# Animating Faces Using Appearance Models

B. Theobald[1], I. Matthews[2], N. Wilkinson[1], J.F. Cohn[2,3], S. Boker[4]
[1]University of East Anglia, Norwich, UK
[2]Robotics Institute, Carnegie Mellon University, Pittsburgh, USA
[3]Depart of Psychology, University of Pittsburgh, Pittsburgh, USA
[4]Department of Psychology, University of Virginia, Charlottesville, USA

**Abstract**

Here we present an overview of a system we are developing for animating faces using appearance models. To facilitate more widespread use of animated characters, we are interested in developing techniques that: 1) require only easy to use and inexpensive data capture equipment, 2) are three-dimensional allowing virtual characters to be animated, 3) allow a range of face models, from near-photoreal to cartoon-like, to be animated depending on the application, 4) allow visual speech information and facial gestures to be easily transferred between models, 5) can be rendered in real-time, and 6) can generate animated sequences directly from motion captured data, lip-synced to audio, or generated entirely from text.

## 1 Introduction

Realistic animation of human faces is challenging as the changes in the features of the face that we interpret as expressions are the product of a complex interaction between various anatomical layers, which include bone, muscle, subcutaneous fat, and skin. The problem is compounded by the fact that we are all expert at detecting and recognising facial expressions and so are sensitive to even the smallest discrepancies from normal behaviour. The central problem then is how best to approximate the intricacies of the face with sufficient detail such that facial gestures synthesised on a model look realistic?

Traditional facial animation approaches are graphics-based, where points on the surface of the face are represented as vertices in three-dimensions (3D) and the skin approximated by connecting the vertices to form a mesh. These mesh vertices are manipulated using time-varying parameters that influence the mesh geometry either directly, or using a physically-based approach [23]. Directly parameterised animation [19, 22] uses geometric interpolation between a collection of hand-crafted face models, known as morph-targets, where each is meticulously designed to be a faithful representation of a change in a particular aspect of the facial anatomy. The drawbacks of this approach are: 1) the morph-targets are usually designed by hand, which is time-consuming, 2) the morph-targets are designed for a particular mesh topology, so are not readily transferable across models. 3) The morph-targets generally are not independent, so care is required to ensure a valid facial expression results from any given combination of the morphs.

Indirectly parameterised models are designed to approximate the anatomical structure of the face, where animation parameters act on physical models, which in turn update the mesh geometry. A popular approach is Waters' pseudo-muscle model [30], where individual mesh vertices are displaced according to the relative vicinity of nearby muscle functions embedded within a mesh. Improved realism has been achieved by extending this approach to use physically-based methods [16, 26]. The limitations of physically-based animation are: 1) it is relatively computationally expensive compared to directly parameterised animation as the influence on each individual vertex must be computed as a function of each muscle. 2) While anatomical models are not tied to a particular mesh topology they must be manually inserted in the mesh. Incorrectly embedding a muscle will produce unexpected results when the model is animated. 3) To prevent artifacts in the rendered mesh care is required to ensure that discontinuities at the boundaries of the regions of interest of the anatomical models are taken into account.

Image-based synthesis can produce animated sequences with a high degree of both static and dynamic realism. Typically animation is achieved either using a data-driven approach, where frames in an existing video sequence are re-ordered [3, 7], by morphing between static images representing key-frames in a video sequence [10], or by warping images using control parameters generated by trajectory synthesis [2, 9]. The main limitations of image-based animation are: 1) it is relatively computationally expensive compared with graphics-based systems, 2) animation is usually confined to re-animating only the face, and 3) transfer of speech and expression information between subjects is relatively difficult and somewhat constrained compared with graphics-based approaches.

The advantages of both graphics and image-based animation can be exploited by unifying the approaches into a single model. Pighin et al. [24] recover the 3D point locations for a sparse set of features on a face in five different views using photogrammetry. A dense geometric mesh is then fitted to the recovered points and the images from each view blended to create view-dependent texture maps. Repeating the process for a number of expressions allows realistic sequences to be animated by interpolating the geometry and morphing the view-dependent images (across both view and time) [24]. Extending this idea to capture the reflectance field of the face allows for a change in illumination in the animated sequences [12]. These methods animate sequences with a stunning degree of realism, but it is not immediately clear how they can be extended to synthesise visual speech, or how the animation can be mapped to different faces. A more flexible approach is to model the variation in the facial features using Active Appearance Models (AAMs) [6] or 3D morphable models (3DMMs) [1]. The main advantage of the AAM over the 3DMM is computational efficiency: near-photorealistic images are generated in real-time from only a few tens of parameters, but this is usually at a cost of image quality. The 3DMM is a dense 3D description of the face and tends to produce higher quality images than the AAM.

The following sections begin by describing AAMs, which are at the heart of our approach for animating faces. Combining scattered data interpolation, as was used in [24], with 2D+3D AAMs allows us to construct dense three-dimensional face models from only a single video camera. We then go on to describe how the model is used to animate faces from video, voice or text.

