Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 44 (2004) 127–140
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Near-videorealistic synthetic talking faces: implementation and evaluation

B.J. Theobald [a,*], J.A. Bangham [a], I.A. Matthews [b], G.C. Cawley [a]

[a] School of Computing Sciences, University of East Anglia, Earlham Road, Norwich, Norfolk NR4 7TJ, UK
[b] Robotics Institute, Carnegie Mellon, Pittsburgh, PA 15123, USA

Received 21 January 2004; received in revised form 14 July 2004; accepted 29 July 2004

Abstract

The application of two-dimensional (2D) shape and appearance models to the problem of creating realistic synthetic talking faces is presented. A sample-based approach is adopted, where the face of a talker articulating a series of phonetically balanced training sentences is mapped to a trajectory in a low-dimensional model-space that has been learnt from the training data. Segments extracted from this trajectory corresponding to the synthesis units (e.g. triphones) are temporally normalised, blended, concatenated and smoothed to form a new trajectory, which is mapped back to the image domain to provide a natural, realistic sequence corresponding to the desired (arbitrary) utterance. The system has undergone early subjective evaluation to determine the naturalness of this synthesis approach. Described are tests to determine the suitability of the parameter smoothing method used to remove discontinuities introduced during synthesis at the concatenation boundaries, and tests used to determine how well long term coarticulation effects are reproduced during synthesis using the adopted unit selection scheme. The system has been extended to animate the face of a 3D virtual character (avatar) and this is also described.

Keywords: Talking faces; Shape and appearance models; Avatars; Dynamic textures

1. Background

It is well known that speech is a multi-modal form of communication; seeing the face of a talker provides information that can significantly influence the perception and understanding of speech, as illustrated by both the McGurk effect (McGurk and MacDonald, 1976) and the speech-reading ability of humans (and machines) (Stork and Hennecke, 1996). It is, therefore, reasonable to focus attention on the visual modality in addition to the auditory in the synthesis of speech. Traditional approaches to synthesising talking faces can be broadly classified

* Corresponding author. Tel.: +44 1603 591122/1502 568386.
E-mail addresses: bjt@cmp.uea.ac.uk (B.J. Theobald), ab@cmp.uea.ac.uk (J.A. Bangham), iainm@cs.cmu.edu (I.A. Matthews), gcc@cmp.uea.ac.uk (G.C. Cawley).

0167-6393/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2004.07.002

as either *model-based* or *image-based*. Model-based systems tend to use techniques from computer graphics, where points on the face are represented as vertices in 3D and the surface itself approximated by connecting the vertices to form a mesh. The mesh then is animated by applying the appropriate time-varying parameters, e.g. (Parke, 1974; Massaro, 1998; Pelachaud, 1991; Waters, 1987). Image-based systems use computer vision or image processing algorithms to analyse images of real faces, which can later be post-processed to re-synthesise the face in a video sequence, e.g. (Bregler et al., 1997; Brand, 1999; Cosatto and Graf, 1998; Ezzat et al., 2002; Huang et al., 2002). There is a trade-off between flexibility and realism. Model-based systems are flexible and can be efficiently rendered, especially on modern graphics processors, but they tend to lack videorealism—a videorealistic system is defined as one that is indistinguishable from a recorded sequence of a talker, regardless of speech content. 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. At the cost of computational expense and a distinct lack of flexibility, image-based systems can achieve close to videorealism providing the correct lip shape is presented for a given sound and the synthesised movements on the face look natural. Image-based systems are also limited in their application. Only the face in a sequence is re-animated, the full character cannot perform novel actions. The approach adopted in this work is based on shape and appearance models and can be considered a hybrid image-based/model-based approach. A statistical model of the appearance of the face is texture mapped onto a 3D mesh model, which in turn is animated by a statistical model of shape. Thus, pose, shape and texture are all animated independently. Potential applications for the 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.

1.1. Previous work

The system proposed here has similarities with several systems reported previously. The Video Rewrite system by Bregler et al. (1997) was one of the first examples of a visual speech synthesiser that approached videorealism. A video sequence of a talker is segmented into short clips that correspond to triphones. Segments of this video can later be selected according to the similarity of a desired triphone and each candidate triphone in the video. The selected segments are then concatenated and the neighbouring regions cross-faded to ensure a smooth transition, resulting in new sequences of a talker uttering novel phrases. The similarity between triphones is measured in terms of the visual confusion of the individual phonemes that form the triphones, and a set of mouth shape parameters defining the triphones (e.g. width and height). The visual confusions used in Video Rewrite are drawn from confusion matrices that had been published previously, e.g. (Owens and Blazek, 1986). The disadvantage of using generic data (i.e. not talker specific) in this way is that the clustering of mouth shapes associated with speech is poorly defined and has been found to be talker dependent (Kricos and Lesner, 1982).

