Visual Speech Synthesis Using Statistical Models of Shape and Appearance

Barry J. Theobald[*], J. Andrew Bangham[*], Iain Matthews[†] and Gavin C. Cawley[*]
[*]*School of Information Systems, University of East Anglia, Norwich, NR4 7TJ, UK.*
[†]*Robotics Institute, Carnegie Mellon, Pittsburgh, PA 15123*
Email: b.theobald@uea.ac.uk, {ab, gcc}@sys.uea.ac.uk, iainm@cs.cmu.edu

ABSTRACT

In this paper we present preliminary results of work towards a video-realistic visual speech synthesizer based on statistical models of shape and appearance. A sequence of images corresponding to an utterance is formed by concatenation of synthesis units (in this case triphones) from a pre-recorded inventory. Initial work has concentrated on a compact representation of human faces, accommodating an extensive visual speech corpus without incurring excessive storage costs. The minimal set of control parameters of a combined appearance model is selected according to formal subjective testing. We also present two methods used to build statistical models that account for the perceptually important regions of the face.

1. INTRODUCTION

Many pre-lingually deaf people find closed caption subtitles in broadcast television of less help than might be expected. Sign language is their first language and subsequently some have difficulties learning to read and write using the conventions of an oral language. The problem is similar to that experienced by hearing people when acquiring a second language [13]. Deaf people value the presence of an on-screen signer [12] using, in the UK, British Sign Language (BSL). This has been recognized by UK legislation, which requires terrestrial digital television to provide on-screen signing. This paper is motivated by the need to develop virtual humans capable of delivering sign language at a quality comparable to high bandwidth video. An important feature of such an avatar will be the realistic reproduction of facial gestures. For television broadcast purposes the whole avatar must be driven at a bandwidth of less than 32 kbits/s. This work is concerned with improving visual speech synthesis systems to enhance a signing avatar for the deaf community.

2. BACKGROUND

Research in computer facial animation began in the early seventies with the pioneering work of Parke [18]. In the last decade the rise in popularity of multimedia technologies has lead to an increase in interest in this field [3, 6, 7, 14, 17, 21], an excellent overview is given in [19]. Facial animation has benefitted greatly from work conducted by psychologists in understanding the relationship between faces and facial expression. In particular Ekman and Freisen [9] developed the Facial Action Coding System (FACS), which has become the basis for facial animation systems [18, 22, 25] and face trackers or recognizers [10, 23].

More recently image processing and computer vision techniques have been applied to modeling and animating the face [4, 11, 16, 20]. Image processing techniques can offer a higher degree of realism since real facial images form the basis of the system. However, producing a realistic talking face from a set of real images is still a difficult problem.

Statistical models of the appearance of the face (or the mouth region) have been used for a visual speech synthesizer [5]. Here we describe a system based on the statistics of the combined shape and appearance of the face.

3. SHAPE AND APPEARANCE MODELS

Following the notation of Cootes [8], a statistical model of shape, the *point distribution model* (PDM), is trained by hand labeling a set of images and performing a principle component analysis (PCA) on the coordinates of the located landmarks (aligned to remove any pose variation). Any training shape can be approximated using $\mathbf{x} = \overline{\mathbf{x}} + \mathbf{P}_s\mathbf{b}_s$. Where $\mathbf{P}_s$ is the matrix of the first t_s eigenvectors of the covariance matrix, chosen to describe some percentage, say 95%, of the total variation, and $\mathbf{b}_s$ is a vector of t_s shape parameters (also known as weights or modes).

A model of appearance is computed by shape normalizing all training images so the landmarks in all images lie in the same position as the mean shape. A PCA is computed on the resulting pixel values.

This shape normalization ensures each example contains the same number of pixels and the pixels are compared like-for-like across the whole training set. With such a model any shape-free image can be approximated using $\mathbf{a} = \overline{\mathbf{a}} + \mathbf{P}_a\mathbf{b}_a$. Where $\overline{\mathbf{a}}$ is the mean shape-free image, $\mathbf{P}_a$ is the matrix of the first t_a eigenvectors of the covariance matrix and $\mathbf{b}_a$ a vector of appearance parameters.