## 2 Active Appearance Models: AAMs

The *shape*, $\mathbf{s}$, of an AAM is defined by the concatenation of the $x$ and $y$-coordinates of $n$ vertices that form a 2D triangulated mesh: $\mathbf{s} = (x_1, y_1, \ldots, x_n, y_n)^T$. A compact model that allows a linear variation in the shape is given by,

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{m} \mathbf{s}_i p_i, \tag{1}$$

where the coefficients $p_i$ are the shape parameters. The model is usually computed by applying principal component analysis (PCA) to a set of shapes hand-labelled in a corresponding set of images [6], where the base shape $\mathbf{s}_0$ is the mean shape and the vectors $\mathbf{s}_i$ are the (reshaped) eigenvectors corresponding to the $m$ largest eigenvalues. A recent extension to AAMs [33], the so called 2D+3D AAM, allows a 3D model of shape to be inferred from the change in shape of the 2D AAM using non-rigid structure from motion. See the top and middle rows of Figure 1 for example 2D and 3D shape models.

The *appearance*, $A(\mathbf{x})$, of an AAM is defined by the pixels that lie inside the base mesh, $\mathbf{x} = (x,y)^{\mathrm{T}} \in \mathbf{s}_0$. A compact model that allows a linear variation in the appearance is given by:

$$A(\mathbf{x}) = A_0(\mathbf{x}) + \sum_{i=1}^{l} \lambda_i A_i(\mathbf{x}) \qquad \forall\, \mathbf{x} \in \mathbf{s}_0, \tag{2}$$

where the coefficients $\lambda_i$ are the appearance parameters. The model is usually computed by applying PCA to the (shape-normalised) training images [6], where the base appearance, $A_0$, is the mean shape-normalised image and the vectors $A_i$ are the (reshaped) eigenvectors corresponding to the $l$ largest eigenvalues. An example is shown in the bottom row of Figure 1.

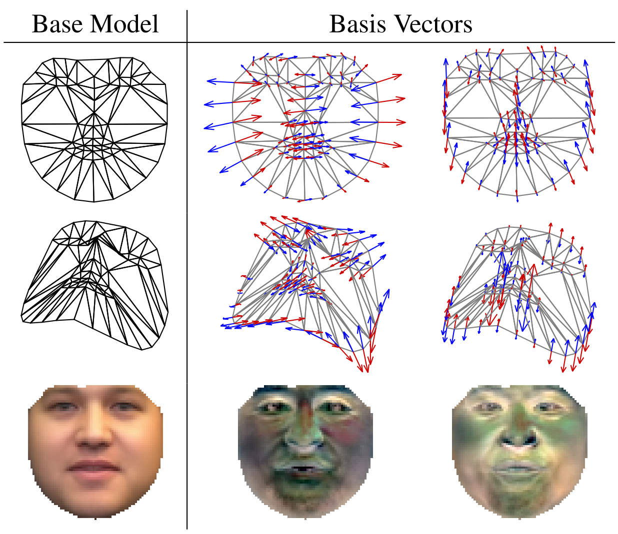



Figure 1: The 2D (top row) and 3D (middle row) shape, and the appearance (bottom row) of an AAM. Shown are the base and the first two basis vectors of the respective models.

## 2.1 Improving the Shape Representation of 2D+3D AAMs

A limitation of the 2D+3D AAM is the vertices that define the shape of an object are relatively sparse, particularly in areas such as the cheeks where there are no defining features that are easy to locate by hand, as illustrated in Figure 1. Consequently these regions are modelled as large planar surfaces, which can give a poor representation of the overall shape — see Figure 3(a). To overcome this we fit a dense generic mesh to the vertices of the shape of the AAM using scattered data interpolation [24]. First, a correspondence is defined between the $n$ vertices in the shape model and the $N >> n$ vertices of the dense mesh. For each of the $n$ constrained vertices in $N$, i.e. those corresponding to a shape model vertex, their displacement is computed by

$$\mathbf{u}_i = \mathbf{x}_i - \mathbf{x}_i^{(0)}, \tag{3}$$

where $\mathbf{x}^{(0)}$ is the dense mesh in the default position and $\mathbf{x}_i$ are the new coordinates for the $i^{th}$ constrained vertex (given by the corresponding vertex coordinates in the AAM shape). A smooth vector-valued function that fits the known displacements, $f(\mathbf{x}_i) = \mathbf{u}_i$ is defined such that the displacements of the remaining (unconstrained) vertices can be found using $f(\mathbf{x}_j) = \mathbf{u}_j$.

