*Proceedings of Mirage 2003, INRIA Rocquencourt, France, March, 10-11 2003*

# VIDEO REALISTIC TALKING HEADS USING HIERARCHICAL NON-LINEAR SPEECH-APPEARANCE MODELS

*Darren Cosker, David Marshall, Paul Rosin and Yulia Hicks*

{D.P.Cosker, Dave.Marshall, Paul.Rosin, Y.A.Hicks}@cs.cf.ac.uk
Cardiff University, Dept. of Computer Science, Queens Buildings,
Newport Road, PO Box 916, Cardiff CF24 3XF, UK

## ABSTRACT

In this paper we present an audio driven system capable of video-realistic synthesis of a speaker uttering novel phrases. The audio input signal requires no phonetic labelling and is speaker independent. The system requires only a small training set of video and produces fully co-articulated realistic facial synthesis. Natural mouth and face dynamics are learned in training to allow new facial poses, unseen in the training video, to be rendered. To improve specificity and synthesis quality the appearance of a speaker's mouth and face are modelled separately and combined to produce the final video. To achieve this we have developed a novel approach which utilizes a hierarchical and non-linear PCA model which couples speech and appearance.

The model is highly compact making it suitable for a wide range of real-time applications in multimedia and telecommunications using standard hardware.

## 1. INTRODUCTION

Realistic computer generated facial animation is a very difficult task. Mouth animation provides perhaps the greatest challenge since the process of speaking involves the sophisticated cooperation of dozens of muscles in the face, mouth and neck [1] and realistic animations must include movement of the tongue and visibility of the teeth.

Classical facial animation methods, first used for cartoon animation and still used in a similar form today in many cutting edge films, are usually tedious processes requiring an animator to examine a speech track, note frame times of significant speech events and produce key frames at these events.

What is desirable is a facial animation system capable of producing animation solely from audio incorporating both realistic facial synthesis and a means of identifying significant audio events and choosing appropriate facial postures.

In this paper we present a low bandwidth image based system capable of producing fully co-articulated video-realistic facial animation from an audio sound track. The system *learns* the facial dynamics of a speaker and uses this as a foundation to synthesize novel facial animations. For training, the process requires only a small corpus of audio and video from a speaker uttering a set of words. These words are chosen to target a set of visually distinguishable speech postures (Viseme). However, no phonetic annotation needs to be applied to the audio before facial synthesis. Once training has been completed new speech can be supplied in the form of an audio input and synchronized video realistic facial animation is produced. After training the system can be applied to synthesize video with previously unheard speakers. The final video is of the person used in the training phase.

During the training phase a human operator must prepare the video data by placing landmarks at key facial features. Thereafter the system is purely data driven and relies on the assumption that our mouth and face appearance data has a strong correlation with our speech data.

To achieve facial synthesis and animation we introduce a hierarchical non-linear speech-appearance model built from data extracted from the training set. The face is decomposed into parts to form a hierarchy where the root corresponds to a non-linear model of the whole face and sub-nodes non-linearly model smaller, more specific facial areas. This structure allows us to better represent small facial variations and learn more precisely their relationship with speech. For the purpose of this paper we only extend the hierarchy to include the face and mouth.

## 2. BACKGROUND

Modern facial animation systems may be classified into one of two categories: model based and image based systems. Model based methods typically consist of 3D polygonal facial models controlled by streams of input parameters. The first parameter driven 3D facial animation system was by Parke in the 1970's [2]. Parke defined a set of parameters that account for observable variation in the face and used these parameters to animate a 3D mesh model. Since then parameterized model based systems have increased in popularity and complexity. Some employ concepts such as Action Units (AU's) [3] [4] and Animation Units [5] to control facial configurations while others attempt to model facial anatomy in more detail by

modelling the tongue and jaw [6] or the facial muscles [7].

When animating model based systems from speech a mapping may be learned between the parameters required for animation and a phoneme stream. Kalberer et al [8] capture facial poses associated with visemes, define a one-to-one mapping between visemes and phonemes and describe a viseme morphing method. Given a phoneme stream a set of visemes is produced which are warped between to provide smooth animation. Text-to-Speech (TTS) systems capable of producing phoneme streams given a typed input are popular devices for animating such systems [9].