Each image is described by a set of shape parameters and a set of appearance parameters, $\mathbf{b}_s$ and $\mathbf{b}_a$ respectively. A combined model of shape and appearance is computed by concatenating the shape and appearance parameters for each image and computing a third PCA. The combined shape and appearance model is given by Equation (1),

$$\mathbf{b} = \mathbf{Q}\mathbf{c}, \tag{1}$$

where $\mathbf{Q}$ is the matrix of the first t eigenvectors of the combined covariance matrix and $\mathbf{c}$ a vector of combined weights. Any image can be reconstructed given a set of weights using Equation (2).

$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{P}_s\mathbf{W}_s\mathbf{Q}_s\mathbf{c}, \quad \mathbf{a} = \overline{\mathbf{a}} + \mathbf{P}_a\mathbf{Q}_a\mathbf{c}, \tag{2}$$

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_s \\ \mathbf{Q}_a \end{pmatrix},$$

where the matrix $\mathbf{W}_s$ takes into account the scaling mismatch between the weights $\mathbf{b}_s$ (which model Euclidean distance) and $\mathbf{b}_a$ (which model pixel RGB intensity). This is described in [8].

Controlling the trajectory through time of the vector $\mathbf{c}$ in Equation (1) creates video-realistic facial animations.

3.1. Data Capture

Models were generated from a database of 9431 images of a single talker uttering 100 sentences constructed to be phonetically rich. The head was not physically constrained but the pose was roughly maintained throughout. The facial expression was held as neutral as possible so the main sources of variation were due to speech. The data was collected in one sitting on a Panasonic DV99B digital camcorder and digitized at a frame rate of 25 frames per second using a Dazzle IEEE 1394 capture card with a frame size of 720x576 (color). The audio was captured at 44.1 kHz stereo and was later used to phonetically segment the video using a hidden Markov model (HMM) speech recognizer run in forced-alignment mode.

4. SYNTHESIS

To obtain the model parameters for synthesizing visual speech a statistical model trained on just 50 images tracked the face across the remaining images using the Active Appearance Model Algorithm (AAM) [8]. The timing information obtained from the audio speech recognizer was used to segment the video. The model parameters for the frames corresponding to each triphone were stored in a look-up table along with the phonetic symbols and durations.

To synthesize a new sentence a text stream is converted to an equivalent set of phonetic symbols and durations. For each phone to be synthesized, the original training data is searched for a phone appearing in the same context, and the original triphone model parameters re-sampled using cubic b-splines [24] to obtain the correct duration. To model coarticulation the neighboring model parameters are overlapped, which also ensures smoother transitions from one phone to the next. To give more significance to the center phones a window is applied to each triphone prior to overlapping the neighboring regions.

At this preliminary stage we only synthesize sentences that contain triphones occurring at least once in the training data. We will address later the problem of a non-matching context using a technique similar to Arslan et al. [2] or Bregler et al. [4].

5. PERCEPTUAL TEST USING EQUIVALENCE

The first few parameters of the appearance model, Equation (1), capture most of the variation of the facial features. To design a talking face requires a judgment on both the number of parameters, t, and how many training examples are required to obtain a stable estimate of these parameters.

Each curve shown in Figure (1) represents the mean number of model parameters from 20 models generated by randomly sampling a number of images from the database. It shows how the number of parameters required to account for $\{30\%, 70\%, 90\%, 91\%, 92\%, 93\%, 94\%$ and $95\%\}$ of the total variation increases with the number of training examples. Clearly, 30% and 70% of the variation is captured very quickly in just a few parameters and if 1000 images are analysed to account for 90% of the variance almost no extra information is then gained by incorporating another 1000 images. To account for 95% of the variation would require in excess of 2000 training examples.

It has been shown elsewhere [9, 18] that as few as 50–60 parameters are required to describe the visible deformation of the face. Figure (1) shows 90% of the variance can be reliably captured in 31 parameters. However, this does not mean that only 31 parameters are required for a good reproduction of the faces for synthesis. To explore this requires perceptual testing. 92% of the variation was captured in 67 parameters, which is reassuringly around the same number of parameters as used by Parke [18] and Ekman [9].

