AVSP 2003 - International Conference
on Audio-Visual Speech Processing
St. Jorioz, France
September 4-7, 2003

Evaluation of a Talking Head based on Appearance Models

Barry-John Theobald[*], *J. Andrew Bangham*[*], *Iain Matthews*[†], *Gavin Cawley*[*]

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

Abstract

In this paper we describe how 2D appearance models can be applied to the problem of creating a near-videorealistic talking head. A speech corpus of a talker uttering a set of phonetically balanced training sentences is analysed using a generative model of the human face. Segments of original parameter trajectories corresponding to the synthesis unit are extracted from a codebook, normalised, blended, concatenated and smoothed before being applied to the model to give natural, realistic animations of novel utterances. We also present some early results of subjective tests conducted to determine the realism of the synthesiser.

1. Background

It is well known that speech is a multi-modal form of communication; seeing the face of the talker provides additional information over and above the auditory signal that significantly influences the perception and understanding of speech [1, 2]. The face is a complex communication device that provides both linguistic and non-linguist cues and we quickly become expert at detecting and recognising subtle changes in the features of the face, making realistic animation of the human face a very difficult problem. Potential applications for a suitably realistic facial animation system include desktop agents, character animation in films or computer games, translation agents, low bandwidth video conferencing and the personalisation of web-based instant messenger clients to name but a few.

Facial animation systems can be broadly classified as either *graphics-based* or *image-based*. Graphics-based systems represent points on the face as vertices in three dimensions and approximate the surface of the face by connecting the vertices. A set of parameters deform the mesh in some controlled manner, where the parameterisation is either direct, as in *terminal analog* synthesis [3, 4, 1], or indirect, as in *physically-based* synthesis [5, 6, 7]. Graphics-based systems can be efficiently rendered, especially on modern graphics processors, however they tend to lack videorealism. Texture mapping an image of a real face onto the mesh generally is still not enough to convince a viewer that the animated sequence is a real face.

Image-based systems use computer vision or image processing algorithms to build facial models and drive animations from images of real faces, for example [8, 9, 10, 11, 12, 13]. Providing the correct lip shape is presented for a given sound and the synthesised movements of the face look natural, image-based synthesis is able to achieve a high degree of videorealism. Bregler and co-workers [8] automatically segment existing footage of a talker into short sequences corresponding to *triphones* and replay these segments in a new order to create novel sequences. The quality of the resultant animations is generally determined by the size of the training corpus. A very large database is required for a reasonable coverage of all possible triphones, and where a triphone is not contained in the database the closest example is used. Brand [9] and Brooke and Scott [10] use hidden Markov models (HMMs) to learn the characteristics of facial deformations associated with speech production. The trained HMMs are used to generate new sequences, where Brand animates both the speech and expression of a (possibly) novel person, while Brooke and Scott generate image sequences of the mouth region of a single talker. Cosatto and Graf [11] populate a hyper-space of facial examples, where the dimensions of the hyper-space correspond to measurements on a talker's face. Example images are extracted from this compact hyper-space based on the phonetic string to be synthesised and these mouth shape images stitched together with images of other facial regions (eyes, cheeks etc.) to create novel sequences of expressive speech. Ezzat and Poggio report a model-based synthesis technique that creates very realistic speech animation of the mouth region of a talker [13]. Here, we describe our alternative technique for achieving very realistic speech animation of the whole face [14, 15], and describe some early subjective tests used to evaluate the naturalness of the synthesiser.

2. Data Capture

To ensure the pose of the head remained constant, the training data was collected using a head mounted camera and transferred from DV tape to computer using an IEEE 1394 compliant capture card with a frame size of 360x288 pixels (one quarter DV-PAL). The audio was captured using the on-camera microphone and digitised at 11025 Hz, 16 bits/sample stereo and was later used to phonetically segment the video using a HMM-based speech recogniser run in forced-alignment mode. Only a single talker was recorded in a single sitting to remove identity variation and to ensure the lighting was even and constant throughout the entire training video. 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, comprising approximately twelve minutes of speech data.

