SPEAKER-INDEPENDENT SPEECH-DRIVEN FACIAL ANIMATION USING A HIERARCHICAL MODEL

D P Cosker, A D Marshall, P L Rosin and Y A Hicks

University of Wales, Cardiff, UK

Abstract

We present a system capable of producing video-realistic videos of a speaker given audio only. The audio input signal requires no phonetic labelling and is speaker independent. The system requires only a small training set of video to achieve convincing realistic facial synthesis. The system learns the natural mouth and face dynamics of a speaker to allow new facial poses, unseen in the training video, to be synthesised. To achieve this we have developed a novel approach which utilises a hierarchical and non-linear PCA model which couples speech and appearance. We show that the model is capable of synthesising videos of a speaker using new audio segments from both previously heard and unheard speakers. 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

Since the pioneering work of Parke [1] in the 1970's, the development of realistic computer facial animation has received a vast amount of attention, crossing over into fields such as movies, video games, mobile and video communications and psychology. The problem not only encompasses the design of a mechanism capable of *representing* a face realistically, but also its *control*. Most computer generated facial systems are based on 3D Models [2] or image based models [3] [4], and are parametric. Using these representations we may animate a face using only a speech signal. This is desirable for many applications such as low-bandwidth network communications and broadcasts, movie lip re-synching and lip-synching for animated movies.

Commonly, automatic speech animation techniques are based on either simplified mappings from phonetically aligned speech signals to viseme key frames (which may then be interpolated) [2] [5], or using a non-linear based model (such as a HMM or a neural network) to define mappings between speech features such as Linear Predictive Coding (LPC) coefficients and facial parameters [6]. The former technique is perhaps the most popular, although it is based on producing only convincing mouth animation and ignores correlations between speech and facial emotion. It also requires pre-processing of the input speech signal before facial synthesis. The latter technique is better suited to producing not only mouth animation, but also facial emotions inherent in the content of the speech signal.

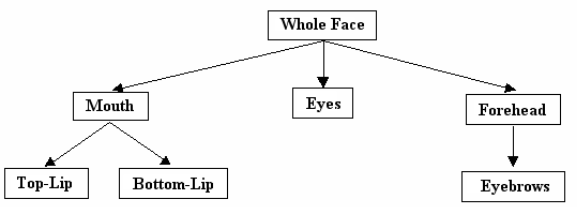

Figure 1: *Hierarchical facial model overview.*

In this paper we present an image based system capable of producing 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 synthesise novel facial animations. During the training phase a small corpus of audio and video is collected of a speaker uttering a list of words that target visually distinguishable speech postures. After training new speech can be supplied, by the original speaker or a new speaker, and synchronised video realistic facial animation is produced. The final video is of the person used in the training phase.

To achieve facial synthesis and animation we introduce a hierarchical non-linear speech-appearance model built from data extracted from the training set. Figure 1 gives an example of a hierarchical model. The face is decomposed into parts to form a hierarchy where the root corresponds to a non-linear model of the whole face and subnodes 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.

2 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. The system then extracts the landmarks from the training set and builds a hierarchical model of the face. For the purpose of this paper we only extend the hierarchy to include the face (as the root node) and the mouth.

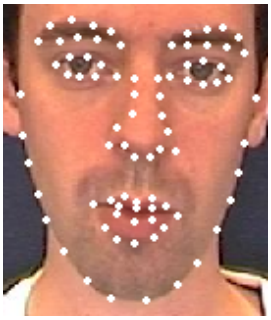

Figure 2: *Annotated training image.*

Given our training set we next extract the data required for each node in the hierarchy. For the representation of a node we introduce a non-linear speech-appearance model. This is an extension of an appearance model introduced by Cootes *et al* [7] encoding relations between appearance parameters and speech vectors allowing the synthesis of facial configurations given new audio. For the root node the model is built using the full facial landmark and image data. For nodes such as the mouth we simply extract corresponding landmarks and texture. For representation of speech signals we process our training audio using Mel-Cepstral analysis.