5.1. Perceptual Test

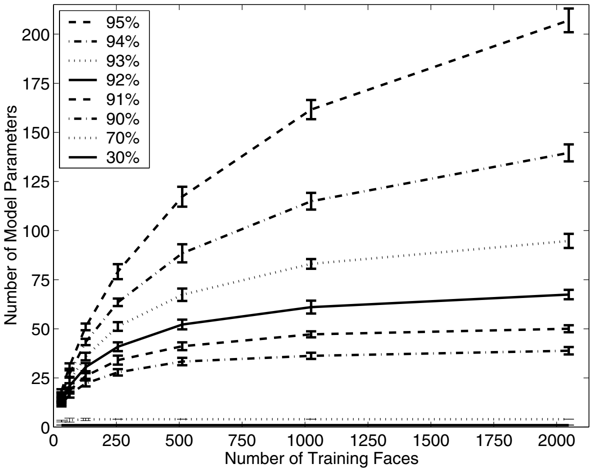
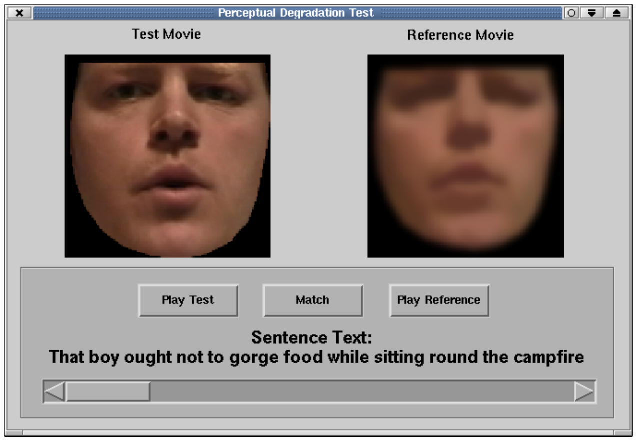


Figure 1: The family of curves show the number of model parameters required to capture a given proportion of the variation seen in the training set increases as the number of training images increases. Each of the curves represents a percentage of the variation captured and the error bars show ±2 standard deviations. At each point 20 models were built, 1120 in total.

The purpose of perceptual testing is to rank the relative quality of different models. Figure (1) shows an extra 140 parameters are required to capture a 3% increase in variation from 92% to 95% and the first 90% of the variation is captured in 31 parameters. For synthesis we seek a model that is as compact as possible, but still able to accurately reconstruct the face. Compactness allows a larger database to be stored and is more computationally efficient. If the higher dimensions of the model contain no perceptual information nothing is gained, in terms of image reconstruction quality, by retaining them.

Standard perceptual testing practice used by, for example, the MPEG committee in assessing the quality of video coding algorithms, is defined in ITU Rec. BT.500 [1] and requires around 20 naive viewers to view the test sequences (double blind) and score quality in comparison with a good reference. Here we propose a modification modeled on that used in text-to-speech synthesis [15]. Rather than implicitly asking viewers to score good-to-bad, they are asked to say when the test and degraded sequences are *equivalently* bad.



Sentence Text:
That boy ought not to gorge food while sitting round the campfire

Figure 2: A GUI that plays either a test video (left) or a reference (right). The slider controls degradation applied to the reference and the user presses 'match' when both movies are judged to be equally bad.

A Matlab GUI is used, Figure (2), that plays a test and a reference movie. The reference movie is the face patch extracted directly from the original video. The slider allows the user to vary the degree of degradation applied to this movie, which increases monotonically from zero (slider value to the right) to maximum (left) so the quality of the movie decreases from good-to-bad. It is important that at one end the reference is worse than, and at the other the reference is better than the models under test. The viewer is asked to select a position for the slider at which point the test and reference are judged to be equally good (or bad).

The test movies are synthesized versions of the original reference video and in a typical experiment 7 tests are compared by showing each 3 times at random: a procedure that takes about 10 minutes. Four different ways of degrading the video were considered: 1) Gaussian blurring, 2) temporal blurring, 3) morphological filtering, and 4) temporal warping, but many others could be devised. Only the results from method 4) are presented here.