3. Modelling the Face

Following the notation of Cootes and co-workers [16], a statistical model of the shape of an object, termed the *point distribution model* (PDM), is trained by manually placing landmarks on a set of images and performing a principal component analysis (PCA) on the coordinates of these landmarks. Typically about 100 points are used for the whole face and 30 images are selected for hand labelling covering a broad range of the mouth shapes associated with speech production.

Any training shape can then be approximated using $\mathbf{x} \approx$

$\bar{\mathbf{x}} + \mathbf{P}_s\mathbf{b}_s$, where $\bar{\mathbf{x}}$ is the mean shape, $\mathbf{P}_s$ is the matrix of the eigenvectors of the covariance matrix associated with the t_s eigenvalues of the greatest magnitude, chosen to describe some preset percentage of the total variation (typically 95%), and $\mathbf{b}_s$ is a vector of t_s shape parameters.

A statistical model of the appearance of the face is computed by warping the labelled images from the landmarks to the mean shape. This normalises the shape of the face in each image, ensures each example has the same number of pixels and ensures a pixel in one example corresponds to the same feature of the face in all other examples, where typically about 40,000 (RGB) pixels are used. A further PCA is performed on the pixel values within the shape-normalised faces, such that any RGB appearance can be approximated using $\mathbf{a} \approx \bar{\mathbf{a}} + \mathbf{P}_a\mathbf{b}_a$, where $\bar{\mathbf{a}}$ 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.

Each image is, therefore, described by a set of shape parameters and a set of appearance parameters, $\mathbf{b}_s$ and $\mathbf{b}_a$ respectively. The shape and appearance spaces are concatenated such that the face in an image maps to a single point in a *face-space*, where some of the dimensions of this face-space correspond to shape variation and some to appearance variation. We do not project the shape and appearance parameters into a combined model space for synthesis as subjective testing of various forms of appearance models have shown that the most *dynamically* realistic models are comprised of independent shape and appearance models [15].

4. Data Preparation

Given the shape and appearance models, the face in all 34000 video frames must be encoded in terms of the parameters b_s and b_a. To project the face onto the principal components requires the landmark positions for each image, which are obtained using the *gradient descent active appearance* search algorithm [17]. This takes as input an image, the shape model and the appearance model, and outputs the corresponding landmarks for each frame. This labelling can be done using any face tracker, however active appearance models and their descendent's have the advantage that they use the same models as used by the synthesiser. Hence the points on the face located by the tracker are exactly the points required by the synthesiser.

Given the landmarks, each image is projected into face-space by computing the shape parameters, warping the image from the landmarks to the mean shape and computing the appearance parameters. Each example image corresponds to a point in face-space and over the course of a sentence the parameters approximate a trajectory through face-space. A continuous parametric representation of this trajectory is obtained using Hermite interpolation [18], and the 279 continuous trajectories, one for each training sentence, are stored in the synthesis codebook. Hermite interpolation is used to fit the data rather than natural cubic splines as the smoothness constraints in the calculation of the natural cubic spline often results in an overshoot of the data points. If, say, a point of curvature along the parameter trajectory corresponds to mouth opening, the overshoot could result in the mouth opening further than actually occurred in the original data and the auditory and visual information could become misaligned.

4.1. Segmenting the Trajectories

The audio component of the training video is passed through the HTK speech recogniser [19], the output of which is a list of the constituent phoneme symbols that form each sentence and their corresponding start and end times. This phonetic information is also stored in the synthesis codebook and is later used to index the parameter trajectories such that segments can be extracted corresponding to individual phonemes.

4.2. Measuring Phoneme Similarity

It is well known that during speech lip shapes depend not only the sound being produced, but also the surrounding sounds — known as phonetic context. The shape and appearance models are 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 an unseen context. This similarity matrix is automatically derived from the training data and each element contains an objective measure of similarity, in terms of the model parameters, between two given phonemes. This idea is similar to that in [20], except we extend their idea to consider the time variation of the parameters, the degree to which phonemes are modified by context and the relative significance of each model parameter.