One issue with many model based facial animation systems, except for the most sophisticated [6], is their inability to realistically model the teeth and tongue. If a mapping between speech and the parameters which drive a 3D facial model is derived, as is popular in many phoneme driven parameterized models [8] [10], then essentially it is only a correlation between the outer shape of a face and the speech that is learned. Information such as the position of the tongue and the visibility of the teeth, which provide helpful features in categorising phonemes and strong visual cues in activities such as lip-reading, are discarded. Image based systems are able to learn the *shape* of a face (e.g. the outline of the outer and inner lips) along with its *appearance* (i.e. are teeth visible?) thus providing a richer set of features with which to train a classification based system.

Image based synthesis systems have applications in tracking [11], face identification [12] , behavioural modelling [13] and animation [14] [15] [16]. An Appearance Model [17] is one such image based system. A single vector of parameters is used as the input to a joint statistical model of shape and texture variation, the outputs of which define the texture of an object and the shape that it is warped to. The drawback of appearance models is that collecting training data can be an arduous task. Each image in the training set must be labelled with landmarks defining features of interest. These landmarks form the shape data for an image. Ezzat et al [14] describe another model of image appearance which they call a Multidimensional Morphable Model (MMM). A MMM is defined using optical flow parameters between a reference image and a set of prototype images that define appearance. Optical flow is used as an alternative to land-marking images as it can be done automatically [18].

Control and animation of image based model parameters using a speech signal may come from a TTS system[10] [14] or from a processed audio input [19] [16] [14]. In [14] the audio signal must first be phonetically aligned. The phoneme stream is then mapped to a trajectory of MMM parameters which are used to synthesize animation in MMM space. The draw back of this system is that the audio must be phonetically analyzed and aligned before synthesis. Also, since it is simply a set of standardized phonetic information that is supplied to the system, information in the speech, such as intonation and emotional content, is ignored. This restricts the application in its ability to deliver novel facial animation in areas apart from those chosen for synthesis. This is further emphasized by the fact that it is only the bottom of the face, from the bottom of the nose to the jaw line, that is modelled. A post-processing step re-attaches the computed area to a segment of the original video. This essentially means that synthesis of an emotional phrase would be computed without facial emotion, unless there existed a part of the original training video where the speaker conveys similar emotion for a similar duration, in which case this would have to be manually identified and attached to the synthesized region.

In Voice Puppetry [19] no phonetic alignment is necessary. A Hidden Markov Model (HMM) is trained using frames from a large video data set. The occupancy matrices of each state in the HMM are then used to determine the audio features of each facial state. Given an audio input a facial state sequence is then estimated and a trajectory of facial configurations is produced using an inverse Viterbi algorithm. This system provides more flexibilty than in [14] since the audio is not in phoneme form, and retains artefacts in the speech not directly associated with mouth animation making it capable of synthesising facial expression. However, the system is restricted in that the animations produced are not video-realistic since it is the mouth and face *shape* which is synthesized and not its appearance.

Our system provides the audio-based flexibilty of Voice Puppetry with the video-realism of Poggio, Gieger and Ezzats MMM based system. Like Poggio et al we construct a model of facial shape and appearance in a high dimensional space based on an appearance model. We choose to use an appearance model because of its dimensional reduction of facial parameters and the low bandwidth video realism obtainable. We associate image parameters with speech parameters from our training set to construct a hierarchical structure of speech-appearance models which we use for synthesis. This synthesis model is then capable of handling speech inputs without any alignment, making it suitable for real time or network applications. The hierarchical structure of our model produces significantly improved facial animation compared with similar non-hierarchical systems, which we show in our results.

The next section gives an overview of the process. Section 4 describes the training process and Sections 5 to 7 describe how our hierarchical facial model is constructed and utilized. Sections 8 to 11 present test cases and discus-

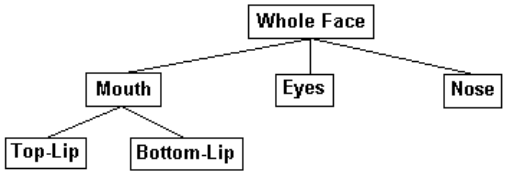


Figure 1: *Hierarchical facial model overview.*

sions based on the results as well as our plans for further work and conclusions.