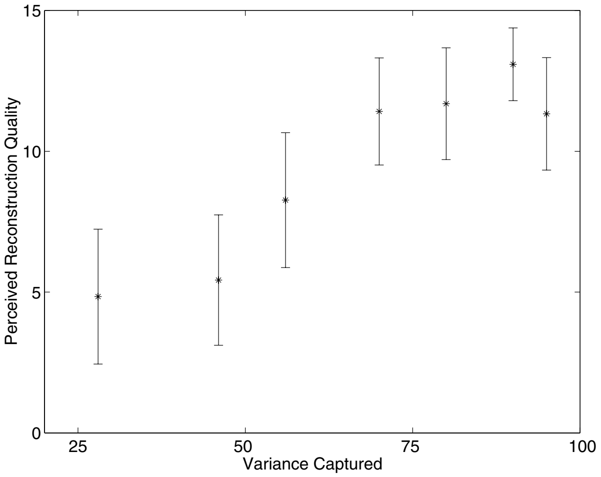


Figure 3: Results from the perceptual test using temporal warping as a degradation. The error bars show ±2 times the standard error.

The test model was trained on 1024 images (the model contained 98 landmark points and the face patch 18000 pixels) and synthesized the reference movie using $t = \{1, 2, 3, 6, 14, 67, 217\}$ parameters, or equivalently {28%, 45%, 56%, 70%, 80%, 90% and 95%} of the total variance.

The results shown in Figure (3) show the perceptual information judged to be obtained from the models capturing 28%–56% is lower than those capturing higher variance — as would be expected. The models capturing the higher dimensions (70%, 80%, 90%, and 95% variance) are judged to convey a similar degree of perceptual information.

There is plenty of room to expand on this experiment. For example, it would be appropriate to include the audio signal degraded sufficiently to force viewers to extract information from the synthesized face in order to understand what is being said. Also other forms of degradation should be investigated and the number of users increased. Twenty users took part in this perceptual test, five for each of the different degradations.

5.2. Perceptual Models

Building a statistical model of the shape and appearance of the face using a linear PCA, as described in Section 3, gives rise to unwanted artifacts. The linear PCA is unable to model non-linear variations in the training data correctly, as can be seen in Figure (4A), where the inside of the mouth and eyes appear blurred. The mouth is the region that conveys most of the perceptually important information during speech and a good reconstruction of the eyes is required for video-realism. Both of these regions are characterized by regions of high spatial bandwidth (the teeth in the mouth and the sclera+pupil combination in the eyes) that may be occluded by the lips or the eyelids.

Assuming that the model has been correctly aligned these features appear as outlying clusters in the greyscale space. We therefore allow the possibility of modeling separate greyscale models in regions that are known to be perceptually important. We call these models Multi-segment Appearance Models, MAMs. They are different from conventional clustering because here each sub-model is forced to occupy a perceptually important subspace. This technique is particularly useful where one needs to represent fine scale features such as eye color.

5.2.1. Modeling the Face using a MAM

To construct a MAM the images are segmented into perceptually important sub-regions. For the face the sub-regions are the whole face, the mouth and each of the eyes. The combined model of shape and appearance denoted as $\mathbf{b}^w = \mathbf{Q}^w \mathbf{c}^w$ is computed first. The shape weights, $\mathbf{b}^w_s$, are computed from, and will control, the position of the landmark points, $\mathbf{x}^w$, around the face. The appearance weights, $\mathbf{b}^w_a$, are obtained from the intensity of the color pixel values in the whole face region (including those contained within the eyes and mouth).

The shape weights in $\mathbf{b}^w_s$ can be used to determine whether or not the mouth and eyes are open. If open, the pixel values contained within the points $\mathbf{x}^{si} \subset \mathbf{x}^w$ that form the (open) segments are sampled and spatially normalized by warping to the corresponding mean position $\overline{\mathbf{x}}^{si}$. A PCA is computed on these subsegments individually, giving three further appearance models $\mathbf{a}^i = \overline{\mathbf{a}}^i + \mathbf{P}_a^i \mathbf{b}_a^i$. Where $\mathbf{a}^i$ denotes the shape-free appearance of i^{th} region.