To build the matrix, first all observations of each phoneme are gathered and the relevant portions of the original trajectories sampled at five equi-distant points over the duration of the phoneme[1] Next, the mean representation of each phoneme is computed and the distances found on a pair-wise basis using,

$$D_{ij} = \sum_k \sum_l \left[\left(v_i P^i_{kl} - v_j P^j_{kl} \right) w_k \right]^2, \qquad (1)$$

where D_{ij} is the distance between phonemes i and j. P^i is the mean matrix representing the i^{th} phoneme and P^j the j^{th} phoneme. The weights v take into account the degree to which the context modifies the lip shape for a phoneme, i.e. how reliable the mean representation is for a phoneme. For each phoneme, its weight is proportional to the total area between the mean trajectory and all of the observed trajectories. The value w_k is the significance of the k^{th} parameter in the model and is proportional to the variance captured by the corresponding principal component.

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

$$S_{ij} = e^{-\gamma D_{ij}}. \qquad (2)$$

The range of similarity is 0 (maximally dissimilar), to 1 (identical) and the variable γ controls the spread of similarity values over the range (0,1). This similarity matrix is stored with the parameter trajectories and phoneme timing information in the synthesis codebook. Some typical similarity values are shown in Table 4.2.

5. Synthesis

A visual sequence corresponding to a new utterance is synthesised by first converting a text stream to a list of phonemes and durations. This can either be from analysis of a real (unseen) utterance, or derived from a text-to-speech (TTS) synthesiser.

[1] The choice of sampling at five equi-distant points follows [20], we have also used the continuous trajectory representations and calculated the distances analytically, which gives similar results.

Phoneme	Rank 1		Rank 2		Rank 3	
m	p	0.869	b	0.850	w	0.830
f	v	0.808	s	0.621	dʒ	0.619
t	d	0.967	ɪ	0.900	z	0.894
tʃ	dʒ	0.898	ʃ	0.852	s	0.767

Table 1: Some typical phoneme similarity scores. The column Rank 1 is the most similar phoneme with its similarity score, Rank 2 the second most similar and so on. Generally the most similar phonemes belong to the same class of sound, for example the bilabials /b/, /m/ and /p/ are all considered similar, as with the labio-dental fricatives, /f/ and /v/.

For each phoneme to be synthesised, the original training data is searched for the n examples of that phoneme in the most similar contexts found in the codebook using

$$\mathbf{s}_j = \sum_{i=1}^{C} \frac{\mathbf{S}_{l_{ij}}}{i+1} + \sum_{i=1}^{C} \frac{\mathbf{S}_{r_{ij}}}{i+1}, \tag{3}$$

where $\mathbf{s}_j$ is the similarity between the desired context and the j^{th} context in the codebook, C is the context width, $\mathbf{S}_{l_{ij}}$ is the similarity between the i^{th} left phoneme in the j^{th} codebook 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} codebook context and the corresponding phoneme in the desired context. This similarity score is attractive since it allows the context width to be easily varied by simply changing an input parameter to the synthesiser (C), the structure of the synthesiser itself requires no modification. In the results presented here a context width of $C = 1$ is used, hence, the synthesis unit is the triphone. 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. A weighted average of these normalised trajectories is computed to give a new trajectory in face-space, where the weights are proportional to the similarity of the codebook context to the synthesis context, ensuring the most similar contexts receive more weight and that the sum of the weights is unity.

The new phoneme trajectories in face-space are concatenated to form a trajectory for the entire sentence, which is sampled at the original frame rate. Since no smoothness constraints were placed on the examples selected from the codebook, smoothing splines [21] 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. The synthesiser itself outputs a sequence of 2D landmarks and a sequence of shape-normalised images. The final synthesised image frames are created by warping the shape-normalised images to the corresponding landmarks.

Example parameter trajectories are shown in Figure 1, where the trajectory for the first parameter for the shape and appearance models are shown for an original (novel) sequence and the synthesised equivalent. While there are systematic differences between the trajectories, the overall shape is generally correct. Formal subjective testing is required in order to determine the significance of the differences between these trajectories. Results of early subjective tests are given in Section 6. A comparison of original and synthesised faces from a real sequence and the corresponding synthesiser output are shown in Figure 2. The data for the original sequence was not included in the synthesis codebook.