## 3. SYSTEM OVERVIEW

The system can be broken into four stages: Training, Model Learning, Facial Synthesis and Video Production. In the training phase a video is captured of speaker uttering a list of words targeting different visemes. A human operator then annotates the training set placing landmarks at the boundaries of facial features. This process is made semi-automatic with the use of an Active Shape Model (ASM) [20]. The system then extracts the landmarks from the training set and builds a hierarchical model of the face.

Figure 1 gives a diagrammatic overview of our hierarchical face model. For the purpose of this paper we extend the hierarchy to only include the overall face and mouth nodes. The root node is built from normalized data for the entire face, which when modelled captures global facial variation. Sub-sets of this data are used to build nodes for the next level in the hierarchy which capture in greater detail variations of facial sub-parts. Full and sub-facial appearance is represented using appearance model parameters while speech is processed using Mel-Cepstral analysis [21] before being dimensionally reduced using Principle Component Analysis (PCA). At each node we build a high dimensional speech-appearance manifold approximated non-linearly with a number of Gaussian probability density functions capable of synthesising texture and shape from speech.

During the facial synthesis stage the incoming audio is processed every 40ms. Mel-Cepstral analysis is then applied and the signal is projected into its reduced dimensional form. For each node in the hierarchical model we then do the following. The input signal is classified into a cluster based on the smallest Mahalanobis distance to the centre of each cluster and appearance parameters are calculated by exploiting the locally linear relationship between speech and appearance in this cluster.

In the final stage synthesized facial information from sub-nodes is combined to construct an entire face.

## 4. TRAINING AND PRE-PROCESSING

The training process requires the capture of approximately 40 seconds of video with which to build our hierarchical model and create associations between appearance parameters and audio parameters. We recorded a speaker uttering a list of words that target different visually distinct speech postures. These words were chosen based on the work of Nitchie [22] and are summarized in Table 1. It should be noted that for training any substantially long viseme rich phrase would also suffice.

**Table 1: Training vocabulary.**

| Word Spoken | Mouth Posture Name |
|---|---|
| “We” | Narrow |
| “Fit”, “Heavy” | Lip to teeth |
| “Mock”, “Bin”, “Spark”, “Pin” | Lips shut |
| “Moon” | Puckered lips |
| “We”, “Wharf”, “Good” | Small oval |
| “Farm” | Widely open |
| “Up”, “Upon”, “Fur” | Elliptical |
| “Free” | Elliptical puckered |
| “Pit”, “Believe”, “Youth” | Relaxed narrow |
| “Thin”, “Then” | Front tongue |
| “Time”, “Dime”, “Nine” | Tongue to gum |
| “Leaf” | Tip to gum |
| “All”, “Form” | Spherical triangle |
| “See”, “Zebra” | Tightened narrow |
| “Mat” | Large oval |
| “Ship”, “Measure”, “Chip”, “Jam”, “Gentle” | Protrusion |
| “Let”, “Care” | Medium oval |
| “Keep”, “Go”, “Rank”, “Rang” | Undefined open |

The audio was captured at 33KHz Mono and the video at 25 fps. The video was converted into images with a resolution 720 x 576 and images where words were spoken were extracted. For hand-marking the images we trained and used an ASM.

Images were annotated with 82 landmarks between the top of the eye-brows and the jaw. Figure 2 shows one of the training images labelled with landmarks. Once the whole training set was labelled we extracted the 18 landmarks on the mouth for building the mouth node of our hierarchical model giving us two sets of shape vectors for the training set.

For each set of shape vectors we then do the following: We first align the vector set with respect to scale, rotation and translation. Each image in the training set is then warped to the mean of the aligned vector set using

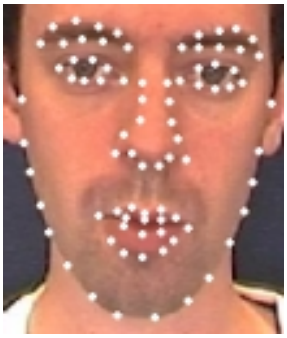

Figure 2: *Annotated training image.*

a piece-wise triangular warping method. This process removes shape variation from the textures, the resulting images are referred to as shape-free patches [17]. In a shape-free patch we are only interested in the texture contained within the convex hull of a set of landmarks. This allows us to model the mouth separately from the face without including any unwanted texture information. The texture from each patch is then normalized to have a mean of zero and a variance of 1. Finally, Mel-Cepstral analysis is applied to the corresponding audio data associated with each video frame. Since we recorded our video data at 25 fps this meant that we processed our audio with Mel-Cepstral analysis every 40ms.

