NEAR-VIDEOREALISTIC SYNTHETIC VISUAL SPEECH USING NON-RIGID APPEARANCE MODELS

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

*School of Information Systems, University of East Anglia, Norwich, NR4 7TJ, UK
†Robotics Institute, Carnegie Mellon, Pittsburgh, PA 15213, USA
b.theobald@uea.ac.uk, {gcc, ab}@sys.uea.ac.uk, iainm@cs.cmu.edu

ABSTRACT

In this paper we present work towards videorealistic synthetic visual speech using non-rigid appearance models. These models are used to track a talking face enunciating a set of training sentences. The resultant parameter trajectories are used in a concatenative synthesis scheme, where samples of original data are extracted from a corpus and concatenated to form new unseen sequences. Here we explore the effect on the synthesiser output of blending several synthesis units considered similar to the desired unit. We present preliminary subjective and objective results used to judge the realism of the system.

1. INTRODUCTION

Potential applications for a realistic visual speech synthesiser include desktop agents, character animation in computer games, translation agents, low bandwidth videoconferencing, and film and media post-production. In many such applications, videorealism may be a requirement. Traditionally to achieve videorealism image-based techniques are employed, for example [1, 2, 3]. Model-based schemes are usually based on a computer graphics representation of the face [4] and are employed where videorealism is not a requirement. The trade-off is the bandwidth required to drive the animations, image-based systems require the greater bandwidth. Recently computer vision techniques have been used to extract a model of the face from video sequences and the model applied to create near-videorealistic synthetic visual speech, for example [5, 6]. Such systems are attractive since they allow the realism of image-based systems to be achieved, while still maintaining the flexibility of model-based systems. They also allow videorealistic synthesis at low bandwidths [7].

Here we describe both a model that creates near-photorealistic facial images, and how this model is used in a sample-based method for synthesising visual speech.

2. DATA CAPTURE

Training data was collected using an ELMO EM-02PAL camera and digitised at a frame rate of 25 frames per second using an IEEE 1394 compliant capture card with a frame size of 720x576. The audio was captured using the on-camera microphone and digitised at 11025 Hz, 16 bits/sample stereo. This was later used to phonetically segment the video using a hidden Markov model (HMM) speech recogniser run in forced-alignment mode.

To minimise unwanted sources of variation when creating the model, the video was recorded using a head mounted camera, recording a single talker in one sitting, thus eliminating variations due to pose, identity and lighting. The speaker held the facial expression as neutral as possible (no emotion) to confine the variation of the facial features to that due to speech. The training data consisted of 279 sentences containing multiple occurrences of 6315 triphones, resulting in approximately 34,000 frontal images of the face.

3. THE FACE MODEL

The face model that forms the basis of the synthesiser is based on shape and appearance models due to Cootes and co-workers [8]. Facial gestures are represented as principal component scores drawn from statistical models of the shape and appearance variation of the face.

Following the notation of Cootes [8], a model of shape, termed 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 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, and $\mathbf{b}_s$ is a vector of shape parameters.

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

A combined shape and appearance model is built by computing and concatenating the shape and appearance parameters for each labelled image and performing a third PCA. The combined shape and appearance model is given by $\mathbf{b} \approx \mathbf{Q}\mathbf{c}$, where $\mathbf{Q}$ is the matrix of eigenvectors of the covariance matrix and $\mathbf{c}$ a vector of parameters that reflect changes in the shape and appearance of the face. Given the combined model, realistic images can be synthesised given a set parameters using

$$\mathbf{x} \approx \overline{\mathbf{x}} + \mathbf{P}_s\mathbf{W}_s\mathbf{Q}_s\mathbf{c}, \quad \mathbf{a} \approx \overline{\mathbf{a}} + \mathbf{P}_a\mathbf{Q}_a\mathbf{c}, \tag{1}$$

where the matrix $\mathbf{W}_s$ takes into account the scaling mismatch between the parameters $\mathbf{b}_s$ (which models Euclidean distance) and $\mathbf{b}_a$ (which models pixel RGB intensity). This is computed as shown in [8].

Given a set of landmarks, $\mathbf{x}$, and a shape-free image, $\mathbf{a}$, the final synthesised image is created by warping $\mathbf{a}$ to $\mathbf{x}$.

4. DATA PREPARATION

The face in the video is first encoded in terms of the model parameters. This requires locating the landmarks in all images, computing the shape parameters, warping each image to the mean shape and computing the appearance parameters, then finally projecting into the combined shape and appearance space (if a combined model is used).

Given that the database contains over 34,000 images, labelling all by hand would be a time consuming and tedious task. Instead, a shape model and an appearance model are built from 100 images labelled by hand and the *gradient descent active appearance* search algorithm [9] used to automatically label the images.