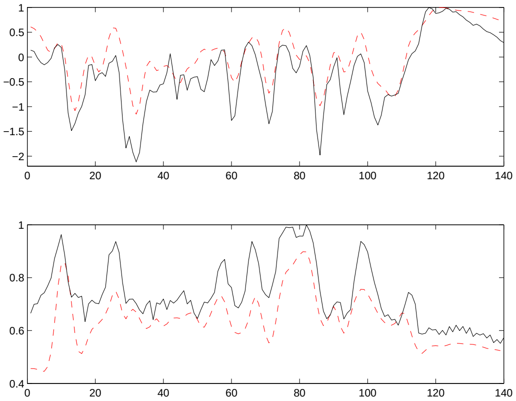

Figure 1: Upper plot shows the first shape model parameter trajectory from an original sequence (solid curve) and a synthetic sequence (dashed curve). The lower plot shows the same information, but for the first appearance parameter. The trajectories correspond to the phrase "Charlie brought his dog out but their only pure intent was to catch churchgoers wearing turquoise."

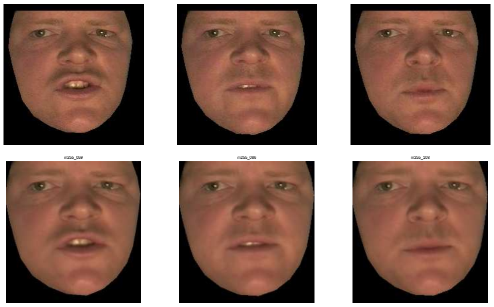

Figure 2: The top rows shows pixel values extracted from selected video frames from a real video sequence not used in training, while the bottom row shows the corresponding face output by the synthesiser.

The synthesis method described here for creating near-videorealistic synthetic visual speech sequences has the advantage over traditional image-based systems in that the manipulation of the original data is much easier in terms of the model parameters than the original images. The resultant sequences are still only 2D image sequences of a talking face however. It just happens that the images are created by the generative model, rather than obtained directly from a camera. The next section describes how the synthesiser can be easily extended to animate a 3D mesh model, providing near-videorealistic 2.5D animations.

6. Evaluation of Talking Faces

The quality of the output of a synthesiser can be measured using both subjective and objective tests. Objective measures of performance are attractive because they are automatic and repeatable. Numerical comparisons are made between some parameterisation of an original utterance and its synthetic equivalent, with the difference giving a measure of the distortion in the synthesised output. Objective measures can only be used as a guide however, since it remains difficult to determine the overall *naturalness* of the synthesiser output using only objective methods. Subjective measures may seem less attractive as they require a panel of users to make judgements regarding the performance of the system, however it should be remembered that it is the human perception of the performance that is the ultimate benchmark.

Subjective measures of quality include the *naturalness, acceptability* and *intelligibility* [22]. Intelligibility is a measure of the information provided the synthesiser, which for visual speech usually requires lipreading tests. Acceptability measures how suitable a system is for a given application. For example, for a particular application the user interface need not be video-realistic and a graphics model may suffice [23]. Naturalness is a general measure of performance that indicates the smoothness and realism of the dynamics of the features of the face. The following sections outline subjective experiments conducted to determine the naturalness of the synthetic visual speech output by the synthesiser.

6.1. Testing the Effect of Parameter Smoothing

The synthesiser described above imposes no smoothness constraints on the units selected from the training corpus. The assumption was made that any discontinuities at the concatenation boundaries will effectively be removed using the smoothing spline [21]. The aim of this test was to determine whether the smoothing spline significantly affects the naturalness of synthesiser output. The test used here follows the *double stimulus continuous quality scale* (DSCQS) method outlined in ITU BT.rec 500 [24]. This is a set of tests designed to evaluate the performance of new video coding techniques against a reference system. In the DSCQS method, sequences are presented in pairs to a viewer who is asked to judge the quality of each. The sequences are rated on a continuous scale from 1 to 5, corresponding to the levels “bad”, “poor”, “fair”, “good” and “excellent”. The scores are usually collected on paper, where users are asked to strike through the scale at the point corresponding to the quality. Here a graphical user interface (GUI) presents the movies and a slider collects the score from the user. The GUI approximates a continuous scale by collecting scores in the range 1 to 50, which are then divided by 10 to give the ITU scale.