To achieve facial synthesis given a new speaker we process the incoming audio every 40ms (yielding 25fps) using Mel-Cepstral analysis. We then project this signal into a low-dimensional space and use the non-linear speech-appearance model at each hierarchy node to synthesise a facial area. In the final stage synthesised facial information from sub-nodes is then combined to construct an entire face.

3 DATA AQUISITION

The training process requires the capture of at least 30 seconds of audio and video of a speaker uttering a set of viseme rich phrases with which to build our hierarchical model. We capture audio at 33KHz Mono and video at 25 fps. Each image in our training set is then labelled with 82 landmarks between the top of the eye-brows and the jaw. Figure 2 shows one of our labelled training images annotated with the 82 landmarks.

4 HIERARCHICAL FACIAL MODELLING

Facial area synthesis for each node in our hierarchical model is based on an appearance model [7]. Given an audio input each node is used, in turn, to synthesise a facial area. Given our training set we begin building our model by extracting landmark shape data, and shape-free texture data, for each facial area. Using this data, and the captured audio data, we then build a non-linear speech-appearance model for that node.

The rest of this section describes how a node in the hierarchy is constructed and used for synthesis given speech.

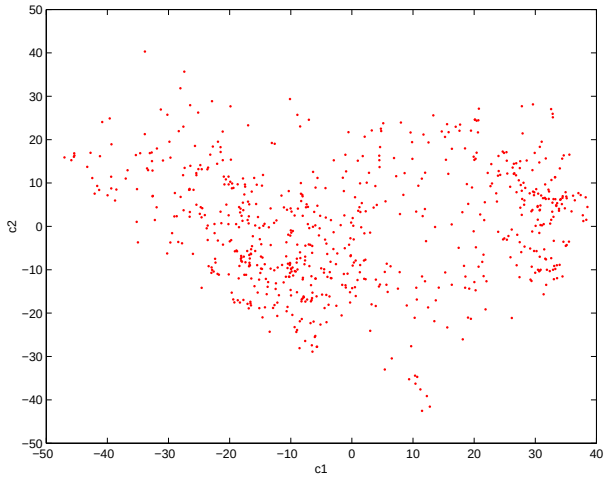

Figure 3: *Distribution of mouth appearance parameters represented by the two highest modes of variation.*

In Section 5 we then describe how nodes are re-combined to construct an entire face.

4.1 FACIAL AREA MODELLING AND NODE INITIALIZATION

Given a subset of facial information from the global training set we first build an appearance model of the corresponding facial area as described in [7]. Statistical PCA models of shape and texture are built using the training set and combined in a joint PCA model. We define this model as follows

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

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

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, $\mathbf{c}$ is the appearance parameter, $\mathbf{W}_s$ is a diagonal scale matrix where each element is a ratio of the eigen-variances of the shape and texture models and $\mathbf{Q}_s$ and $\mathbf{Q}_g$ are the shape and texture parts of the eigenvectors $\mathbf{Q}$. Using this model we then project each shape and texture vector associated with a node into appearance parameter space using

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

giving us n appearance parameters $\mathbf{c}$ for a given node. An example of the distribution of the two highest modes of appearance variation for our mouth node training set is shown in Figure 3.

4.2 NON-LINEAR APPEARANCE MODELLING

By examining Figure 3 we see that our appearance parameter distribution is highly non-linear. Modelling this data

set using a single linear model would degrade its specificity and generalisation [8]. We therefore model the distribution using a mixture of Gaussians [9] initialised using a k-means algorithm. After initial experimentation we found that the natural number of clusters was approximately 60. We discovered that our model tended to be unstable given a data set of less than approximately 400 samples and reached a satisfactory level of stability at around 700 samples.