The process results in two normalized training sets for building the mouth and full-face nodes of our hierarchy where each set contains shape, texture and associated speech vectors.

In the following two sections we describe how we construct and use our synthesis models and how we have improved the specificity and generalisation of our models by hierarchicaly modelling the face and mouth. In Section 7 we then describe the post-processing stage where the outputs of both node-models are re-combined.

## 5. FACIAL MODELLING

The facial synthesis method used in our system is based on Cootes et al's Appearance Model [17]. In the following section we describe how we begin constructing a node in our hierarchical model by first building an appearance model of that nodes facial area. We then describe how this appearance manifold is non-linearly approximated and coupled with speech in a speech-appearance model.

### 5.1. Facial Appearance

Firstly, a PCA is applied separately to the shape and texture vector sets to define two linear models for shape and texture respectively

Figure 3: *Highest mode of face appearance variation at + (left) and - (right) 2 s.d. from the mean.*

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}_s \mathbf{b}_s \quad (1)$$

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g \mathbf{b}_g \quad (2)$$

where $\mathbf{x}$ and $\mathbf{g}$ are examples of shape and texture, $\bar{\mathbf{x}}$ and $\bar{\mathbf{g}}$ are the mean normalized shape and texture vectors, $\mathbf{P}_s$ and $\mathbf{P}_g$ are the eigenvectors of each training sample distribution and $\mathbf{b}_s$ and $\mathbf{b}_g$ are shape and texture parameters.

Each training example is then dimensionally reduced through its respective model and represented by the parameter $\mathbf{b}_s$ for a shape vector and $\mathbf{b}_g$ for a texture vector. For our mouth and full face models we reduce the dimensionality of the shape and texture vectors by retaining 99% of their variation. We keep this high percentage of variation to maximize the quality of the final video. We estimate a global scaling for the shape parameters based on the ratio of variances for the shape and grey level eigenmodels and use this to scale the shape parameters with respect to the texture parameters. Corresponding $\mathbf{b}_s$ and $\mathbf{b}_g$ parameters are then concatenated to form vectors $\mathbf{b}$ and a PCA is performed on this new vector set defining a new model of joint shape and texture variation. This new PCA model is called an appearance model where shape and texture are expressed linearly as functions of $\mathbf{c}$. We use the notation:

$$\mathbf{b} = \mathbf{Q}\mathbf{c} \quad (3)$$

to describe our joint model, where $\mathbf{Q}$ are the eigenvectors of our joint distribution and we represent examples of shape and texture as

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}_s \mathbf{W}_s^{-1} \mathbf{Q}_s \mathbf{c} \quad (4)$$

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g \mathbf{Q}_g \mathbf{c} \quad (5)$$

where $\mathbf{W}_s$ is our scale matrix and $\mathbf{Q}_s$ and $\mathbf{Q}_g$ are the shape and texture parts of the eigenvectors $\mathbf{Q}$.

Figure 3 shows the highest mode of full-face appearance variation at $\pm 2$ s.d. from the mean. Figure 4 shows our mouth training set represented by the two highest modes of appearance variation $c_1$ and $c_2$.

### 5.2. Improved Generalization and Specificity

In [19] Brand noted that his results were improved when the lower and the upper face were modelled seperately. In [23] they built a tracking system based on a hierarchical model of the body where each limb was a node and each node represented a separate eigenspace. The top-level node therefore captured major variations in the entire body while nodes at lower levels captured varitaions in the limbs and the torso.
This hierarchical structure of eigenspaces was found to be advantageous since dimensionality could be further reduced and important minor variations in the limbs, which may have been to small to accurately captured with a single eigenmodel, were better modelled. We have found that the same is true in modelling faces. Minor variations in the mouth, which are important communicators when speaking, are better modelled using a hierarchical model.
For these reasons we build separate models for the face and the mouth. We have also found this system to be better at modelling correlations between speech and the mouth since redundant data in the rest of the face, which often remains motionless over a many number of frames, does not bias our model. Our results in Section 8 give strong evidence to suggest this.