The sequences presented in this test were the original video projected into face-space and the same sequence with the parameters smoothed using the smoothing spline. This is essentially a video coding problem, where the unsmoothed sequences represent the reference system and the smoothed sequences the system under test.

Eight subjects took part in this test and all were asked to watch the sequences and rate the naturalness of the dynamics of the face. The original acoustic signal was played with the visual sequences in order to provide a reference for the presented material. In all twenty sentences were presented in pairs (smoothed and unsmoothed), where the order of the pair is randomised.

6.1.1. Results

The result of a two-sample Wilcoxon’s signed rank test [25] on individual viewer responses is shown in Table 2, where N is the total number of observations, n is the number of observations used (sequence pairs with a difference in naturalness rating not equal to zero), W is the Wilcoxon test statistic and p the probability value. Viewers 1–3 detected a significant reduction in the naturalness of the smoothed sequences, the unsmoothed sequences were *always* rated more natural than the smoothed. The remaining five of the eight viewers did not detect a significant reduction in the naturalness of the smoothed sequences ($p < 0.01$), indeed viewer 6 preferred the smoothed sequences overall. Feedback from the subjects suggested that the smoothing splines gives the effect of “lazy” speech, i.e. the articulation strength is lower for the smoothed sequences and movements appear slower.

N	n	W	p	Median
20	20	0.0	0.000	-20.50
20	20	0.0	0.000	-16.50
20	20	0.0	0.000	-13.50
20	20	38.0	0.013	-15.00
20	16	23.5	0.023	-6.500
20	18	122.0	0.117	3.750
20	20	67.5	0.167	-3.500
20	20	92.5	0.654	-1.000

Table 2: Result of the per-viewer Wilcoxon signed rank test to determine the effect of the smoothing spline on the synthesiser output.

6.2. Testing the Naturalness of Sentence Level Synthesis

The purpose of this experiment was to determine the effect of the number of observations extracted and blended from the synthesis corpus on the naturalness of the synthesised output. In order to determine the how well the longer term effects of coarticulation are modelled, sentences were synthesised and played back to the viewer.

Five test conditions were used, random lip movements synchronised to the original acoustic speech signal, a single example ($N = 1$) extracted from the corpus for each synthesis phoneme, $N = 3$ and $N = 5$ observations extracted and blended, and the original (smoothed) parameters. In all cases the original speech signal was played back with the synthetic output. The random lip movements and original parameters were included to provide a upper and lower bound on the performance of the synthesiser.

The test data consisted of 20 sentences drawn at random from the training corpus and held out from training the synthesiser; each sentence was presented five times and played back in a randomised order. The viewers were again asked to watch the sequences and rate the naturalness of the dynamics of the face.

6.2.1. Results

The responses were subjected to a Kruskal-Wallis test[2][25] in order to determine whether there were significant differences between the various synthesis methods, the original sequences and the random sequences. The result of the test for all sequence types is shown in Table 3. It is clear that the distribution of at least one of the sequence types differs ($p = 0$). The median naturalness score for the random mouth movements (6) is considerably less than for the other four sequence types (> 30). This is promising in that the naturalness of the synthesiser output is significantly better than random movements (worst case), and is close to the original smoothed sequences (best case). A point to note from this experiment, the smoothed sequences here are

[2] For the case of testing only two distributions, the Wilcoxon signed-rank test and the Kruskal-Wallis test are equivalent.

judged more natural than the smoothed sequences in the test described in Section 6.1, and as natural as the original unsmoothed sequences. This is most likely because in the previous experiment the smoothed sequences were compared *directly* to the unsmoothed and the loss of subtle movements would be less obvious if the sequences were compared indirectly.