For each frame in the video the face is mapped to the corresponding point in the model space. Over the course of a sentence the discrete points corresponding to the parameters for each frame map a trajectory through the face-space. A continuous parametric representation of this trajectory is obtained using Hermite interpolation, and the 279 trajectories, one for each training sentence, are stored in the synthesis codebook.

The audio component of the training video is passed through a speech recogniser, the output from which is a list of the constituent phoneme symbols that form each sentence and their corresponding start and stop times. This phonetic information is also stored in the synthesis codebook and is later used to index the trajectories such that segments can be extracted corresponding to any particular phoneme.

4.1. Measuring Phoneme Similarity

The model is to be used in a sample-based synthesis scheme, so the synthesiser must be able to account for phonemes appearing in unseen contexts. To allow for this, a similarity matrix is used to find contexts in the training data that are 'closest' to the desired context. The scheme described here is similar to that described by Arslan and Talkin [10].

The similarity matrix is automatically derived from the training data and each element contains an objective measure of similarity between two given phonemes. To measure the similarities, first all observations of each phoneme are gathered and the portions of the original trajectories sampled at five evenly spaced intervals. Next the mean representation of each phoneme is computed. For a model with N parameters, each phoneme is represented by an $N \times 5$ matrix.

The mean representation may not however, be a reliable representation of a phoneme, particularly if it is significantly modified by context. To allow for this, the dispersion of each phoneme in the face-space is computed by calculating the total area between the normalised parameter trajectories and the corresponding mean trajectory. The distance between two phonemes is given by the sum of the squared differences between the matrix elements, where the matrices are weighted by their relative stability. Since the model is based on PCA, errors in the lower dimensions of the face-space are more significant than errors in the higher dimensions. The rows of the matrices are also weighted by the significance of the corresponding parameter in the model. The distance between any phoneme pair is given by

$$D_{ij} = \sum_k \sum_l \left[\left(v_i P^i_{kl} - v_j P^j_{kl} \right) w_k \right]^2, \tag{2}$$

where P^i is the matrix representing the i^{th} phoneme and P^j the matrix representing the j^{th} phoneme. The value w_k is the significance of the k^{th} parameter in the model, v_i is the variability of the i^{th} phoneme and v_j the variability of the j^{th} phoneme.

Given the matrix of distance values, the similarities are computed using

$$S_{ij} = e^{-\nu D_{ij}}. \tag{3}$$

The range of similarity is 0 (maximally dissimilar), to 1 (identical). The variable ν controls the spread of similarity values over the range (0,1). The similarity matrix is stored with the parameter trajectories and phoneme timing information in the synthesis codebook.

5. SYNTHESIS

The synthesiser can be driven in one of three ways; by the shape model alone, the shape and appearance model, or the combined model. These systems differ only in what the parameters model.

For a synthesiser driven by the shape model alone, a linear mapping is assumed to exist between the shape parameters and the appearance parameters, such that the appearance parameters can be predicted from the shape parameters output from the synthesiser using

$$\mathbf{b}_a = \mathbf{A}\mathbf{b}_s. \tag{4}$$

Where $\mathbf{A}$ is computed given the original parameter trajectories

$$\mathbf{A} = \mathbf{b}_a \mathbf{b}'_s \left(\mathbf{b}_s \mathbf{b}'_s \right)^{-1}. \tag{5}$$

For a synthesiser driven by the shape and appearance models, each sentence is represented by two trajectories. The $N \times 5$ phoneme matrices in the similarity calculations become $2N \times 5$ matrices, where N shape and N appearance parameters are concatenated to form the phoneme observations.

The combined model case is identical to the shape only case, each sentence is represented as a single trajectory and phoneme observations are represented by an $N \times 5$ matrix. In this case however, the appearance is implicitly modelled by the parameters.

5.1. Synthesising a New Utterance

The visual sequence corresponding to a new utterance is synthesised by first converting a text stream to a list of phonemes and durations. For each phoneme to be synthesised, the original training data is searched for the n examples in the most similar contexts in the codebook using

$$\mathbf{s}_j = \sum_{i=1}^{C} \frac{\mathbf{S}_{l_{ij}}}{i+1} + \sum_{i=1}^{C} \frac{\mathbf{S}_{r_{ij}}}{i+1}, \tag{6}$$

where $\mathbf{s}_j$ is the similarity between the desired context and the j^{th} context in the inventory, C is the context width, $\mathbf{S}_{l_{ij}}$ is the similarity between the i^{th} left phoneme in the j^{th} inventory context and the corresponding phoneme in the desired context, $\mathbf{S}_{r_{ij}}$ is the similarity between the i^{th} right phoneme of the j^{th} inventory context and the corresponding phoneme in the desired context.