Arslan and Talkin (1998) also proposed a concatenative synthesiser. An optical tracking system tracks the (xyz) positions of a set of markers attached to the face of talker enunciating a series of training sentences. The trajectory of the marker positions is segmented into a lookup codebook, where each entry corresponds to an observation of a phone and contains the marker positions, the phoneme symbol and a context label (two phoneme symbols either side of the centre-phone). A similarity score is then computed between each phoneme pair, so, during synthesis, examples can be selected from the codebook in contexts that are closest to a desired context. Concatenating the selected segments creates a new trajectory for the markers, which can be used to animate a three-dimensional (3D) graphics model of the face. The advantage of this approach is that the similarity of speech segments is entirely speaker dependent. However, since the animation is based on graphics models, this approach lacks the static realism (photorealism) of systems such as Video Rewrite.

Ezzat et al. (2002) proposed a (potentially) videorealistic system based on *multi-dimensional mor-*

phable models (MMMs). Images are selected to represent the key mouth shapes associated with speech production and a further image selected as a reference. The set of optical flow vectors that morph the reference image to each mouth shape image are computed, which form the shape component of the MMM. The shape parameters define the linear contribution of the optical flow vectors that when applied to the reference image generate a set of morphed images, and the appearance parameters define the contribution of these morphed images in the synthesis of the final frame.

To train the synthesiser, images of the mouth of a talker in a training video are projected into model-space and each phoneme represented as a multi-dimensional Gaussian. A trajectory of shape parameters and a trajectory of appearance parameters for a novel utterance are computed using regularisation. The parameters are then applied to the model to generate an image sequence of a talker's mouth, and these images are re-composited back into an original video sequence to create realistic synthetic visual speech sequences. Using original background scenes gives natural eye and eyebrow movements, which serves to improve the realism of the system. However, the disadvantage of this approach is that only the face can be re-animated, the character cannot perform novel actions. Also, the entire phoneme sequence forming the target utterance must be known before the trajectory can be computed, making a real-time implementation difficult.

The system proposed in this work is closely related to the three described above. The basis of the system is a shape and appearance model, see Section 2, which, like the model in (Ezzat et al., 2002) is a generative model. The advantage of the shape and appearance model over the MMM is that the shape component of the model describes the movement of the features of the face directly in terms of a coordinate system. Thus, the geometry of the face of a complete character can be animated using the shape component of the model. As in (Bregler et al., 1997), the idea is to select segments from the training data based on the similarity of triphone segments and concatenate the selected segments to form new sequences. The unit selection approach is based on that in (Arslan and Talkin, 1998), however both the shape and appearance of the face is considered, whereas in (Arslan and Talkin, 1998) only the shape is considered. Also, the temporal evolution of each phoneme, the degree to which each phoneme is modified by context and the relative significance of each model parameter are also considered. The final output from the synthesiser can be in the form of a 2D image sequence, which can re-composited into an existing video sequence (as in (Ezzat et al., 2002)), or can be used to create realistic sequences that can be combined with other manual gestures (e.g. deaf-signing) by animating a complete character.

2. Shape and appearance models

To construct a shape and appearance model, a set of images are first hand-annotated with landmarks that delineate the shape. The vector, $\boldsymbol{s}$, describing the shape in an image is given by the concatenation of the x and y-coordinates of the landmarks that define the shape: $\boldsymbol{s} = (x_1, y_1, \ldots, x_n, y_n)^{\mathrm{T}}$. A compact model that allows a linear variation in the shape is given by

$$\boldsymbol{s} = \boldsymbol{s}_0 + \sum_{i=1}^{k} \boldsymbol{s}_i \boldsymbol{p}_i, \tag{1}$$

where $\boldsymbol{s}_i$ form an orthonormal set of basis shapes and $\boldsymbol{p}$ is a set of k shape parameters that define the contribution of each basis in the representation of $\boldsymbol{s}$. The basis shapes may be derived from the hand-annotated examples using *principal component analysis* (PCA), and the shape model is often referred to as the *point distribution model* (PDM) (Cootes et al., 1998). In this case, the base shape, $\boldsymbol{s}_0$, is the mean shape and the basis shapes are the k eigenvectors of the covariance matrix corresponding to the largest eigenvalues.

To ensure that the model generates legal shapes, in the sense of the training examples, the parameters are constrained to lie within some limit, typically ±2 standard deviations, from the mean. The result of varying the first four parameters of a typical shape model is shown in Fig. 1.