A radially symmetric basis function that falls off smoothly with distance is used so the displacement of unconstrained vertices are more influenced by the displacement of constrained vertices lying closer by. The function $f(\mathbf{x})$ used here follows [24]:

$$f(\mathbf{x}) = \sum_i \mathbf{c}_i \phi(\| \mathbf{x} - \mathbf{x}_i \|), \tag{4}$$

where the basis function takes the form $\phi(r) = e^{-r/64}$ and the coefficients $\mathbf{c}$ are found by multiplying the coordinates of $\mathbf{u}_i$ with the matrix $\Phi^{-1}$, where $\Phi_{ik} = \phi(\| \mathbf{p}_i - \mathbf{p}_k \|)$, with $\mathbf{p}_i$ the $i^{th}$ constrained vertex and $\mathbf{p}_k$ the $k^{th}$ constrained vertex.

Fitting the dense mesh to each hand-labelled frame, see Figure 2, and constructing a 2D+3D AAM provides a smoother model of shape, as shown in Figure 3. We are currently investigating methods for automatically constructing dense, person-specific meshes from a single video source to better model the subtleties of the shape of each individual.

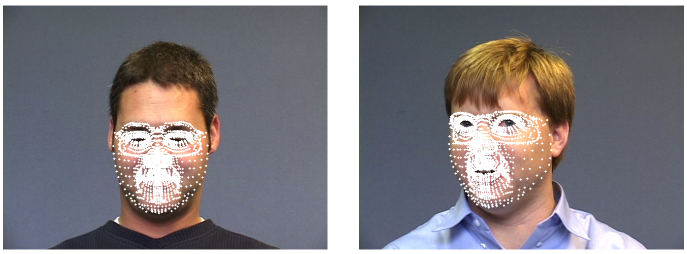
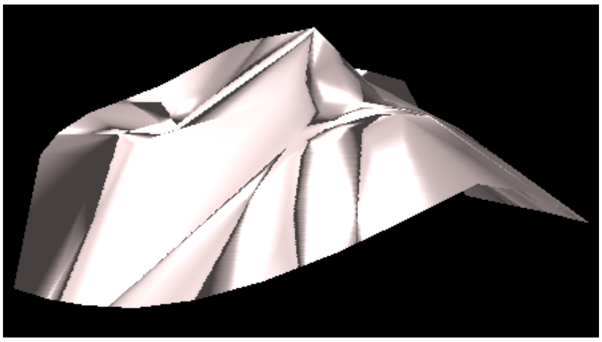


Figure 2: The same (generic) mesh, adapted from [24], matched automatically to two subjects using the position of the sparse 2D vertices of the shape of an AAM. The vertices of this generic mesh may now define the vertices of a dense 2D+3D AAM.



(a)

(b)

Figure 3: The 3D shape of (a) a sparse and (b) a dense 2D+3D AAM. Notice the poor representation of shape in the sparse model in the regions of the face around the cheeks and chin. These regions are modelled as large triangles as there are no defining features that can be represented as vertices in the AAM. The overall shape captured using the dense model is much smoother than that captured by the sparse model.

# 3 Animating Faces using AAMs

A near-photorealistic image is rendered using an AAM by first applying the shape parameters $\mathbf{p} = (p_1, \ldots, p_m)^{\mathrm{T}}$ to the model, Equation (1), to generate the shape, $\mathbf{s}$, of the AAM, then applying the appearance parameters $\lambda = (\lambda_1, \ldots, \lambda_l)^{\mathrm{T}}$ to the model, Equation (2) to generate the AAM image, $A(\mathbf{x})$. Finally, a piece-wise affine warp is used to warp $A(\mathbf{x})$ from $\mathbf{s}_0$ to $\mathbf{s}$. Animating faces using AAMs is then simply a matter of applying the appropriate time-varying parameters to the model. These can be obtained directly from a video source, as in performance-driven animation, from a novel speech signal, or entirely from text.

## 3.1 Performance-Driven Animation

Performance-driven facial animation maps facial gestures (speech and expression) from the face of an actor to either the face of a graphics-based model [5, 11, 20, 32], or a face in an image [4, 17, 35]. This is often preferred to synthesis-based approaches as the sequences generally look more natural. However, methods that generate results close to photo-real are relatively computationally expensive and it is difficult to manipulate facial gestures, such as exaggerate or attenuate the degree of expressiveness. In addition, the result is usually the gesture, as produced by the source identity, transferred directly to the target face, which is unlikely to look the same as the gesture produced by the target identity. This issue was recently addressed in [29], where multi-linear models are constructed from a number of people speaking and displaying pre-specified facial expressions. The model is matched to new faces and expressions cloned on these new faces based on statistics learned during training. In particular the multi-linear model captures the variation due to identity, expression and speech independently. This approach was also approximated in 2D and shown to work with AAMs [18].

