model there are a number of considerations to be taken into account. In particular, PCA is not invariant to any type of variation. As much of the original variation as possible is captured in the first few modes. In creating a visual speech synthesiser based on PCA, the only variations we wish the model to capture are those variations of the face due to speech. Lighting changes, identity changes of the face and pose variations of the head are all considered noise. We can ensure these variations are not captured by training the model on a single face. Secondly, the data should be captured in a single sitting under controlled lighting conditions and pose variations can be removed by constraining the head to lie in the same place in all images.

The training data for the face model presented here was collected in one sitting on a Panasonic DV99B digital camcorder and digitized at a frame rate of 25 frames per second with a frame size of 720x576 (color). The audio was captured at 44.1 kHz stereo and was later used to segment the video using a hidden Markov model (HMM) speech recogniser run in forced-alignment mode. The training data consisted of a single talker uttering 100 sentences chosen to be phonetically rich. The sentences themselves were drawn from the British Telecom Messiah corpus. The speaker was not physically constrained during recording but the pose of the head was roughly maintained throughout. The facial expression was held as neutral as possible so the main sources of variation were due only to speech. The total database consisted of 9431 images.

4. SYNTHESIS

To obtain model parameters for synthesising visual speech a statistical model trained on 50 images tracked the face in all 9431 images using the Active Appearance Model Algorithm (AAM) [11]. Hermite interpolation [12] was used to give a continuous representation of the time trajectory of each parameter in each sentence. A look-up table was then created using the speech recogniser output. For each phone the start time in the sentence, the end time and an index indicating in which sentence the phone occurred are stored along with the phone symbols appearing immediately either side.

To synthesise a new utterance a text stream is converted to a set of phones and durations. For each phone to be synthesised, the original training data is searched for the phone appearing in the same context and the corresponding portion of the original spline curves extracted and temporally warped to the desired duration. In the event that a phone has occurred in the correct context more than once in training, the phone with the closest duration to the intended duration is selected.

The spline segments for each of the synthesis phones are concatenated to form a series of new trajectories, which are sampled at regular intervals to give a series of model parameters (one set per frame). Smoothing splines [13] 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 2) to produce the synthetic image sequence of the talking face.

4.1. Results

As a first approximation to synthesis, only utterances appearing in contexts that have been previously seen are synthesised. The result of this is shown in Figure 3. The trajectory of the first model parameter from the original data (the face tracker output) is plotted with the trajectory of the first model parameter output from the synthesiser. If the context of a synthesis phone has occurred anywhere other than the original sentence, that data is used by the synthesiser. In the event that the context has only occurred in the original sentence (currently being synthesised) the synthesiser will fall back on that data. The aim of this experiment was to determine whether the concatenation boundaries were visible in the synthetic visual speech sequence.

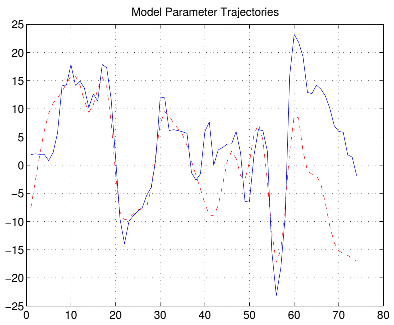

Fig. 3. The trajectory of the first parameter from the original sequence (solid line) and the synthesiser output (dashed line).

This approach to visual speech synthesis is rather simplistic, although effective. It is hoped that the effects of coarticulation are captured by concatenating segments of the original data. This approach relies on the context of the phone having appeared elsewhere in training however. In English there are approximately 45 phonemes, making possible a total of 45^3 different contexts (if the context width is taken as one phone either side of the center phone), although in practise not all of these will appear in spoken language. To ensure any of these contexts could be recreated using the method described would require a large amount of training data. Also, previous studies have shown that a context of only a single phone either side of the current phone may not be sufficient as lip shapes may be influenced by those up to six phones away [14].

In Figure 3, the largest error appears at frame 42 (other than the silence region at the end of the sentence where the trajectory of the model parameter appears to be of the correct shape, but offset from the desired level). The two faces in Figure 4 are synthesised using the model parameters from the original sentence and the synthesiser output. It appears that although the trajectories of the synthesised and original model parameters are different at frame 42, the correct lip shape has been created, although it appears somewhat over articulated.

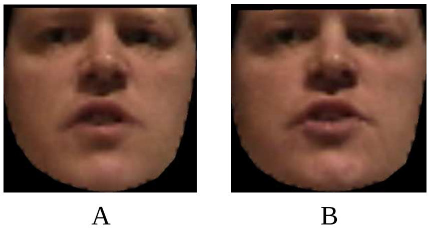

Fig. 4. Synthetic faces created using A) the tracker output and B) the synthesiser output for frame 42, which seems a major source of error between the trajectories in Figure 3.

4.2. 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 et al. [15]. The similarity matrix is automatically derived from the training data and therefore contains an objective measure of similarity between each phone and all others. The matrix in [15] was built based on a pairwise Euclidean distance between the principal components of each phone in a compact shape space. Here we tested the method against both the shape component of the model and the combined shape and appearance component of the model. The resulting synthesis trajectories across the two tests were identical, for all synthesis contexts the same training context was extracted as the closest match in both experiments. The result of the test in the combined space is shown in Figure 5.