A compact model that allows a linear variation in the appearance of the face is given by

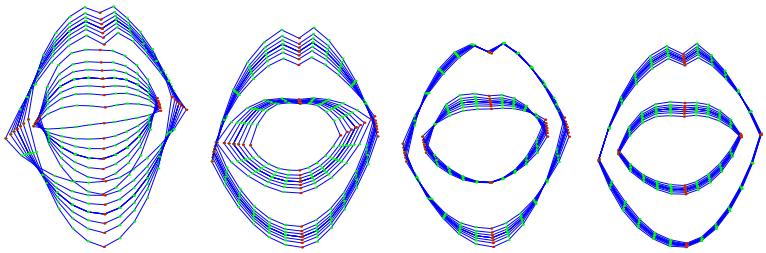

Fig. 1. Varying the first four parameters of a shape model through ±2 standard deviations from the mean. Typically, 10 parameters are required to account for 95% of the shape variation. For the model shown here, the first parameter appears to capture the opening and closing of the mouth and the second the degree of lip rounding. Subsequent modes capture more subtle variations in the mouth shape.

$$\boldsymbol{A} = \boldsymbol{A}_0 + \sum_{i=1}^{l} \boldsymbol{A}_i \boldsymbol{q}_i, \tag{2}$$

where $\boldsymbol{A}_i$ form an orthonormal set of basis images and $\boldsymbol{q}$ is a set of l appearance parameters that define the contribution of each basis in the representation of $\boldsymbol{A}$. The elements of the vector $\boldsymbol{A}$ are the (RGB) pixel values that are bound by the base shape, $\boldsymbol{s}_0$. Again, as with the shape, the orthonormal set of basis images can be found using PCA. In (Cootes et al., 1998), shape variation is first removed from the images on which the model is trained by warping each example from the hand-annotated landmarks, $\boldsymbol{s}$, to the base shape. This ensures each example has the same number of pixels and that a pixel in one example corresponds to the same feature of the face in all other examples. An example of an appearance model, the base and first three basis images, is shown in Fig. 2.

Any example face image can be described by a set of shape parameters and a set of appearance parameters, $\boldsymbol{p}$ and $\boldsymbol{q}$ respectively. Applying the shape parameters to the shape model, Eq. (1), generates a set of landmarks, $\boldsymbol{s}$. Applying the appearance parameters to the appearance model, Eq. (2), generates an appearance image. The final synthesised image of the face is generated by warping the new appearance image from the base shape to the landmarks $\boldsymbol{s}$. Note, we do not project the shape and appearance parameters into a combined space (as in (Cootes et al., 1998)) 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 (Theobald et al., 2003). Also, although the models shown here are of the whole face, the synthesiser is primarily concerned only with the synthesis of the visible articulators.

3. Data capture and preparation

The training data for the talking head consists of a single talker enunciating approximately 300 sentences (around twelve minutes of speech, in accordance with similar systems, e.g. (Ezzat et al., 2002; Huang et al., 2002)). To confine the variation of the facial features to only the gestures related directly to speech production, the talker was instructed to maintain a neutral expression

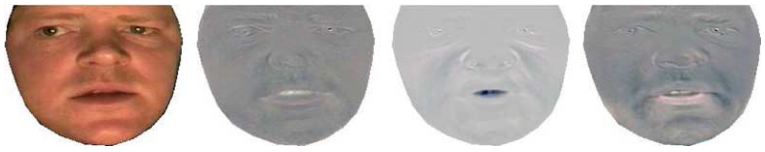

Fig. 2. The base and first three basis images of an appearance model. Note: the basis images have been suitably scaled for visualisation. Typically, between 15 and 30 parameters are required to account for 95% of the appearance variation.

(no emotion) throughout the recording. The video was captured using a head mounted camera to ensure the pose of the head remained as constant as possible and was transferred from DV tape to computer using an IEEE 1394 compliant capture card with a frame size of 360×288 pixels at a frame rate of 25 fps (i.e. one quarter DV-PAL). The audio was captured using the on-camera microphone and digitised at 11 025 Hz, 16 bits/sample stereo. The capture conditions were controlled such that unwanted sources of variation, i.e. identity, pose and lighting, were, as far as possible, minimised. The training video contained approximately 34 000 frames (including both speech and silence) and each frame was mapped to the corresponding point in the space spanned by the shape and appearance model using the *gradient descent active appearance model* search algorithm (Baker and Matthews, 2001). [1] This uses gradient descent optimisation to automatically find the parameters that generate a synthetic (model-generated) face image that is as close as possible to the face in the corresponding video frame—i.e. the error between the original and synthesised face images is minimised. Given an initial set of parameters, the algorithm iteratively solves for updates to the parameters until there is little or no change between iterations. Since the face in a video frame forms a point in the model-space, the movements of the face corresponding to a sentence approximate a trajectory through the model-space—it is only an approximation since the trajectory exists only at each frame. A continuous parametric representation of this trajectory is obtained using Hermite interpolation (Bartels et al., 1987) and is stored in a synthesis codebook. Hermite interpolation is used to fit the data rather than natural cubic splines as the second order smoothness constraints in the calculation of the natural cubic spline often results in an over-smoothed fit of the data points. This is particularly significant when rapid changes in the trajectory of the parameters (or in the visible articulators) is required, during the plosives /b/ and /p/ for example. Using a natural cubic spline, the acceleration of the articulators would be required to be smooth.