We have developed a more convenient solution for expression transfer that does not require a database of speech and expression for a number of individuals, as is required in [18, 29]. Rather, we require only a source AAM and a model for each of the faces we wish to transfer facial gestures to. In particular, we use a mapping that is intuitive given the nature of the vectors that span the shape and appearance space of an AAM. Each component of a shape vector is an offset from the mean shape (resp. appearance) and the vector itself represents the overall displacement that gives rise to a specific type of gesture — see Figure 1. For example, one vector might open/close the mouth and rotate the jaw. If the correspondence between models were one-to-one we could simply apply the parameters for one model directly to the shape/appearance vectors of another (ignoring scale). However, it is extremely unlikely that the vectors will correspond in this way. Indeed it could be that a specific source of variation captured by a *single* basis vector for one model is represented as a *combination* of basis vectors for another model. To map the meaning of the parameters from one model to another we compute the relationship between the basis vectors in the two model-spaces to determine the combination of vectors in the target space that produces the corresponding change in shape (or appearance) when moving along a single vector in the source space. As the basis vectors are unit length and can be constrained to lie in the same dimension Euclidean space when the models are built, the alignment of a source basis vector with the target vector-space is given simply by the inner products $<\mathbf{s}_i^s, \mathbf{s}_j^*>$ between a source vector and each of the target vectors. Thus, a vector (a displacement from the mean) in the source space is a weighted average of the vectors (displacements from the mean) in the target-space, and the weights are obtained from the inner-products. More formally, expressing Equation 1 in matrix form and including the mapping gives:

$$\mathbf{s}^* = \mathbf{s}_0^* + \mathbf{S}^* (\mathbf{R}\mathbf{p}_s), \quad (5)$$

where the columns of $\mathbf{S}^*$ are the basis vectors spanning the **target** space, $\mathbf{R}$ is a $q \times r$ matrix of inner products (the target space is of dimension $q$ and the source of dimension $r$), and $\mathbf{p}_s$ are the parameters representing the expression in the **source** space. Note: $\mathbf{R}$ does not depend on the parameters to be mapped, so can be pre-computed. Therefore, the cost of this mapping is only a matrix-vector product. Notice we do not need to provide semantic labels describing facial expressions, as is required in [18, 29]. This information is implicit from the vectors in the source and target spaces. Also note we are not concerned with the direction of the eigenvectors. For example, an increasingly positive value for a source parameter might, say, open the mouth, while the same action could be defined by an increasingly negative value in the target space. In this instance the inner product for that combination of vectors is negative (the displacements from the mean are largely in opposite directions), so the sign of the parameter value is flipped when the parameter is mapped. Another consideration when mapping parameters is moving too far along the target vectors, which could generate implausible faces. An obvious example is the upper and lower lip boundaries intersecting. However, the mapped parameters can be constrained within the limits of the original (target) training data (typically they must lie within $\pm 3\sigma$ from the mean), which will ensure only valid faces with the target appearance are generated. Example images generated using this cloning are shown in Figure 4.

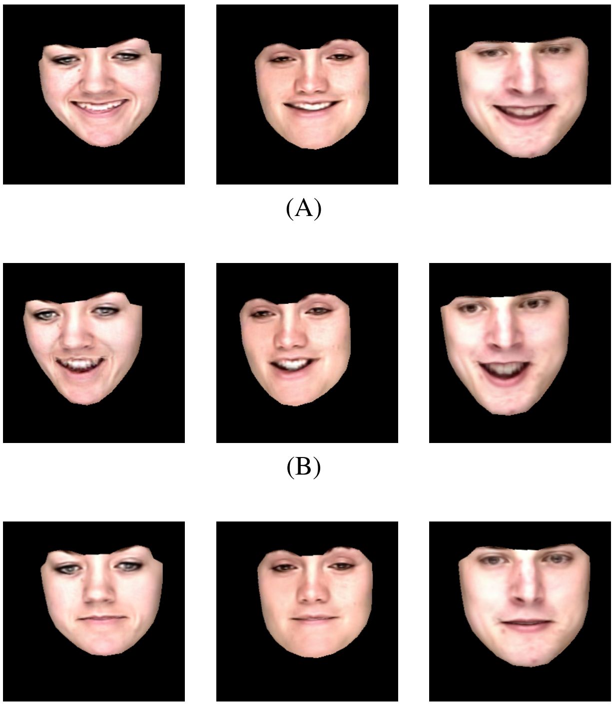



Figure 4: Mapping facial gestures from a source model (left column) to two target faces: one is female (Target (F)), and one is male (Target (M)). The examples show: (A) a smile, (B) a laugh, and (C) a bilabial stop during speech, mapped from the source to the target faces.

Linear mapping in this way will undoubtedly lose information as it is possible that not all of the variation captured by the source model will be described by the target. In practise very little of the variation cannot be mapped since the movements of the facial features are highly constrained. Missing components of each source vector, which are known *a priori*, can be appended to the target model if desired (with suitable orthonormalisation). Performance-driven facial animation using AAMs in this way operates in real-time and preliminary experiments using live face-to-face conversation with our face models suggest this mapping is effective. Our system was used in a videoconference experiment in which participants naive to the existence of the cloning procedure spoke with what they were told were six different people but were actually two confederates cloned onto six different identities. After the conversations the naive participants were asked if "anything was strange" and no one out of the 24 in the experiment guessed that they were talking to fewer than six people. These experiments will be the focus of a future publication.