### 5.3. Non-Linear Hierarchical Modelling

Linear PDM models are incapable of properly generalising data sets which are non-linear in distribution [24]. In particular it is the specificity of the model that is most affected, with the generation of illegal training examples made possible. In these cases it has been shown that a mixture of Gaussians [25] or a non-linear PCA model [26] can better generalize and improve the specificity of the model. In modelling our mouth texture distribution we require that as much variation as possible is captured if we are to produce convincing animation. From Figure 4 it is seen that our appearance model distribution is highly non-linear. We therefore decided to model it using a mixture of Gaussians.
We first used (3) to project our training set into appearance parameters **c**, and then used a k-means clustering algorithm to initialise a set of centres for an Expectation Maximization (EM) algorithm. We found that our model was quite sensitive to the initial choice of cluster centres and the number of centres used. We therefore carried out a set of experiments with which to find the most stable number of centres and initial centre positions. We began with a low number of centres and calculated the lowest error rating from a series of 20 initial centre positions. We then increased the number of centres and repeated the experiment. We found that the error rating degraded gracefully and began to smooth out between 60 and 70 centres. We therefore chose to use 60 centres with their initial position where the EM algorithm's error output was lowest.

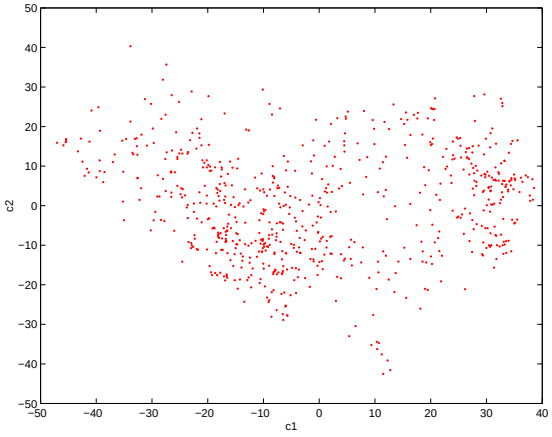


Figure 4: *Distribution of mouth appearance parameters represented by the two highest modes of variation.*

### 5.4. Coupling Speech to Appearance

Our aim is to be able to create an association between our speech vectors and our appearance parameters such that given a speech vector we may estimate its associated appearance parameter. We therefore define a relationship between speech and appearance in each cluster of our model and use a simple classification scheme to identify which cluster to utilize given new speech parameters.
We first perform a PCA on the training set of speech vectors to define the linear model

$$\mathbf{a} = \bar{\mathbf{a}} + \mathbf{P}_a \mathbf{s} \tag{6}$$

where $\mathbf{a}$ is a speech vector, $\bar{\mathbf{a}}$ is the mean speech vector in our training set, $\mathbf{P}_a$ are the eigenvectors of our speech distribution and $\mathbf{s}$ is a speech parameter. We then reduce the dimensionality of each speech vector using

$$\mathbf{s} = \mathbf{P}_a^T (\mathbf{a} - \bar{\mathbf{a}}) \tag{7}$$

and retain 99% of the variation over the initial training set. To ensure that one set of parameters does not bias our model we scale the appearance parameters based on the ratio of the sums of the variances of the speech and appearance models. We then take each appearance parameter and concatenate it with its associated speech parameter giving

$$\mathbf{M}_j = [\mathbf{W}_c \mathbf{c}_j^T, \mathbf{s}_j^T]^T \qquad j = 1, \ldots, k \tag{8}$$

where $\mathbf{M}_j$ is the concatenation of appearance parameter $\mathbf{c}_j$ and speech parameter $\mathbf{s}_j$, $\mathbf{W}_c$ is our scale matrix and $k$ is the number of images in our training set. This gives

us *n* clusters of vectors **M**. We then perform a PCA on each cluster of $\mathbf{M}_j$ parameters to give us *n* joint models of appearance and speech

$$\mathbf{M} = \bar{\mathbf{M}}_i + \mathbf{R}_i\mathbf{d} \qquad i = 1, \ldots, n \tag{9}$$

where $\bar{\mathbf{M}}_i$ is the mean of cluster *i*, $\mathbf{R}_i$ are the eigenvectors of cluster *i* and **d** is a speech-appearance parameter constrained to be within $\pm 3$ s.d. from the mean of cluster *i*. Our non-linear model therefore gives us *n* sets of basis vectors with which to represent facial configurations in appearance space. Hence, an animation sequence is a trajectory through this space defined at each video frame by the basis vectors of a particular cluster.