The HTK speech recogniser was used in forced alignment mode (Young et al., 1999) to segment the trajectory by aligning the constituent phoneme symbols that form the sentences to the audio component of the training video. The timing information returned by the recogniser is stored in the synthesis codebook and is later used to index the trajectory in the model-space such that segments can be extracted corresponding to individual phones, or groups of phones.

3.1. Measuring phoneme similarity

It is well known that during speech lip shapes depend not on only the sound being produced, but also the surrounding sounds—known as phonetic context. The trajectory in the synthesis codebook is formed from the limited number of training sentences and so contains only a subset of all possible contexts in which each phoneme may appear. A synthesiser must be capable of synthesising entirely arbitrary utterances, so some method of selecting contexts from the training data that are 'closest' to a previously unseen context is required. The scheme adopted here is similar to that in (Arslan and Talkin, 1998) as it is automatically learned from the data on which the synthesiser is trained. There is no manual specification of the similarity between synthesis units, as in (Bregler et al., 1997) for example. The scheme in (Arslan and Talkin, 1998) is extended here to consider the time variation of the synthesis parameters, the appearance information in the face, the degree to which each phoneme is modified by context and the relative significance of each model parameter. A similarity matrix is automatically constructed from the training data, where each element contains an objective measure of similarity, in terms of the shape and appearance parameters, between each phoneme pair. To build the matrix, first all observations of each phoneme are gathered and the relevant sub-trajectories are extracted from the original trajectory.

[1] In principle any face tracker that uses shape and appearance models can be used. The choice is arbitrary. The gradient descent active appearance model was selected as it has proved to be both fast and reliable.

These are then sampled at five equi-distant points over the duration of each observation. [2] Next, the mean representation of each phoneme is computed and the distances found on a pair-wise basis using,

$$D_{ij} = \sum_{m=1}^{k+l} \sum_{n=1}^{5} [(v_i P^i_{mn} - v_j P^j_{mn}) w_m]^2, \quad (3)$$

where the first summation is over the number of parameters in the shape and appearance model and the second over the five equally space samples. The value D_{ij} is the distance between phonemes i and j and P^i is the mean representation of the ith phoneme and P^j the jth 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 each phoneme, the weight is proportional to the variance of the area between the mean and observed sub-trajectories, so those that are more modified by context are penalised more heavily. This ensures that two phonemes with the same mean but unequal variances are not considered identical. In practise, phonemes that belong to the same class of sound (e.g. bilabial) have approximately equal means *and* equal variances and this scaling effectively amplifies the similarity value. The value w_m is the significance of the mth parameter in the model and is proportional to the variance captured by the corresponding principal component, i.e. how significant the parameter is in the representation of the data. Note, $\sqrt{D_{ij}}$ is the Frobenius norm of the weighted difference between the phoneme representations and this formulation approximates computing the area between the multi-dimensional curves representing each phoneme in the model-space (computing the area analytically results in similar distance scores).

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

$$S_{ij} = e^{-\gamma D_{ij}}. \quad (4)$$

[2] The choice of sampling at five equi-distant points follows (Arslan and Talkin, 1998). The effect of increasing the number of samples was tested in (Theobald, 2003) and found to have no significant change in the performance.

Table 1
Some typical phoneme similarity scores

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

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 are the labio-dental fricatives, /f/ and /v/, and so on.

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 trajectory and phoneme timing information in the synthesis codebook. Typical similarity values are given in Table 1.

4. Synthesis

The visual synthesiser is driven by a sequence of phoneme symbols that form the desired utterance and the duration of each. The input can be either an auditory utterance or a text stream. For an auditory input, an automatic speech recogniser (ASR) converts the utterance to the constituent phoneme symbols and durations, whereas for a textual input, a text-to-speech (TTS) synthesiser, e.g. (Black and Taylor, 1997), converts the input. 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

$$\delta_j = \sum_{i=1}^{C} \frac{\boldsymbol{S}_{l_{ij}}}{i+1} + \sum_{i=1}^{C} \frac{\boldsymbol{S}_{r_{ij}}}{i+1}, \quad (5)$$

where δ_j is the similarity between the desired context and the jth context in the codebook, C is the context width, $\boldsymbol{S}_{l_{ij}}$ is the similarity between the ith left phoneme in the jth codebook context and the corresponding phoneme in the desired context, $\boldsymbol{S}_{r_{ij}}$ is the similarity between the ith right phoneme of the jth codebook context and the corresponding