Performance-driven animation can produce very life-like sequences, but the approach is not particularly flexible. Gestures from an actor are transferred directly to the model, so it is difficult to synthesise unseen sequences. The animation of a new sequence of gestures would first require the gestures to be motion captured. A more flexible solution is to adopt a synthesis-based approach that attempts to create unseen sequences based on some (often limited) training data. The following gives an overview of our visual speech synthesiser, which generates synthesised visual speech from either voice or text.

## 3.2 Speech-Driven Animation

Speech is multi-model in nature — the sounds of speech arise, in part, from the physical movements of the speech articulators. Several studies [2, 8, 21, 14, 15, 25, 31, 34] have investigated animating face models directly from various parameters that encode auditory speech. To animate AAMs from voice we have conducted a series of experiments designed to measure the strength of the relationship between common auditory parameters (formant frequencies, linear prediction coefficients (LPCs), line spectral frequencies (LSFs), and Mel frequency cepstral coefficients (MFCCs)), and the abstract parameters of visual speech provided by the AAM. The advantage of the AAM over models used elsewhere is that: 1) it provides an (almost) complete description of visual speech (in the sense that near-photorealistic images can be reconstructed from the parameters), 2) using AAMs we are able to separate shape and appearance information, and 3) we do not need to specify the parameters in which we are interested (e.g. degree of mouth opening, etc.). The parameters are entirely data-driven and capture the subtleties of speech production.

To measure the strength of the relationship between the auditory and AAM parameters we first recorded a subject reciting approximately 280 sentences and divided the audio signal into frames of 40ms duration (to match the 25Hz frame-rate of the video). The signal in each frame was parameterised as formant frequencies, MFCCs, LPCs and LSFs and the face in each video frame encoded in terms of shape and appearance parameters of the AAM. Canonical correlation analysis (CCA) [13] was used to measure the strength of the relationship between randomly selected frames from the audio and visual data. In our experiments [28] we have found MFCCs and LSFs are best correlated with visual speech, and a stable estimate of the strength of the correlation can be obtained in just a few hundred frames (a few seconds of speech) — see Figure 5. Following this we use MFCCs to animate an AAM from audio using either a linear, or a non-linear mapping.

### 3.2.1 Linear Mapping of Audio to Visual Speech

To synthesise visual speech from novel auditory speech using a linear mapping, we first use CCA to calculate the canonical factor pairs $\mathbf{W}_a$ and $\mathbf{W}_v$ for the auditory and visual

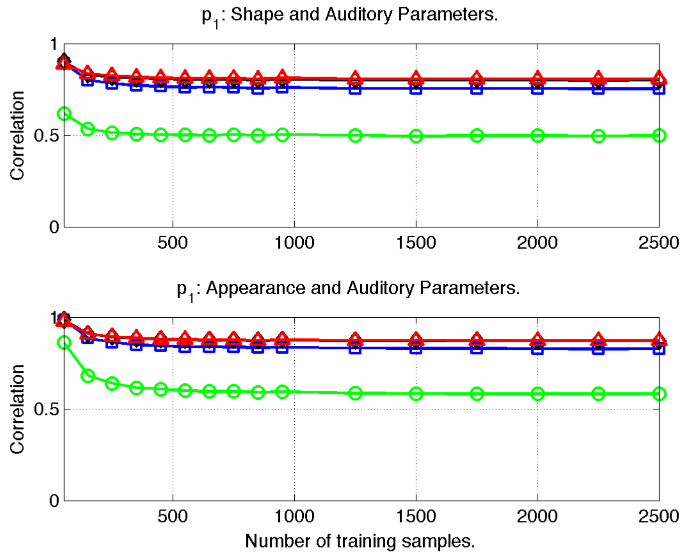



Figure 5: The value of the first canonical correlation coefficient calculated from randomly selected audio-visual feature vector pairs. The auditory parameters are: formant frequencies (green circles), LPCs (blue squares), LSFs (black diamonds), and MFCCs (red triangles). Each point represents the mean value averaged over 100 trails for the specified number of audio-visual feature frames.

parameters respectively. Next the auditory parameters are projected onto $\mathbf{W}_a$:

$$\mathbf{A}_p = \mathbf{W}_a^T \mathbf{A}, \tag{6}$$

(assuming the mean has been removed) and the transformation matrix that maps auditory to visual parameters is found using regression, as follows:

$$\tau = \mathbf{V}\mathbf{A}_p^T \left(\mathbf{A}_p \mathbf{A}_p^T\right)^{-1}. \tag{7}$$

Given a novel speech signal, it is divided in to windows of 40mS duration from which MFCCs are calculated. Next the visual parameters are calculated using the regression matrix in Eq. (7) as follows:

$$\mathbf{v} = \tau \mathbf{W}_a^T \mathbf{a}. \tag{8}$$

where $\mathbf{a}$ are the auditory parameters and $\mathbf{v}$ are the visual parameters. The visual parameters are then applied to the AAM to generate a visual sequence lip-synched to the auditory signal. An example original and synthesised parameter trajectory (for the first shape parameter) are shown in Figure 6. CCA coupled with this linear mapping does a reasonable job of capturing the overall shape of the parameter sequence, but the synthesised visual speech using this approach is under-articulated. We might use an approach like in this situations where training data is very limited, but for better performance in a speech-driven application we usually map from auditory to visual speech using a non-linear mapping, described in the next section.