This similarity score is attractive since it allows the context width to be easily varied without modifying the synthesiser. In the results presented here a context width of $C = 1$ is used. Hence, the synthesis unit is the triphone.

5.1.1. Creating New Parameter Trajectories

Given the n closest matches in the codebook for each synthesis phoneme, the corresponding portions of the original parameter trajectories are extracted and temporally warped to the desired duration. These normalised trajectories are blended to form a new trajectory in the face-space, where the blending is weighted such that the most similar contexts receive more weight and the sum of the weights is unity.

These new phoneme trajectories in the face-space are concatenated and sampled at the original frame rate. Smoothing splines [11] are fitted through the model parameters to ensure a smooth transition between synthesis units and the smoothed parameters are applied to the model to produce the synthetic image sequence of the talking face.

Figure 1 shows an example of a real and synthetic shape parameter trajectory, while Figure 2 shows example faces output from the tracker and the corresponding faces output from the synthesiser.

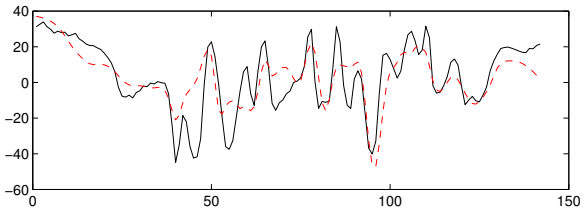

Fig. 1. A shape parameter trajectory from the tracker (black solid line) and synthesiser output (red dashed line). In this example, the synthesiser blended the $n = 3$ closest examples in the codebook.

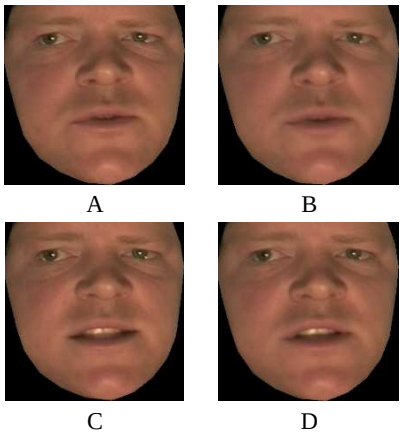

Fig. 2. Example faces output by the tracker (A and C) and the corresponding examples output by the synthesiser (B and D).

6. RESULTS

We present here subjective and objective results of tests performed on the synthesiser.

6.1. Subjective Results

A Turing test was used in order to select the model from the three outlined in Section 3 that produces the most natural looking animations. In the test, original sequences and the corresponding synthesiser output were randonly intermixed and played to viewers (without audio). Viewers were then asked to watch the sequences and identify which were real and which were synthetic.

Ezzat performed similar tests in [5] and noted post-processing of the synthetic output was required. The process of averaging image frames in the synthesis removes camera noise from the synthetic sequences and gives the face a zombie-like effect due to the removal of subtle eye and eye brow movements. To overcome this, an estimate of the camera noise is made and added to the synthetic images and a face mask defined such that mouth regions from synthesised images can be re-composited into a background sequence from the original video. This ensures a fair comparison between real and synthetic sequences.

Here, since we are testing the *dynamics* of the various models, camera noise is removed from the original sequences by playing the tracker output (model encoded video) rather than the video itself. Pixels from the eyes and above are also removed from the sequences, viewer attention is then focused on the movements of the face around the mouth region.

Twenty sequences (ten original and ten synthetic) were played to 12 viewers, with the results shown in Table 1.

Synthesis Type	% Correct	χ^2	p $\leq$
Shape driven	71.67%	45.067	0.001
Combined driven	60.42%	10.4232	0.01
Separate driven	52.92%	0.8236	1

Table 1. Percent correct identification of real and synthetic sequences.

It is clear that the shape driven model performed the worst of the three as viewers were able to correctly identify the real sequences from the synthetic. The hypothesis that the model and video are indistinguishable has a probability of p≤ 0.001, and so is rejected. This is perhaps due to the assumption of a linear relationship between the shape and appearance parameters. It was shown in [12] that a non-linear mapping results in a better approximation of the original appearance parameter trajectories.

The combined model performed better than the shape only model, however since the hypothesis that the model and video are indistinguishable has a probability of p≤ 0.01, this is also rejected. The best performance was obtained by driving the synthesiser with separate shape and appearance information. For this synthesis scheme, the error between real/synthetic judgements approached the chance level (50% correct) and the hypothesis that the model and video are indistinguishable has a probability of p≤ 1, and is therefore accepted.