## 6. MOUTH SYNTHESIS FROM SPEECH

Using our non-linear speech-appearance model we would like to calculate the associated appearance parameter of an input speech parameter for every video frame. This is done in two stages. First, we classify a speech parameter into a speech-appearance cluster based on the smallest Mahalanobis distance to the centre of each cluster

$$D = (\mathbf{s}_{input} - \bar{\mathbf{s}}_i)^T \Sigma (\mathbf{s}_{input} - \bar{\mathbf{s}}_i) \tag{10}$$

where $\bar{\mathbf{s}}_i$ is the mean speech parameter in cluster *i* and $\Sigma$ is the covariance matrix of the speech parameter training set. We then use the chosen cluster to map from our speech parameter $\mathbf{s}_{input}$ to an appearance parameter **c**.
The mapping process we used was first described by Bowden in [24]. Given two sets of strongly correlated variables we can build a statistical model which encodes their relationship allowing us to estimate the value of one from a previously unobserved value of another. This is based on a linear relation between the variables in a data cluster. Bowden extends the idea to a non-linear model where a relationship between variables is formed linearly in each cluster. We adopt the same approach here. We model speech and appearance non-linearly, assume that their relationship is locally linear and use the PDM model for the chosen cluster to estimate an appearance parameter from a speech parameter.
Given a cluster *i* we split its matrix of eigenvectors $\mathbf{R}_i$ into two parts where the top part corresponds to appearance and the bottom part to speech. We then denote the linear relationship between speech and appearance in each cluster as

$$\mathbf{W}_c\mathbf{c} = \bar{\mathbf{c}}_i + \mathbf{R}_{c,i}\mathbf{d} \tag{11}$$

$$\mathbf{s} = \bar{\mathbf{s}}_i + \mathbf{R}_{s,i}\mathbf{d} \tag{12}$$

where $\bar{\mathbf{c}}_i$ and $\bar{\mathbf{s}}_i$ are the mean appearance and speech parameters of cluster *i* and $\mathbf{R}_{c,i}$ and $\mathbf{R}_{s,i}$ are those parts of the eigenvectors of $\mathbf{R}_i$ associated with appearance and speech respectively.

Given a parameter $\mathbf{s}_{input}$ we would therefore like to calculate the associated parameter **d** and use this to calculate **c**. We calculate **d** as follows

$$\mathbf{d} = \mathbf{R}_{s,i}^T(\mathbf{s}_{input} - \bar{\mathbf{s}}_i) \tag{13}$$

and then use **d** in (11) to calculate the appearance parameter **c**. As a final step we then constrain **c** to be within $\pm 3$ s.d's from the mean of its respective cluster. Using our **c** parameter trajectory we calculate shape **x** and texture **g** using (4) and (5).

### 6.1. Trajectory Smoothing

The synthesis technique described is carried out every 40ms for a 25 fps video. We therefore rely on a steady audio signal with which to calculate steady appearance parameter trajectories. Since appearance estimation is based on cluster choice it is important the the correct cluster choice is made for every video frame. However, since our synthesis method is not based on learned cluster routes this is not always guaranteed to happen. The outcome of this is that unwanted mouth configurations and *noisy* shape vectors may appear. To compensate for this we perform a local smoothing of the calculated shape and texture vectors

$$\mathbf{x}_i = (\mathbf{x}_{i-1} + \mathbf{x}_i + \mathbf{x}_{i+1})/3 \tag{14}$$

$$\mathbf{g}_i = (\mathbf{g}_{i-1} + \mathbf{g}_i + \mathbf{g}_{i+1})/3 \tag{15}$$

Smoothing operations in facial animation systems are commonly used [8]. We also tried the same smoothing operations on the **c** parameter trajectory but found that the overall quality of the videos produced was significantly reduced.