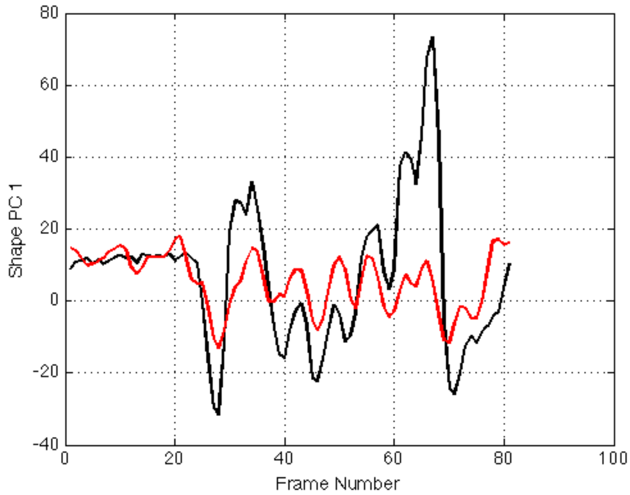



Figure 6: Real (black) and synthesised (red) trajectory for the first parameter in the shape-space of an AAM over an utterance. The synthesised trajectory was generated using a linear mapping from MFCCs projected onto canonical factors and the shape and appearance parameters of an AAM.

#### 3.2.2 Non-linear Mapping of Audio to Visual Speech

To improve upon the linear mapping described previously our system also allows non-linear mapping from auditory parameters to AAM parameters using a neural network. We have systematically evaluated the required amount of training data, the topology of the network, and the influence of context on the performance of the system. Typically we use a three-layer back propagation network with ten units in the hidden layer and 3–5 frames either side of a synthesis frame as contextual information. Mapping in this way requires significantly more training data than the linear mapping described previously, but the articulation strength in resultant synthesised sequences more closely resembles natural speech. The evaluation of this system is the focus of a pending publication, but an example original and synthesised trajectory corresponding to those shown in Figure 6 are shown in Figure 7).

### 3.3 Text-Driven Animation

To generate synthesised visual speech from text we pre-compute the similarity of visual speech gestures (observed during real speech) in terms of the parameters of an AAM, where the similarity measure is designed to take into account the degree to which speech gestures are coarticulated [27]. The similarity scores for our data-driven approach typically reflect viseme groupings observed using human judgement of similarity, where 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 — see Table 1 for examples.

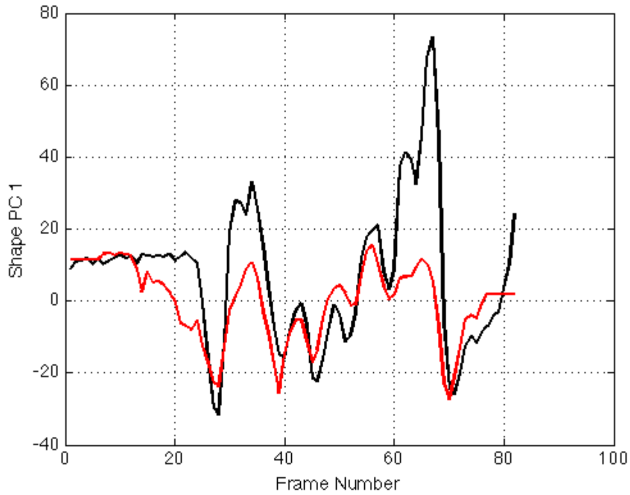



Figure 7: Real (black) and synthesised (red) trajectory for the first parameter in the shape-space of an AAM over an utterance. The synthesised trajectory was generated by mapping MFCCs to AAM parameters using a three-layer back propagation neural net with 10 hidden units.

| 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: Typical phoneme similarity scores, where the range of similarity is 0 (maximally dissimilar), to 1 (identical). Rank 1 is the most similar phoneme with its similarity score, Rank 2 the second most similar and so.

To generate the parameter sequence for a new utterance from text, text-to-speech (TTS) rules are used to map the text to a sequence of phonemes and the original parameters are searched for the $k$ examples of that phoneme in the most similar contexts (surrounding phonemes) based on the visual similarity of the phonemes. These $k$ closest matches for each phoneme are temporally warped to the desired duration and averaged. This is repeated for each phoneme in the utterance and the new trajectories representing these phonemes are concatenated and smoothed to provide a new trajectory in the model-space, which is applied to the model. The parameter trajectories corresponding to those shown for the speech driven approaches in Figures 6 and 7 are shown in Figure 8, and example frames for various mouth shapes in synthesised sequences are shown in Figure 9.