4.3 ASSOCIATING APPEARANCE WITH SPEECH

Our aim is to encode relationships between our appearance model and our speech training set so that given speech we may estimate an appearance parameter **c** for a nodes facial area. Each cluster in our non-linear appearance model allows the synthesis of a specific set of facial area configurations. Therefore, given a speech signal the first thing we would like to do is to find the cluster which can best synthesise an accurate facial configuration. To achieve this mapping we first reduce the dimensionality of our speech training set using principle component analysis (PCA) yielding the model

$$\mathbf{a} = \bar{\mathbf{a}} + \mathbf{P}_a \mathbf{s} \tag{4}$$

where **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 **s** is a speech parameter. We then reduce the dimensionality of each speech vector using

$$\mathbf{s} = \mathbf{P}_a^T(\mathbf{a} - \bar{\mathbf{a}}) \tag{5}$$

Speech parameters **s** are then concatenated with scaled appearance parameters **c** giving *n* vectors $\mathbf{M}_j$ defined as

$$\mathbf{M}_j = [\mathbf{W}_c \mathbf{c}_j^T, \mathbf{s}_j^T]^T \qquad j = 1, \ldots, n \tag{6}$$

where $\mathbf{W}_c$ is a diagonal matrix where each element is a ratio of the eigenvariances of the speech and appearance models. This gives us *k* clusters of vectors **M**. We then perform a PCA on each cluster to give us *k* joint models of appearance and speech

$$\mathbf{M} = \bar{\mathbf{M}}_i + \mathbf{R}_i \mathbf{d} \qquad i = 1, \ldots, k \tag{7}$$

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.

4.4 APPEARANCE SYNTHESIS FROM SPEECH

Using the joint model of speech and appearance defined for each node we can now calculate a facial areas appearance parameter **c** given $\mathbf{s}_{input}$. First we choose which cluster in our speech-appearance model can best synthesize the facial area by finding the smallest Mahalanobis distance center of each cluster using

$$D = (\mathbf{s}_{input} - \bar{\mathbf{s}}_i)\Sigma^{-1}(\mathbf{s}_{input} - \bar{\mathbf{s}}_i) \tag{8}$$

where $\bar{\mathbf{s}}_i$ is the mean speech parameter in cluster *i* and Σ is the covariance matrix of the speech parameter training set.

Given an appropriate cluster we may now estimate **c** using $\mathbf{s}_{input}$. The mapping process we used was first described by Bowden in [8]. It should however be noted that our process differs in that we cluster our data based only on appearance parameters and not on combined correlated data from different distributions. This difference is due to the nature of our data, we also found that it improves the stability and synthesis quality of our model.

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

$$\mathbf{W}_c \mathbf{c} = \bar{\mathbf{c}}_i + \mathbf{R}_{c,i}\mathbf{d} \tag{9}$$

$$\mathbf{s} = \bar{\mathbf{s}}_i + \mathbf{R}_{s,i}\mathbf{d} \tag{10}$$

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 $\mathbf{s}_{input}$ we then calculate **d** using

$$\mathbf{d} = \mathbf{R}_{s,i}^T(\mathbf{s}_{input} - \bar{\mathbf{s}}_i) \tag{11}$$

and use **d** in (9) to calculate **c**. Finally, we constrain **c** to be within ± 3 s.d's from the mean of its respective cluster and then calculate shape **x** and texture **g** using (1) and (2).

4.5 POST PROCESSING

Given a speech signal a node in the hierarchy is used to synthesise a facial area every 40ms. Given that we are selecting clusters with missing information (the appearance parameters), with no consideration for the prior probabilities of the speech belonging to a cluster, it is often the case that there are a number of possible clusters candidates for selection - and the best choice is not necessarily chosen. The symptom of this is the synthesis of incorrect appearance parameters. However, since this is an infrequent occurance we eliminate the problem visually by performing local averaging of estimated **x** and **g** vectors. In Section 6 we discuss this problem in the context of speech-coarticulation and suggest alternative solutions.

5 RECONSTRUCTING THE HIERARCHY

Reconstruction of the face from its synthesised sub-parts is done in a top-down fashion beginning with reconstruction of the root node. Shape and texture data from the root node is then substituted with data synthesised from sub-nodes by first warping the root to its mean overall shape, warping the sub-facial data with respect to this mean, directly substituting corresponding texture information and then finally warping the concatenated data with respect to the synthesised sub-facial areas new shape. In order to account for possible intensity discrepencies between concatenated areas the pixels in each sub-facial areas are scaled to have a mean and variance equal to that of the root texture.