## 7. RECONSTRUCTING THE HIERARCHY

This stage is concerned with reconstruction of a face image by combining the textures calculated at each node in the hierarchy. This is done in a top-down fashion and exploits the fact that the mean shape of the mouth node model and the mean shape of the mouth in the face node model is the same. We first construct the face texture from the root node corresponding to the current speech parameter which is shape free but contains the relevant texture information and extract from it the mouth texture. We then construct the mouth texture from the mouth node model and scale its values to lie within the range of the mouth texture from the root node.
We then substitute the mouth texture and shape from our root node with the mouth texture and shape from the mouth node. This composite shape-free patch is then warped according to the new composite shape coordinates.

Figure 5: *Synthesized facial reconstructions for the word 'Dime' compared with ground-truth images.*

Figure 6: *Synthesized facial reconstructions for the word 'Youth' compared with ground-truth images.*

## 8. EVALUATION

We recorded 28 seconds of video of a speaker uttering the words listed in Table 1. We then extracted those frames associated with spoken words and labelled them using our land-marking tool giving us 968 video frames. We used 706 of these frames, along with each frames corresponding speech vector, to build a linear speech-appearance model for the whole face, a non-linear speech-appearance model for the whole face and a hierarchical speech-appearance model with nodes for the whole face and the mouth. We then synthesized three video-reconstructions using the audio associated with the 262 frames we left out of training. This gave us a set of ground truth images with which to compare the synthesized frames from our models.

Figures 5 and 6 show synthesized facial reconstructions produced from our models compared with their ground truth images. The top rows show the training images while the second, third and forth rows show reconstructions from the hierarchical, non-linear and linear models respectively (the linear model results are ommitted from Figure 6. Downloadable movies showing our results may be found at *http://www.cs.cf.ac.uk/user/D.P.Cosker/research.html*.

From our tests we found that the worst reconstructions came from the linear model. Videos reconstructed with this model show little or no variation from their mean image. This is most likely caused by the linear models inability to properly represent the non-linear appearance manifold. Also, the small variations that do occur are almost invisible due to the smoothing operation performed. The non-linear model shows improved lip-synch with the ground truth video but often displays wrong textures, distrupting the flow of the animation. The hierarchical model gives the best results with near-perfect lip-synch, faithful texture reconstructions and strong coarticulation.

We also recorded a number of speakers uttering sets of words not contained in the training set and synthesized video-reconstructions. We found that the hierarchical model produced high-quality lip-synched animations to the new speech segments. As in our other tests we also found that animation quality reduced with the non-linear and linear models. We discovered that accent played an important part in the quality of the synthesized videos. In the future we plan further analysis utilizing greater speech descriptors which will lead to greater discrimination of speakers and emotive aspects of speech.

## 9. DISCUSSION

The strong performance of the hierarchical model is mostly attributed to the non-linear clustering that is performed for each node in the hierarchy. When clustering a non-linear model for the whole face the centre choices do not model well the variation in the mouth, since data in the rest of the face biases decisions. In the hierarchical model the mouth is better represented since it is modelled without extraneous facial information. The hierarchical model therefore better models the associations between the speech training set and mouth variation.

Reconstructions using speakers that the system is not familiar with were significantly better when the speaker had the same accent as the person the system had been trained on.

## 10. FUTURE WORK

It is our intention in future work to model the face in greater detail. This may include a more detailed model of the mouth, eye and eyebrow models and models for the forehead and the chin and jaw. A more detailed model should also better capture emotion present in the face during speaking. In later work we intend to train a new model using a variety of speakers uttering phrases with emotional

emphasis in a hope that we may learn a mapping between the nodes of our model and emotion inherent in speech. To achieve this we will have to introduce other forms of speech analysis to allow us to model speech intensity and pitch, for example.

## 11. CONCLUSIONS

We have introduced a non-linear hierarchical speech-appearance model of the face capable of producing high-quality video-realistic animation given a speech input. The model is purely data driven and the audio requires no phonetic-alignment before video-production. We have shown that the hierarchical model better captures variations in sub-facial areas such as the mouth than non-linear and linear models of the entire face. We have also shown that the model is capable of synthesizing video from previously unheard speakers.

## 12. REFERENCES

[1] F. I. Parke and K. Waters, "Computer Facial Animation," A. K. Peters, Ltd. ,1996.