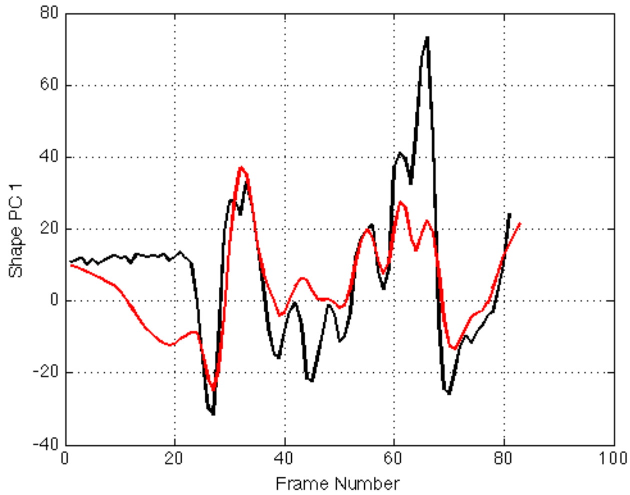



Figure 8: Real (black) and synthesised (red) trajectory for the first parameter in the shape-space of an AAM over an utterance. The synthesised trajectory was generated from a sequence of phoneme symbols using the visual similarity of phonemes measures in terms of AAM parameters.

# 4 Summary and Further Work

We have given a broad overview of our system for modelling and animating faces. At the heart of our system is a dense 2D+3D AAM, which is constructed by matching automatically a dense generic mesh to images hand-labelled with sparse vertices. The advantage is expensive laser scanning equipment, or a complex multi-camera capture environment is not required. The models are constructed from a single, standard camera. We have also described how AAM parameters representing new phrases are generated from either text or voice. In addition we can map the parameters from one model to another and render in real-time, meaning large data-sets for each individual to be animated are not required. The advantage of the text-based approach is faces can be animated by either a TTS engine, or from a voice signal (which is transcribed using a speech recogniser). Since the input is ultimately a sequence of phonemes the model can be lip-synched to any voice and any language. The disadvantage of text-driven synthesis is it does not operate in real-time as the training data must be searched to find the best matching examples.

Further work will investigate the generation of *expressive* speech from text and voice. Currently the synthesised AAM parameter sequences reflect only expressionless speech (i.e. no emotion). For use in interactive environments, the addition of expressive information is paramount. We are also currently investigating speaker-independent parameters for auditory speech, such that our speech-driven system can be animated from any voice. Traditionally this is achieved by first transcribing the utterance using a speech recogniser and animating from the resultant phoneme sequence. However, we are interested in animating the face from any voice directly. We are also formally evaluating the various synthesis approaches using both objective and subjective testing.

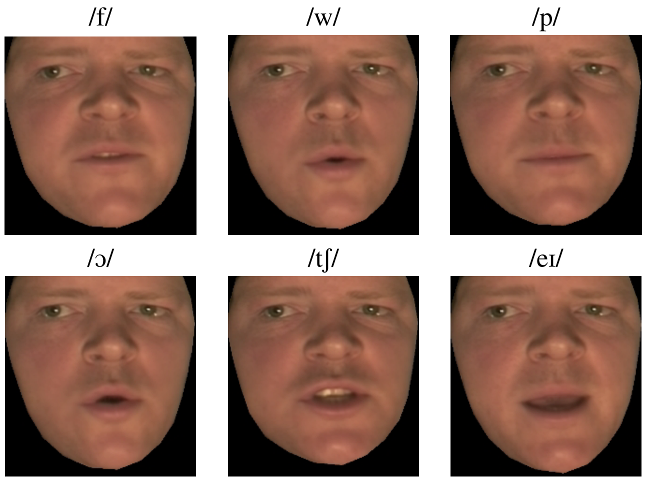



Figure 9: Selected frames from animated sequences synthesised from text illustrating various visual speech poses.

## Acknowledgements

The research described in this paper was supported in part by NSF Grant BCS–0527485, NSF Grant HSD–0(49)527444, and EPSRC Grant EP/D0490751.

## References

[1] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *Proceedings of SIGGRAPH*, pages 187–194, Los Angeles, California, August 1999.

[2] M. Brand. Voice puppetry. In *Proceedings of SIGGRAPH*, pages 21–28, Los Angeles, California, 1999.

[3] C. Bregler, M. Covell, and M. Slaney. Video rewrite: Driving visual speech with audio. In *Proceedings of SIGGRAPH*, pages 353–360, Los Angeles, California, August 1997.

[4] Y. Chang and T. Ezzat. Transferable videorealistic speech animation. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 29–31, Los Angeles, July 2005.

[5] E. Chuang and C. Bregler. Performance driven facial animation using blendshape interpolation. Technical Report CS-TR-2002-02, Stanford University, April 2002.

[6] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, June 2001.

[7] E. Cosatto and H. Graf. Sample-based synthesis of photorealistic talking heads. In *Proceedings of Computer Animation*, pages 103–110, Philadelphia, Pennsylvania, June 1998.