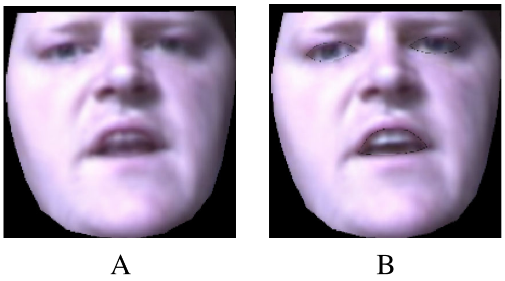


Figure 4: Two synthesized video frames. The left frame is synthesized using conventional statistical modeling techniques, while the right has been synthesized using a perceptual approach. The model segments are outlined in black in the right frame.

Given a set of weights to reconstruct the whole face, $\mathbf{c}^w$, the landmark points and the shape-free appearance can be generated using Equation (2). Then, $\mathbf{x}^w$ is used to find which, if any, of the eyes and mouth are open. If open, a local appearance is generated and warped to the corresponding points. If shut, the sub-segment is generated by taking the pixels from the global appearance model $\mathbf{a}^w$.

In summary, a conventional combined statistical model of shape and appearance codes an image as a single vector ($\mathbf{c}$), as in Equation (1). Whereas, a multi-segment appearance model (MAM) codes an image as $\mathbf{d} = \{\mathbf{c}, \mathbf{b}_a^i\}$ $i = 1 : N$ segments.

5.2.2. A Further Perceptual Model

The reference shape to which the training images are warped when building the appearance model is generally the mean shape, $\overline{\mathbf{x}}$. More significance can be given to the mouth and eye regions prior to computing the appearance PCA (on the pixel intensities) by choosing a normalizing canonical shape other than simply the mean. The shape normalization and reconstruction warps then retain relatively more pixels for these areas. An example of an image warped to an artificial set of points is shown in Figure (5A) along with a face synthesized using this model.

The face in Figure (5B) is again better reconstructed than that using a conventional shape and appearance model. The eyes and inside of the mouth appear clearer. There are artifacts in the synthesized face in Figure (5B) that are undesirable. Namely the 'scratches' that are apparent on the right cheek and along the edge of the left cheek. This would suggest care must be taken when selecting the set of points to which the training images are warped.

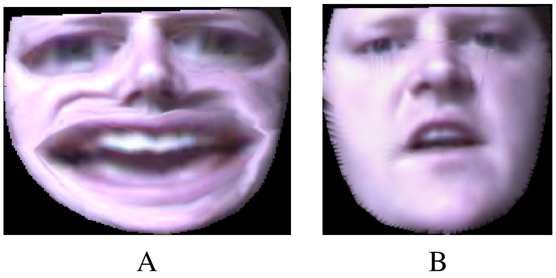


Figure 5: The left frame shows an image warped to an artificial set of points, while to the right shows a synthesized face recreated using this model.

6. CONCLUSIONS

Combined models of shape and appearance have been successfully used elsewhere for tracking and recognizing faces in images and video sequences. This paper has discussed how such a model can be used for synthesizing video-realistic visual speech, Section 4.

When building a model a given percentage of the total variation to be modeled is pre-selected and sufficient model parameters to describe this percentage are retained. It has been shown, in Section 5, that as the percentage variation to be modeled increases the number of parameters required to capture this variation dramatically increases. For a face tracker the model must capture as much variation as possible to ensure robustness. For synthesis only the parameters that capture the perceptual changes in the face need be retained. This not only ensures efficient data storage, but also greatly reduces the computational requirements in reconstructing the face.

Preliminary investigations into determining which model parameters are perceptually important have been outlined, Section 5.1. In particular a test that allows a user to judge when the perceptual information gained from a model is equivalent to the perceptual information gained from a degraded video was outlined. This differs from standard test procedures, for example those used by MPEG, where users are asked to judge quality rather than equivalence.

The users who took part in this experiment found the higher dimensions of the face space to be perceptually insignificant. Further work requires the investigation of different distortion types.

Preliminary work on building models of the face that give significance to the perceptually important regions was outlined in Section 5.2. These results look promising compared to conventional statistical modeling techniques. Further work will include evaluating the perceptual improvement gained by these new methods.