Again we have differences between the two trajectories. Significant errors appear at frames 14, 30, 34, 42, 46 and 56. As in Figure 4, when viewing the synthetic visual speech sequence, most of the errors appear to be an under or over articulation of a particular mouth shape. In frame 34 however, the wrong lip shape has been synthesised, as can be seen in Figure 6. This suggests more work is required in determining which context to extract from the database of training phones, the similarity matrix alone may not be reliable enough.

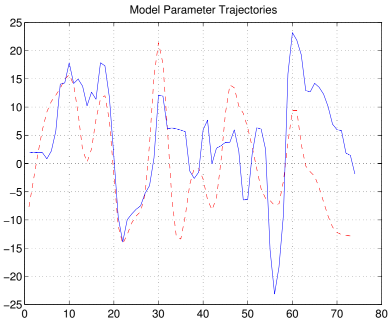

Fig. 5. The trajectory of the first model parameter for the original data (solid line) and the synthesised data (dashed line). The original data was completely held out and the visual phoneme similarity matrix used to find context closest to unseen contexts.

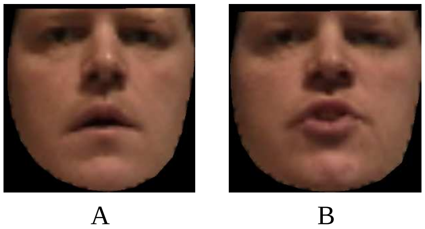

Fig. 6. The synthetic face using A) the tracker output and B) the synthesiser output for frame 34, which seems one of the major sources of error between the trajectories in Figure 5.

5. CONCLUSIONS

In this paper we have presented initial work towards video realistic synthetic visual speech. Early results look promising, we have a face model trained on real facial images and driven by real speech data. Synthesis is based on the concatenation of model parameters corresponding to phone segments. In order to account for phones in unseen contexts, we have implemented an algorithm by Arlsan et al. [15], which determines the most similar context in the training data to the unseen context. We have shown that it appears that this strategy has problems on a small (100 sentence) training set and that wrong lip shapes can be generated.

We are presently gathering a larger training database to give a better coverage of phoneme contexts. We are also looking at other ways to select an unseen context. For example, we saw that often correct lip shapes are extracted from the training data, but are over or under articulated. To overcome this we are adding stress levels to each phone in the training database such that the stress level can be used to select a particular phone if it has appeared in a particular context more than once. Presently the phone with the closest duration is selected.

Further work will also involve evaluating the quality of the synthesiser. In [2] we outlined a perceptual test to determine a trade off between synthesis quality and bandwidth. We are currently refining the test to include a test of perceived synthesis quality and lip readability.

Demos of our system can be found on the web at http://www.facial-animation.co.uk

6. REFERENCES

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

[2] B.J. Theobald, G.C. Cawley, S.M. Kruse, and J.A. Bangham, "Towards a low bandwidth talking face using appearance models," in *Proceeding of the British Machine Vision Conference*, 2001, pp. 583–592.

[3] M. Cohen and D Massaro, "Modeling coarticualtion in synthetic visual speech," in *Models and Techniques in Computer Animation*, N.M. Thalmann and Thalmann D, Eds., pp. 141–155. Springer-Verlag, 1994.

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

[5] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: driving visual speech with audio," in *Proceedings of SIGGRAPH*, 1997, pp. 353–360.

[6] N.M. Brooke and S.D. Scott, "Two- and three-dimensional audio-visual speech synthesis," in *Proceedings of Auditory-Visual Speech Processing*, 1998, pp. 213–218.

[7] E. Cosatto and H.P. Graf, "Sample-based synthesis of photo-realistic talking heads," in *Proceedings of Computer Animation*, 1998, pp. 103–110.

[8] T. Ezzat and T. Poggio, "Miketalk: A talking facial display based on morphing visemes," in *Proceedings of the Computer Animation Conference*, 1998.

[9] Z. Liu, Z. Zhang, C. Jacobs, and M. Cohen, "Rapid modeling of animated faces from video," Tech. Rep. MSR-TR-2000-11, Microsoft Corporation, 2000.

[10] F. Pighin, J. Hecker, D. Lischinski, R Szeliski, and D. Salesin, "Synthesizing realistic facial expressions from photographs," in *Proceedings of SIGGRAPH*, 1998.

[11] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active appearance models," in *Proceedings of the European Conference on Computer Vision*, H. Burkhardt and B. Neumann, Eds. 1998, vol. 2, pp. 484–498, Springer-Verlag.

[12] R.H. Bartels, J.C. Beatty, and B.A. Barsky, *An Introduction to Splines for Use in Computer Graphics and Geometric Modeling*, Morgan Kaufmann, 1987.

[13] C. de Boor, “Calculation of the smoothing spline with weighted roughness measure,” *Mathematical Models and Methods in Applied Sciences*, vol. 11, no. 1, pp. 33–41, 2001.

[14] R.D. Kent and F.D. Minifie, “Coarticulation in recent speech production,” *Journal of Phonetics*, vol. 5, pp. 115–133, 1977.

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