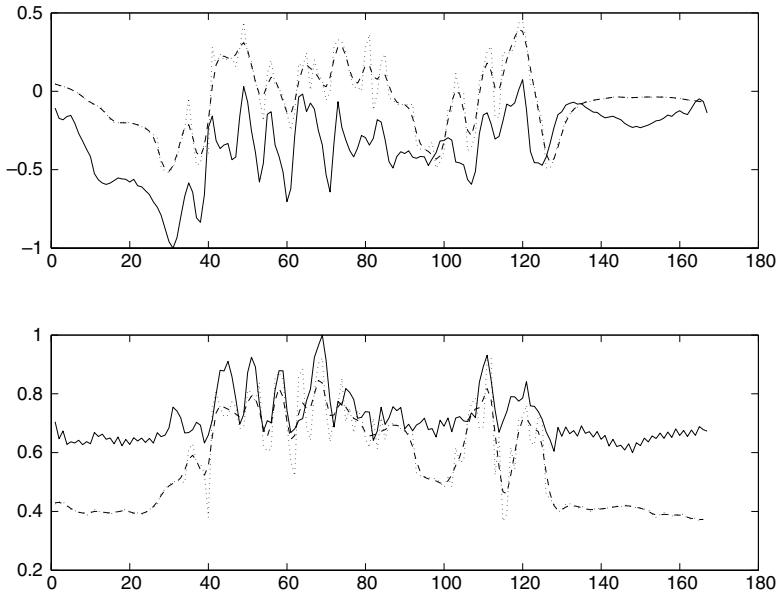

Fig. 3. Upper plot shows the first shape model parameter trajectory from an original sequence (solid curve), an unsmoothed synthesised sequence (dotted curve) and a smoothed synthesised sequence (dashed curve). The lower plot shows the same information, but for the first appearance parameter.

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. The effect of varying the context width has been tested using both objective and subjective testing, the results of which can be found in (Theobald, 2003). However, it was found that increasing the context width did not significantly improve the synthesiser output. Given the N closest matches in the codebook for each synthesis phoneme, the corresponding sub-trajectories from the original parameter trajectory are extracted and temporally warped to the desired duration. A weighted average of these normalised trajectories is computed to give a new trajectory in the model-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. Results of subjective tests to determine the effect of varying N (where $N = \{1, 3, 5\}$) on the naturalness of the synthesiser output are given in Section 5.

The new phoneme sub-trajectories in the model-space are concatenated to form a trajectory for the new sentence, which is sampled at the original frame rate. Since no smoothness constraints were placed on the examples selected from the codebook, cubic smoothing splines 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. [3] The synthesiser itself outputs a sequence of 2D landmarks and a sequence of appearance images. The final synthesised image frames are created by warping the appearance images to the corresponding landmarks.

Example parameter trajectories are shown in Fig. 3, where the trajectory for the first parameter

[3] The cubic smoothing spline is an approximating technique. Interpolating techniques, e.g. Hermite, as used previously, cannot be used as sampling the interpolating functional results in exactly the unsmoothed data.

Fig. 4. The top row shows pixel values extracted from selected video frames from an original video sequence not used in training, while the bottom row shows the corresponding face output by the synthesiser.

for the shape and appearance models are shown for an original (unseen) sequence and the synthesised equivalent. While there are systematic differences between the trajectories, the overall shapes are generally correct. The difference in bias manifests itself in a difference in articulation strength between the original and synthesised mouth gestures, as shown in Fig. 4. 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 5. In generating these examples, the data for the original sequence was not included in the synthesis codebook.

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 having been obtained directly from a camera. The next section describes some early subjective tests used to evaluate the naturalness of the synthesiser output, followed by an extension to the system that allows a full-bodied 3D talking person to be created.

5. 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 synthesised equivalent, with the difference giving a measure of the distortion in the synthesised output. Objective measures can be used only 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* (Benoît and Pols, 1992, in Bailly et al., 1992). Intelligibility is a measure of the information provided by the

synthesiser. Acceptability measures how suitable a system is for a given application. For example, for a particular application the user interface need not be videorealistic and a graphics model may suffice. 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. Experiments to determine the acceptability and intelligibility are on-going and will be presented in a future publication.

5.1. Testing the effect of parameter smoothing

The synthesiser described above imposes no smoothness constraints on units selected from the training corpus. The assumption was made that discontinuities at the concatenation boundaries will effectively be removed using the natural cubic smoothing spline. The aim of this test was to determine whether this smoothing significantly affects the naturalness of synthesiser output, i.e. how natural do the sequences appear if the required discontinuities found in natural speech are also removed? The test used here follows the *double stimulus continuous quality scale* (DSCQS) method outlined in ITU BT.rec 500 (Union, 1974–1978–1982–1986–1990–1992–1994–1995–1998–1998–2000). 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 1–50, i.e. approximating the continuous scale 1–5 to one decimal place.