[8] Y. Du and X. Lin. "Realistic mouth synthesis based on shape appearance dependence mapping," *Pattern Recognition Letters*, **23**(14), pp. 1875–1885, December 2002.

[9] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. In *Proceedings of SIGGRAPH*, pages 388–398, San Antonio, Texas, July 2002.

[10] T. Ezzat and T. Poggio. Miketalk: A talking facial display based on morphing visemes. In *Proceedings of the Computer Animation Conference*, pages 96–103, Philadelphia, Pennsylvania, 1998.

[11] B. Guenter, C. Grimm, D. Wood, H. Malvar, and F. Pighin. Making faces. In *Proceedings of SIGGRAPH*, pages 55–66, Orlando, Florida, 1998.

[12] T. Hawkins, A. Wenger, C. Tchou, A. Gardner, F. . Goransson, and P. Debevec. Animatable facial reflectance fields. In *Eurographics Symposium on Rendering*, June 2004.

[13] H. Hotelling, "Relations between two sets of variates," *Biometrika*, bf 8, pp. 321–377, 1936.

[14] C. Hsieh and Y. Chen. "Partial linear regression for speech-driven talking head application," *Signal Processing: Image Communication*, **21**, pp. 1–12, 2006.

[15] T. Kuratate, K. Munhall, P. Rubin, E. Vatikiotis-Bateson, and H. Yehia. "Audio-visual synthesis of talking faces from speech production correlates". In *Proceedings of Eurospeech*, pages 1279–1282, 1999.

[16] Y. Lee, D. Terzopoulos, and K. Waters. Realistic modeling for facial animation. In *Proceedings of SIGGRAPH*, pages 55–62, 1995.

[17] Z. Liu, Y. Shan, and Z. Zhang. Expressive expression mapping with ratio images. In *SIGGRAPH*, pages 271–276, Los Angeles, August 2001.

[18] I. Macedo, E. Vital Brazil, and L. Velho. Expression transfer between photographs through multilinear aam's. In *Brazilian Symposium on Computer Graphics and Image Processing*, pages 239–246, 2006.

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

[20] J. Noh and U. Neumann. Expression cloning. In *SIGGRAPH*, pages 277–288, 2001.

[21] R. Gutierrez-Osuna, P. Kakumanu, A. Esposito, O. Garcia, A. Bojorquez and I. Rudomin. "Speech-driven facial animation with realistic dynamics," *IEEE Transactions on Multimedia*, **7**(1), pp. 33–42, February, 2005.

[22] F. Parke. Parametric models for facial animation. *Computer Graphics and Applications*, 2(9):61–68, 1982.

[23] F. Parke and K. Waters. *Computer Facial Animation*. A K Peters, 1996.

[24] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. Salesin. Synthesizing realistic facial expressions from photographs. In *Proceedings of SIGGRAPH*, pages 75–84, Orlando, Florida, 1998.

[25] M.E. Sargin, E. Erzin, Y. Yemez, A.M. Tekalp. "Lip feature extraction based on audio visual correlation". In *Proceedings of European Signal Processing Conference*, 2005.

[26] D. Terzopoulos and K. Waters. Physically-based facial modelling, analysis and animation. *Journal of Visualization and Computer Animation*, 1(2):73–80, 1990.

[27] B. Theobald, J. Bangham, I. Matthews, and G. Cawley. Near-videorealistic synthetic talking faces: Implementation and evaluation. *Speech Communication*, 44:127–140, 2004.

[28] B. Theobald, and N. Wilkinson. Real-time visual speech synthesis using Active Appearance Models. To Appear in *Proceedings of Auditory Visual Speech Processing*, 2007.

[29] D. Vlasic, M. Brand, H. Pfister, and J. Popovic. Face transfer with multilinear models. *ACM Transactions on Graphcs*, 24(3):426–433, 2005.

[30] K. Waters. A muscle model for animating three-dimensional facial expressions. In *Proceedings of SIGGRAPH*, pages 17–24, 1987.

[31] Z. Wen, P. Hong and T. Huang. "Real time speech driven facial animation using formant analysis," In *Proceedings of the International Conference on Multimedia and Expo*, pages 817–820, 2001.

[32] L. Williams. Performance driven facial animation. *Computer Graphics*, 24(2):235–242, 1990.

[33] J. Xiao, S. Baker, I. Matthews, and T. Kanade. Real-Time Combined 2D+3D Active Appearance Models. In *Proceedings of Computer Vision and Pattern Recognition*, pages 535–542, 2004.

[34] H. Yehia, P. Rubin and E. Vatikiotis-Bateson. "Quantitative association of vocal-tract and facial behaviour," *Speech Communication*, **26**, pp. 23–43, 1998.

[35] Q. Zhang, Z. Liu, G. Quo, D. Terzopoulos, and H. Shum. Geometry-driven photorealistic facial expression synthesis. *IEEE Transactions on Visualization and Compuer Graphics*, 12(1):48–60, January/February 2006.