Sequence Type	n	Median	Ave Rank	Z
Random	160	6	90.5	-18.97
$N = 1$	160	30	420.2	1.20
$N = 3$	160	33	462.5	3.79
$N = 5$	160	32	444.7	2.70
Original	160	37	584.7	11.27
Overall	800		400.5	
$H = 408.04$		$p = 0.000$		

Table 3: Result of the Kruskal-Wallis analysis for the synthesis conditions; random lip movements, $N = \{1, 3, 5\}$ observations extracted and blended from the synthesis corpus and the original (smoothed) parameter trajectories in the presentation of sentences.

To test for a significant difference in the naturalness of original and synthesised sequences, the Kruskal-Wallis test was repeated with the random lip movements removed, shown in Table 4. To determine if there is any significance when varying the top N examples extracted from the codebook on the naturalness, the test was again repeated without the random lip movements and the original sequences, shown in Table 5.

Sequence	n	Median	Ave Rank	Z
$N = 1$	160	30	263.7	-4.48
$N = 3$	160	33	304.9	-1.23
$N = 5$	160	32	287.6	-2.60
Original	160	37	425.8	8.32
Overall	640		320.5	
$H = 73.18$		$p = 0.000$		

Table 4: Result of the Kruskal-Wallis analysis for the synthesis conditions; $N = \{1, 3, 5\}$ observations extracted and blended from the synthesis corpus and the original (smoothed) parameter trajectories in the presentation of sentences.

Sequence	n	Median	Ave Rank	Z
$N = 1$	160	30	223.3	-1.92
$N = 3$	160	33	256.4	1.78
$N = 5$	160	32	241.8	0.14
Overall	480		240.5	
$H = 4.60$		$p = 0.1$		

Table 5: Result of the Kruskal-Wallis analysis for the synthesis conditions; $N = \{1, 3, 5\}$ observations extracted and blended from the synthesis corpus in the presentation of sentences.

The result of these tests show that the naturalness scores for the synthesiser output are significantly lower than the original sequences, but there is no significant difference in selecting the top N, for $N = \{1, 3, 5\}$ examples, from the codebook. The difference between the synthesiser output and the original sequences could be attributed to the fact that the original audio signal was played back to the viewer with the visual sequences. In this case the original audio and visual information come from the same video sequence. If an utterance is spoken more than once and analysed in terms of the model parameters, there will undoubtedly be differences between the parameters due to the natural variability in the speech production process. It is therefore unfair to expect the synthesiser to exactly replicate the original sequence, and it would be useful to repeat this experiment using synthetic audio rather than the original.

Although the difference is not significant in a statistical sense, selecting only one example from the codebook resulted in less natural sequences than selecting five examples, which in turn was less natural that selecting three examples. The reason selecting a single observation performs worst is because the particular example extracted could be an over (or under) articulation of a particular mouth shape. Selecting more than one and generating a new trajectory as a weighted average ensures that over and under articulations are removed. The reason selecting more and more examples does not significantly affect the naturalness is because the new trajectory is a weighted average of the selected examples, hence as more and more are selected their influence in the new trajectory becomes less and less. The new trajectory is always formed from examples of the correct phoneme, but as more examples are used the subtle differences due to context are averaged out in the less similar examples.

Similar tests were conducted using mono-syllabic words as stimulus rather than sentences, where mono-syllabic words provide the smallest meaningful unit that can be tested. The idea being to test the short term naturalness of the synthesiser. Removing over articulations makes longer sequences look more natural overall, but may have more of a significant effect on shorter sequences since these are usually articulated more clearly. The results are not shown here due to a lack of space, but were similar to those presented for sentences.

7. Conclusions

In this paper we have presented an alternative to existing techniques for creating highly realistic synthetic visual speech. The synthesiser generates a new trajectory in face-space corresponding to a novel utterance from example parameter trajectories in a corpus. The parameters are applied to the model to create a 2D set of landmarks and a shape-normalised image. The final synthetic video frame is generated by warping the shape-normalised image to the 2D landmarks. The synthesis strategy is very simple and creates highly realistic animations, see http://bjtpc.sys.uea.ac.uk for demos.