The sequences presented in this test were the original video projected into the model-space and the same sequence with the parameters smoothed. 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, all postgraduate students and all non-expert in auditory or visual speech synthesis, 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 auditory signal from the training video was played with the visual sequences in order to provide a reference for the spoken material. In all 20 sentences were presented in pairs (smoothed and unsmoothed), where the order of the pair is randomised.

5.1.1. Results

The result of a two-sample Wilcoxon's signed rank test (Wakerly et al., 2002) 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 six 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.

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

Viewer	N	n	W	$p<$
1	20	20	0.0	0.001
2	20	20	0.0	0.001
3	20	20	0.0	0.001
4	20	20	38.0	0.01
5	20	16	23.5	0.02
6	20	18	122.0	0.12
7	20	20	67.5	0.17
8	20	20	92.5	0.65

5.2. Testing the naturalness of sentence level synthesis

Recall that during synthesis, the most similar contexts available in the codebook to the desired context are found. The original parameter sub-trajectories for those examples extracted from original trajectory and temporally normalised to the desired duration. They are then blended by computing a weighted average, and concatenated to form the synthesised parameter trajectory. The purpose of this experiment was to determine the effect of varying the number of observations extracted from the synthesis codebook. In order to determine 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 an upper and lower bound on the performance of the synthesiser. Note that the original video frames were not used. Rather, the video was projected into the model-space. The aim was to ensure the viewer makes judgements based on the dynamics of the synthesised sequences. Original video frames and frames synthesised using the model differ in pictorial quality—camera noise is removed during the averaging process in computing the PCA. During initial trials using original video, subjects stated that judgements were made based on the pictorial quality of the video frames and not necessarily the naturalness of the synthetic movements of the mouth.

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 (once for each test condition) to eight viewers (the same viewers from the previous experiment) and played back in a randomised order. The test is broadly similar to the previous experiment, but where sequences were presented in pairs (in Section 5.1), the order of the sequences in this test is randomised over all 100 sequences. The viewers were again asked to watch the sequences and rate the naturalness of the dynamics of the face.

5.2.1. Results

The responses were subjected to a Kruskal–Wallis test (Wakerly et al., 2002) 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, where n is the number of sequences, Median represents the median naturalness score (in the range 1–50) and Z is the Krukal–Wallis test statistic (Wakerly et al., 2002). It is clear that the distribution of at least one of the sequence types differs ($p < 0.001$). 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 judged more natural than the smoothed sequences in the test described in Section 5.1, and as natural as the original unsmoothed

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

Sequence type	n	Median	Z
Random	160	6	−18.97
$N = 1$	160	30	1.20
$N = 3$	160	33	3.79
$N = 5$	160	32	2.70
Original	160	37	11.27
$H = 408.04$		$p < 0.001$	

n is the number of sequences in the test, and Z the Kruskal–Wallis test statistic. Median is the median naturalness score for each test condition. Note, as with all non-parametric tests, the median is used rather than mean (since the median is a non-parametric quantity).

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.

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 significant change in the naturalness when varying the top N examples extracted from the codebook, the test was again repeated without the random lip movements and the original sequences, shown in Table 5.

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.

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	Z
$N = 1$	160	30	−4.48
$N = 3$	160	33	−1.23
$N = 5$	160	32	−2.60
Original	160	37	8.32
$H = 73.18$		$p < 0.001$	

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

Sequence	n	Median	Z
$N = 1$	160	30	−1.92
$N = 3$	160	33	1.78
$N = 5$	160	32	0.14
$H = 4.60$		$p < 0.1$	

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 has been noted (Bailly et al., 2003) that simply "stretching" parameter trajectories to align visual gestures to an audio track may not be sufficient to maintain coherence between the auditory and visual modalities. 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, or aligning audio from a second recording to the test video.

Although the difference is not significant, it could be expected that selecting a single observation should perform worst because the example extracted could be an over (or under) articulation of a mouth shape. Selecting more than one and generating a new trajectory as a weighted average should ensure that over and under articulations are attenuated. 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.

6. Extending the synthesis to 2.5D

The synthesiser described in Section 4 provides very realistic 2D speech animation of the human face (2D in the sense that an image sequence is generated). The resultant synthetic faces can be composited back into an original video sequence, as with other 2D synthesis systems (Bregler et al., 1997; Ezzat et al., 2002). However, it is desirable to animate the face of a full-bodied 3D virtual character, rather than simply re-animating the mouth in an existing sequence. The character is then free to move around and interact with a virtual environment. Here, we adopt a technique based on scattered data interpolation used in

(Pighin et al., 1998) for animating facial expression. The 3D coordinates of a sparse set of points defined on the face of an individual are recovered from multiple camera views. These sparse points are then used to adapt a dense generic 3D mesh to the individual. Here, we use the same model used by the face tracker and synthesiser, where the sparse shape model landmarks are used to drive the dense 3D mesh and the appearance images provide a *dynamic texture map*. These texture maps are warped to the 3D vertices of the face mesh rather than the 2D landmarks of the shape model, providing near-videorealism in three dimensions. The actual animations produced by the synthesiser are essentially 2.5D since the shape model contains no depth variation—the depth information on the 3D avatar is held constant. The resultant animations are, however, still very realistic for moderate rotations of the head as the depth cues are captured in the subtle changes in the dynamic texture map.