[2] F. I. Parke, "A parametric model for human faces," ,PhD Thesis, University of Utah, Department of Computer Science, 1974.

[3] M. Rydfalk, "CANDIDE, a parameterized face,". Technical Report No. LiTH-ISY-I-0866, Dept. of Electrical Engineering, Linkping University, Sweden, 1987.

[4] J. Ahlberg, "CANDIDE-3, an updated parameterized face," Technical Report No. LiTH-ISY-R-2326, Dept. of Electrical Engineering, Linkping University, Sweden, 2001.

[5] MPEG Working Group on Visual, International Standard on Coding of Audio-Visual Objects, Part2 (Visual), ISO-14496-2, 1999.

[6] L. Reveret, G. Bailly and P. Badin, "MOTHER: A new generation of talking heads for providing a flexible articulatory control for video-realistic speech animation," Proc. of the 6th Int. Conference of Spoken Language Processing, ICSLP'2000, Beijing, China, pp. 755-758, Oct. 16-20, 2000.

[7] K. Waters, "A muscle model for animating 3D facial expression," ACM Computer Graphics, Vol. 21, No 4, pp. 17-24, 1987.

[8] G. A. Kalberer and L. Van Gool, "Face animation based on observed 3D speech dynamics," Procs. of The Fourteenth IEEE Conf. on Computer Animation, 2001.

[9] B. Le Goff, C. Benoit, "A text-to-audiovisual-speech synthesizer for french," Proc. of ICSLP96, 1996.

[10] T. Ezzat and T. Poggio, "MikeTalk: A talking facial display based on morphing visemes," In Proc. of Computer Animation Conference, 1998.

[11] G. J. Edwards. C. J. Taylor and T. F. Cootes, "Learning to Identify and Track Faces in Image Sequences," Procs. of British Machine Vision Conference, 1997.

[12] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," Journal of Cognitive Neuroscience, Vol 3, No.1, pp 71-86, 1991.

[13] V. E. Devin and D. C. Hogg, "Reactive Memories: An interactive talking-head," Procs. of British Machine Vision Conference, 2001.

[14] T.Ezzat, G. Geiger and T. Poggio, "Trainable videorealistic speech animation," ACM Transactions on Graphics, Vol. 21, No. 3, pp. 388-398, 2002.

[15] B. Theobald, G. Cawley, S. Kruse and J. A. Bangham, "Towards a low bandwidth talking face using appearance models," Procs. of British Machine Vision Conference, 2001.

[16] C. Bregler, M. Covell and M. Slanry, "Video Rewrite: Driving visual speech with audio," In Proc. ACM SIGGRAPH, 1997.

[17] T. F. Cootes, G. J. Edwards and C. J. Taylor, "Active Appearance Models," In Proc. 5th European Conference on Computer Vision, Freiburg, Germany, 1998.

[18] T. Ezzat and T. Poggio, "Facial Analysis and Synthesis Using Image-Based models," In Proc. 2nd International Conference on Automatic Face and Gesture Recognition, 1996.

[19] M. Brand, "Voice Puppetry,", In Proc ACM SIGGRAPH, pp. 21-28, August 1999

[20] T. F. Cootes, C. J. Taylor, D. H. Cooper and J. Graham, "Active Shape Models - Their training and applications," Computer Vision and Image Understanding, Vol. 61, No. 1, pp 38-59, 1995.

[21] J. R. Deller, J. G. Proakis and J. H. L. Hansen, "Discrete-time processing of speech signals," Macmillan Publishing Company, 1993.

[22] E. B. Nitchie, "How to read lips for fun and profit," Hawthorn Books, New York, 1972

[23] Y. A. Karaulova, P. M. Hall and A. D. Marshall, "A hierachical model of human dynamics for tracking people with a single video camera," Procs. of British Machine Vision Conference, 2000.

[24] R. Bowden, "Learning non-linear Models of Shape and Motion," PhD Thesis, Dept Systems Engineering, Brunel University, 2000.

[25] T. F. Cootes and C. J. Taylor, "A mixture model for representing shape variation," Image and Vision Computing 17, No. 8, pp 567-574, 1999.

[26] T. Heap and D. Hogg, "Improving Specificity in PDM's using a Hierarchical Approach," Procs. of British Machine Vision Conference, 1997.