Formal subjective testing of the synthesiser shows that the naturalness is approaching that of original sequences coded in terms of the model parameters. A Turing test reported in [15] showed that by simply judging the dynamics of the system, the synthetic sequences are indistinguishable from model encoded sequences. The short-fall in the naturalness in the tests reported here could be attributed to the fact the original audio was used in the test rather than synthetic auditory speech. The tests will be repeated with synthetic audio to see how using real auditory speech influences the perceived naturalness. Also, the original sequences could be captured twice, and one set of sequences used for synthesis and the other for testing. The auditory component for the test sequences would come from the training sequence, but re-synchronised to the test sequences. Another factor that could possibly influence the naturalness is the face in the test sequences is presented as a patch against a black background, see Figure 2. Re-compositing the face into an original video sequence may further improve the realism [13]. The face is then seen in the correct context, i.e. part of a complete body, and with hair etc.

The synthesiser described here has been extended to animate the face of a full-bodied 3D avatar, details will be published in a separate paper.

8. References

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

[2] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” *Nature*, vol. 264, pp. 746–748, 1976.

[3] F. Parke, “A parametric model for human faces,” Ph.D. dissertation, University of Utah, Saltlake City, Utah, 1974.

[4] B. Le Goff and C. Benoît, “A text-to-audiovisual-speech synthesizer for french,” in *Proceedings of the International Conference on Spoken Language Processing*, Philadelphia, Pennsylvania, 1996, pp. 2163–2166.

[5] S. Platt and N. Badler, “Animating facial expression,” *Computer Graphics*, vol. 15, no. 3, pp. 245–252, 1981.

[6] C. Pelachaud, “Communication and coarticualtion in facial animation,” Ph.D. dissertation, The Institute for Research in Cognitive Science, University of Pennsylania, 1991.

[7] K. Waters, “A muscle model for animating three-dimensional facial expressions,” *Proceedings of the International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, vol. 21, no. 4, pp. 17–24, 1987.

[8] 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.

[9] M. Brand, “Voice puppetry,” in *Proceedings of the International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, Los Angeles, California, 1999, pp. 21–28.

[10] N. Brooke and S. Scott, “Two- and three-dimensional audio-visual speech synthesis,” in *Proceedings of Auditory-Visual Speech Processing*, Terrigal, Australia, December 1998, pp. 213–218.

[11] E. Cosatto and H. Graf, “Sample-based synthesis of photorealistic talking heads,” in *Proceedings of Computer Animation*, Philadelphia, Pennsylvania, June 1998, pp. 103–110.

[12] T. Ezzat and T. Poggio, “Miketalk: A talking facial display based on morphing visemes,” in *Proceedings of the Computer Animation Conference*, Philadelphia, Pennsylvania, 1998, pp. 96–103.

[13] T. Ezzat, G. Geiger, and T. Poggio, “Trainable videorealistic speech animation,” in *Proceedings of the International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, San Antonio, Texas, July 2002, pp. 388–398.

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

[15] ——, “Near-videorealistic synthetic visual speech using non-rigid appearance models,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, 2003.

[16] T. Cootes, G. Edwards, and C. Taylor, “Active appearance models,” in *Proceedings of the European Conference on Computer Vision*, H. Burkhardt and B. Neumann, Eds., vol. 2. Freiburg, Germany: Springer-Verlag, 1998, pp. 484–498.

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

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

[19] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge: Entropic Ltd., 1999.

[20] L. 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.

[21] 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.

[22] C. Benoît and L. Pols, “On the assessment of synthetic speech,” in *Talking Machines: Theories, Models and Designs*, G. Bailly, C. Benoît, and T. R. Sawallis, Eds. Amsterdam: North-Holland, 1992.

[23] I. Pandzic, J. Ostermann, and D. Millen, “User evaluation: Synthetic talking faces for interactive services,” *The Visual Computer*, vol. 15, pp. 330–340, 1999.

[24] “Methodology for the subjective assessment of the quality of television pictures,” 1974-1978-1982-1986-1990-1992-1994-1995-1998-1998-2000.

[25] D. Wakerly, W. Mendenhall, and R. Scheaffer, *Mathematical Statistics with Applications*. Duxbury Advanced Series, 2002.

[26] G. Bailly, C. Benoît, and T. R. Sawallis, Eds., *Talking Machines: Theories, Models and Designs*. Amsterdam: North-Holland, 1992.