First, a correspondence must be defined between the N 2D landmarks in the shape model and the $M \gg N$ vertices of the 3D mesh. This is done manually prior to synthesis and informs the synthesiser which vertex belongs to which point in the shape model. Vertices on the 3D mesh mapped to a point in the shape model are known as *constrained* vertices and the displacements for these vertices are known, they take the coordinates of the corresponding shape model points. The displacements for the constrained vertices are given by

$$\boldsymbol{u}_i = \boldsymbol{p}_i - \boldsymbol{p}_i^{(0)}, \tag{6}$$

where $\boldsymbol{p}^{(0)}$ is the 3D mesh in the default position, i.e. adapted to the mean shape in the shape model, and $\boldsymbol{p}_i$ are the new 3D coordinates for the ith constrained vertex. A smooth vector-valued function that fits the known displacements, $f(\boldsymbol{p}_i) = \boldsymbol{u}_i$, is defined such that the displacements of the remaining (unconstrained) vertices can be found using $f(\boldsymbol{p}_j) = \boldsymbol{u}_j$.

A radially symmetric basis function is used in (Pighin et al., 1998), which falls off smoothly with distance, thus the displacement of unconstrained vertices are more influenced by the displacement of constrained vertices lying closer by. The function $f(p)$ is defined as

$$f(p) = \sum_i \boldsymbol{c}_i \phi(\|\boldsymbol{p} - \boldsymbol{p}_i\|), \tag{7}$$

where, following (Pighin et al., 1998), the basis function takes the form $\phi(r) = \mathrm{e}^{-r/64}$ and the coefficients $\boldsymbol{c}$ are found by multiplying the (x, y, z) coordinates of $\boldsymbol{u}_i$ with the matrix Φ^{-1}, where $\Phi_{ik} = \phi(\|\boldsymbol{p}_i - \boldsymbol{p}_k\|)$, with $\boldsymbol{p}_i$ the ith constrained vertex and $\boldsymbol{p}_k$ the kth constrained vertex.

The original synthesiser training data was captured using a head mounted camera to minimise unwanted pose variation from the face model. In the synthesised sequences, pose information (translation and rotation) can be applied to the 3D mesh prior to rendering the face, hence the pose of the face is independent of the synthesis parameters. The mesh used to drive the model need not contain only a face, it could form part of a full virtual character, shown in Fig. 6. In this instance vertices are tagged prior to synthesis as belonging to the face of the avatar, or not. Those not forming the face are ignored, while those belonging to the face are displaced as described above. Example frames from an animated sequence using a generic mesh model are shown in Fig. 5 and an example of a full bodied talking avatar is shown in Fig. 6.

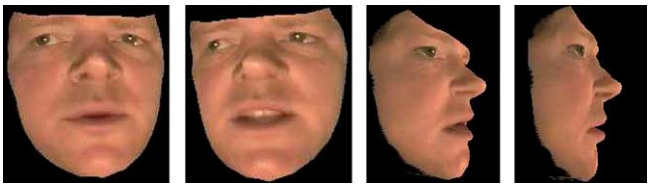

Fig. 5. Example frames from a sequence, where a generic mesh is deformed according to the 2D shape model landmarks and textured with the appearance images output by the synthesiser.

Fig. 6. Example frames from a sequence, where the face of a complete avatar is animated using a shape and appearance model.

7. Future work

Future work will include an investigation of how expressive speech can be animated using the model. Currently the synthesiser is trained on speech without emotional context. One approach to animating expressive speech would be to include existing graphics rules, for example Waters' muscle model (Waters, 1987), and apply the graphics rules to the 3D mesh after the speech animation has been generated. A second approach could be to capture a database of images with emotional expression in the same session as a speech corpus is captured. A separate shape and appearance model could then be trained on this database, and the leading modes of variation added to the speech model. One of the major limitations of image-based synthesis is the lack of generalisation—only the face(s) in the synthesiser corpus may be animated. Since our animation parameters are offsets from the mean shape and appearance, we will investigate how displacing the mean to a new position in the model-space affects the perceptual quality of the synthesiser output when animating new faces.

8. 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 an appearance image. The final synthesised video frame is generated by warping the appearance image to the 2D landmarks, or by adapting a generic 3D mesh to the landmarks and warping the shape-normalised image to the new mesh vertices. The latter allows the face of a complete virtual character to be animated with a high degree of realism, i.e. a "talking person", rather than a "talking head".