Figure 4: *Reconstruction of the untrained word "Go" using the training speaker (view from left to right, top to bottom).*

6 EVALUATION

We recorded 28 seconds of video of a speaker uttering a list of words chosen to target specific mouth configurations and labelled 706 of the frames with the aid of an Active Shape Model (ASM) [10]. Using this data we constructed a hierarchical model with a root node for the whole face and a sub node for the mouth. We then recorded the same speaker uttering a set of new words unheard in training and a new speaker uttering a different set of words. Using this new audio we then synthesised video-realistic reconstructions using our hierarchical model. Figures 4 and 5 show a selection of frames from the reconstructions of both speakers which may be found at *http://www.cs.cf.ac.uk/user/D.P.Cosker/research.html* The animations generated using the model are both convincing and realistic, showing strong lip-synch to the audio. However, since the model does not look ahead to the next speech segment in order to modify the current mouth configuration in anticipation of it, we believe that a time based model of coarticulation would improve animation quality. As well as improved coarticulation such a model may also be used for animation of a mouth in the anticipation of a speech sound during silence, i.e. when there is no speech to drive our animation model and would reduce the probability of selecting inappropriate appearance clusters (see Section 4.5).

7 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 capable of synthesising convincing animation given new audio from either the training speaker or a new speaker. The system is also purely data driven requiring no phonetic-alignment before video-synthesis. In future work we hope to extend the model by encoding relations between sub-facial areas and emotional content derived from speech. We also plan to improve animation co-articulation with the inclusion of a time-series based model.

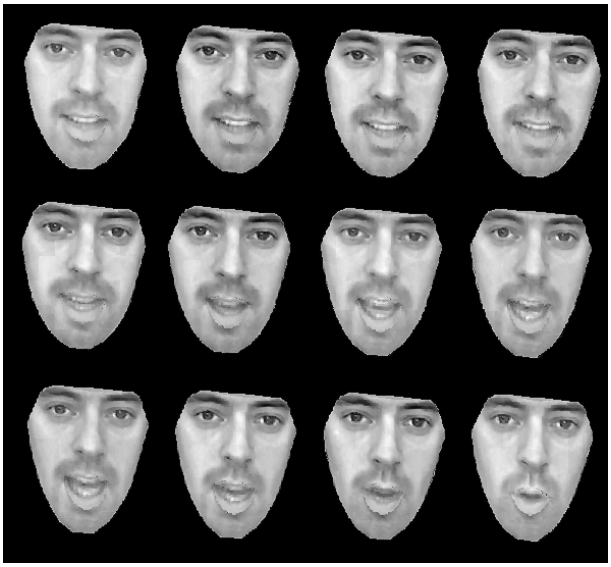

Figure 5: *Reconstruction of the untrained word "Cow" using a new speaker (view from left to right, top to bottom).*

References

[1] Parke F, 1972, "Computer Generated Animation Of Faces", In Proc. ACM National Conf..

[2] Kalberer G, and Van Gool L, 2001, "Face animation based on observed 3D speech dynamics", In Proc. Fourteenth IEEE Conf. on Computer Animation.

[3] Ezzat T, Geiger G and Poggio T, 2002, "Trainable Videorealistic Speech Animation", In Proc. ACM SIGGRAPH.

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

[5] Reveret L, Bailly G and Badin P, 2000, "MOTHER: A new generation of talking heads for providing a flexible articulatory control for video-realistic speech animation", ICSLP'2000, pp. 755-758.

[6] Brand M, 1999, "Voice Puppetry", In Proc ACM SIGGRAPH.

[7] Cootes T, Edwards G and Taylor C, 1998, "Active Appearance Models", In Proc. 5th ECCV.

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

[9] Cootes T and Taylor C, 1999, "A mixture model for representing shape variation", IVC 17, 8, 567-574.

[10] Cootes T, Taylor C, Cooper D and Graham J, 1995, "Active Shape Models - Their training and applications", CVIU, 61, 38-59.