6.2. Objective Testing

One possible objective measure of performance is the correlation between real and synthetic parameter trajectories. Figure 3 shows the mean correlation coefficients for the first five parameters of the shape and appearance models in the synthesis of 279 sentences. Also shown is the effect of varying the number of examples extracted and blended from the codebook in forming the synthetic trajectories, described in Section 5.1.1.

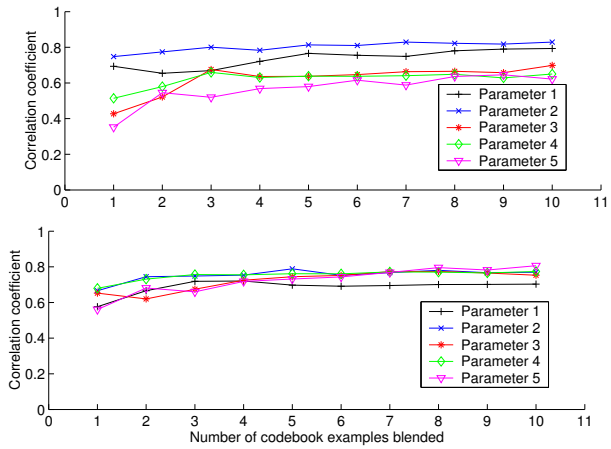

Fig. 3. The mean correlation coefficient between original and synthesised parameter trajectories for the shape (upper) and appearance (lower) models, averaged over the synthesis of 279 sentences.

Objective measures of the synthesiser output, such as the correlation between parameter trajectories, are difficult to quantify. It is unclear at what point the difference between real and synthetic trajectories becomes significant. Some parameters will be better correlated than others, however, the model parameters are not independent and good quality synthesis requires all parameters to have a high correlation.

A perhaps more meaningful measure of performance is the RMS error between points on the face in real and synthetic sequences, and the RMS error between pixel values in the real and synthetic images. Figure 4 shows these errors for the same sequences used to create Figure 3.

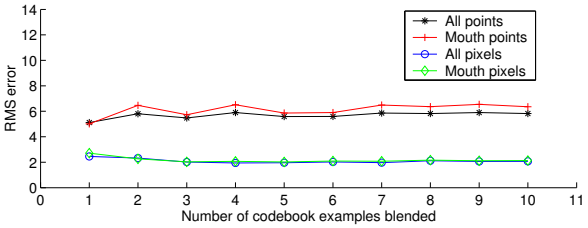

Fig. 4. Mean RMS error between synthesised and original landmarks for the whole face (*) and mouth only (+), and the pixels contained in the whole face (o) and the mouth only (◇).

7. CONCLUSIONS

In this paper we have described three models that produce near-photorealistic facial images and explained how these models can applied to create near-videorealistic synthetic visual speech. Section 6.1 gave results of a Turing test used to select the model that produces the most natural facial movements. We are presently conducting formal subjective testing to determine the intelligibility of these models. In lieu of these results, Section 6.2 presented objective measures of performance. Caution should be taken when using these as a direct measure of synthesis performance since the same sentence spoken more than once by the same talker will itself never be identical. Demos of the system can be found at http://www.facial-animation.co.uk.

8. ACKNOWLEDGEMENTS

The authors would like to thank everyone who took part in the subjective tests and Dr. Rob Foxall for his input.

9. REFERENCES

[1] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *Computer Graphics Annual Conference Series (SIGGRAPH)*, Los Angeles, California, August 1997, pp. 353–360.

[2] E. Cosatto and H.P. Graf, "Photo-realistic talking-heads from image samples," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 152–163, September 2000.

[3] T. Ezzat and T. Poggio, "Visual speech synthesis by morphing visemes," Tech. Rep. 1658, Massachusetts Institute of Technology, 1999.

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

[5] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," in *Proceedings of SIGGRAPH*, San Antonio, Texas, July 2002, pp. 388–398.

[6] B.J. Theobald, J.A. Bangham, I.A. Matthews, and G.C. Cawley, "Towards video realistic synthetic visual speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, USA, 2002, pp. 3892–3895.

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

[8] 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., Freiburg, Germany, 1998, vol. 2, pp. 484–498, Springer-Verlag.

[9] S. Baker and I. Matthews, "Equivalence and efficiency of image alignment algorithms," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, Kauai, Hawaii, 2001, pp. 1090–1097.

[10] L.M. Arslan and D. Talkin, "3D face point trajectory synthesis using an automatically derived visual phoneme similarity matrix," in *Proceedings of Auditory-Visual Speech Processing*, Terrigal, Australia, December 1998, pp. 175–180.

[11] 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.

[12] Y. Du and X. Lin, "Realistic mouth synthesis based on shape appearance dependence mapping," *Pattern Recognition Letters*, vol. 23, no. 14, pp. 1875–1885, December 2002.