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 (Theobald et al., 2003) showed that by simply judging the dynamics of the system, the synthesised 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 auditory speech to determine how real 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 Fig. 4. Re-compositing the face into an original video sequence may further improve the realism (Ezzat et al., 2002). The face is then seen in the correct context, i.e. part of a complete body, and with hair etc.

Acknowledgments

The authors would like to thank all persons who took part in the perceptual tests and Vince Jennings for his help in mapping the output from

the synthesiser to the face of the avatar (shown in Fig. 6). The authors are also grateful to Dr. Adrian Clark (Essex University, UK) for his helpful comments and suggestions.

References

Arslan, L., Talkin, D., 1998. 3D face point trajectory synthesis using an automatically derived visual phoneme similarity matrix. In: Proceedings of auditory-visual speech processing. Terrigal, Australia, pp. 175–180.

Bailly, G., Benoît, C., Sawallis, T.R. (Eds.), 1992. Talking Machines: Theories, Models and Designs. North-Holland, Amsterdam.

Bailly, G., Bérar, M., Elisei, F., Odisio, M., 2003. Audiovisual speech synthesis. International Journal of Speech Technology 6, 331–346.

Baker, S., Matthews, I., 2001. Equivalence and efficiency of image alignment algorithms. In: Proceedings of the international conference on computer vision and pattern recognition, Kauai, Hawaii, pp. 1090–1097.

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

Benoît, C., Pols, L., 1992. On the assessment of synthetic speech. In: Bailly, G., Benoît, C., Sawallis, T.R. (Eds.), Talking Machines: Theories Models and Designs. North-Holland, Amsterdam, pp. 435–441.

Black, A., Taylor, P., 1997. The festival speech synthesis system. Technical Report No. HCRC/R-83. University of Edinburgh.

Brand, M., 1999. Voice puppetry. In: Proceedings of the international conference on computer graphics and interactive techniques (SIGGRAPH), Los Angeles, California, pp. 21–28.

Bregler, C., Covell, M., Slaney, M., 1997. Video rewrite: driving visual speech with audio. In: Computer Graphics Annual Conference Series (SIGGRAPH), Los Angeles, California, pp. 353–360.

Cootes, T., Edwards, G., Taylor, C., 1998. Active appearance models. In: Burkhardt, H., Neumann, B. (Eds.), Proceedings of the European Conference on Computer Vision, Vol. 2. Springer-Verlag, Freiburg, Germany, pp. 484–498.

Cosatto, E., Graf, H., 1998. Sample-based synthesis of photo-realistic talking heads. In: Proceedings of Computer Animation, Philadelphia, Pennsylvania, pp. 103–110.

Ezzat, T., Geiger, G., Poggio, T., 2002. Trainable videorealistic speech animation. In: Proceedings of the International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), San Antonio, Texas, pp. 388–398.

Huang, F., Cosatto, E., Graf, H., 2002. Triphone based unit selection for concatenative visual speech synthesis. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, USA, Vol. II, pp. 2037–2040.

Kricos, P., Lesner, S., 1982. Differences in visual intelligibility across talkers. Volta Review 84, 219–225.

Massaro, D., 1998. Perceiving Talking Faces. The MIT Press.

McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. Nature 264, 746–748.

Owens, E., Blazek, B., 1986. Visemes observed by the hearing-impaired and normal-hearing adult viewers. Journal of Speech and Hearing Research 28, 381–393.

Parke, F., 1974. A parametric model for human faces. Unpublished doctoral dissertation, University of Utah, Saltlake City, Utah.

Pelachaud, C., 1991. Communication and coarticualtion in facial animation. Unpublished doctoral dissertation, The Institute for Research in Cognitive Science, University of Pennsylania.

Pighin, F., Hecker, J., Lischinski, D., Szeliski, R., Salesin, D., 1998. Synthesizing realistic facial expressions from photographs. In: Proceedings of the International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH), Orlando, FL, pp. 75–84.

Stork, D.G., Hennecke, M.E. (Eds.), 1996. Speechreading by Humans and Machines: Models, Systems and Applications, Vol. 150. Springer-Verlag, Berlin.

Theobald, B., 2003. Visual speech synthesis using shape and appearance models. Unpublished doctoral dissertation, University of East Anglia, Norwich, UK.

Theobald, B., Bangham, J., Matthews, I., Cawley, G., (2003). 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, pp. 800–803.

Union, I.T., (1974–1978–1982–1986–1990–1992–1994–1995–1998–1998–2000). Methodology for the subjective assessment of the quality of television pictures.

Wakerly, D., Mendenhall, W., Scheaffer, R., 2002. Mathematical statistics with applications. Duxbury Advanced Series.

Waters, K., 1987. A muscle model for animating three-dimensional facial expressions. Proceedings of the International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH) 21 (4), 17–24.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 1999. The HTK Book. Entropic Ltd, Cambridge.