7. ACKNOWLEDGMENTS

The authors would like to thank all persons who took part in the perceptual tests.

8. REFERENCES

1. Methedology for the subjective assessment of the quality of television pictures. Technical report, Recommendation ITU-R BT.500-10, 1974-1978-1982-1986-1990-1992-1994-1995-1998-1998-2000.
2. L.M. Arslan and D. Talkin. 3-d face point trajectory synthesis using an automatically derived visual phoneme similarity matrix. In *Proceedings of Auditory-Visual Speech Processing*, pages 175–180, 1998.
3. J. Beskow. Talking heads - communication, articulation and animation. In *Proceedings of Fonetik-96*, pages 29–31, 1996.
4. C. Bregler, M. Covell, and M. Slaney. Video rewrite: driving visual speech with audio. In *Proceedings of SIGGRAPH*, pages 353–360, 1997.
5. N.M. Brooke and S.D. Scott. Two- and three-dimensional audio-visual speech synthesis. In *Proceedings of Auditory-Visual Speech Processing*, pages 213–218, 1998.
6. J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achron, T. Becket, B. Douvill, S. Prevost, and M. Stone. Animated conversation: rule based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *Proceedings of SIGGRAPH*, pages 413–420, 1996.
7. M. Cohen and D Massaro. Modeling coarticualtion in synthetic visual speech. In N.M. Thalmann and Thalmann D, editors, *Models and Techniques in Computer Animation*, pages 141–155. Springer-Verlag, 1994.
8. T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. In H. Burkhardt and B. Neumann, editors, *Proceedings of the European Conference on Computer Vision*, volume 2, pages 484–498. Springer-Verlag, 1998.
9. P. Ekman and W.V. Friesen. *Manual for the Facial Action Coding System*. Consulting Psychologists Press Inc., 1978.
10. I. Essa and A. Pentland. Coding, analysis, interpretation and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):757–763, July 1997.
11. T. Ezzat and T. Poggio. Visual speech synthesis by morphing visemes. Technical Report 1658/CBCL, MIT, 1999.
12. Bristol University Centre for Deaf Studies. Deaf people and television. *Research Notes*, 10, 1995.
13. S. Gregory, J. Bishop, and L. Sheldon. *Deaf Young People and their Families: Developing Understanding*. Cambridge University Press., 1995.
14. A. Hallgren and B. Lyberg. Visual speech synthesis with concatenative speech. In *Proceedings of Auditory-Visual Speech Processing*, pages 181–183, 1998.
15. R.D. Johnston. Beyond intelligibility - the performance of text-to-speech synthesisers. *BT Technology Journal*, 14(1):100–111, 1996.
16. Z. Liu, Z. Zhang, C. Jacobs, and M. Cohen. Rapid modeling of animated faces from video. Technical Report MSR-TR-2000-11, Microsoft Corporation, 2000.
17. S. Morishima. Real-time talking head driven by voice and its application to communication and entertainment. In *Proceedings of Auditory-Visual Speech Processing*, pages 195–199, 1998.
18. F.I. Parke. *A Parametric Model for Human Faces*. PhD thesis, University of Utah, Saltlake City, Utah, 1974.
19. F.I. Parke and K. Waters. *Comptuer Facial Animation*. A K Peters, 1996.
20. F. Pighin, J. Hecker, D. Lischinski, R Szeliski, and D. Salesin. Synthesizing realistic facial expressions from photographs. In *Proceedings of SIGGRAPH*, 1998.
21. L. Reveret and C. Benoit. A new 3d lip model for analysis and synthesis of lip motion in speech production. In *Proceedings of Auditory-Visual Speech Processing*, 1998.
22. D. Terzopoulos and K. Waters. Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(6):569–579, 1993.
23. Y. Tian, T. Kanade, and J.F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, February 2001.
24. M. Unser. Splines: a perfect fit for signal and image processing. *IEEE Signal Processing Magazine*, 16(6):22–38, 1999.
25. K. Waters. A muscle model for animating three-dimensional facial expressions. *Computer Graphics (SIGGRAPH '87)*, 21(4):17–24, 1987.