

# Large Vocabulary Audio -Visual Speech Recognition Using Active Shape Models

Tanveer A Faruque    Abhik Majumdar    Nitendra Rajput    L V Subramaniam  
*IBM India Research Lab, New Delhi, India*  
*{ftanveer,rnitendr,lvsubram}@in.ibm.com*

## Abstract

*Orthogonal information present in the video signal associated with the audio helps in improving the accuracy of a speech recognition system. Audio-visual speech recognition involves extraction of both the audio as well as visual features from the input signal. Extraction of visual parameters is done by the recognition of speech dependent features from the video sequence. This paper uses geometrical features to describe the lip shapes. Curve-based Active Shape Models are used to extract the geometry. These geometrically represented visual parameters are used along with the audio cepstral features to perform an audio-visual classification. It is shown that the bimodal system presented here gives an improvement in the classification results over classification using only the audio features.*

## 1. Introduction

With the penetration of machines into human life, ease of use becomes a critical factor for the computer to be more widely acceptable. Speech recognition is one of those critical areas which provide a human like interface with the machine. Considerable research is being done to improve the speech recognition accuracy while catering to a wide variety of speakers. Recently work has been done on Audio-Visual speech recognition [1][2][8][10] where the aim is to use the lip and other facial features in addition to the audio for further improving the accuracy of a speech recognition system. In addition to extracting the audio features, this involves extraction of lip shapes from the face in the video which can then be parameterised in accordance with the different audio units. Several techniques exist for fitting contours in objects of an image [3][4][5], however, if there is a priori knowledge about the shape of the object being recognized, Active Shape Models (ASM) can be utilized to fit a contour on the shape of interest. ASMs have been widely used as a model-based method for extracting shape information about deformable shapes from image sequences, including lips [5]. The shapes to which it can

deform are constrained by a Point Distribution Model (PDM) [11]. PDMs represent objects as a set of points labeled over the object of interest.

We propose to substitute the PDM by a model in which we use a set of parametric curves to mark out the contours of a deformable shape, instead of points as in a PDM. By representing the shape with parametric curves it is possible to reduce substantially the number of parameters required to represent the model.

The rest of the paper is organised as follows. In Section 2 we first present the basic technique of ASMs. The ASM is then modified to incorporate color information and finally we discuss the proposed extension to a curve-based ASM for providing accurate lip shapes. In Section 3 the set of visual features used to represent the lip shape variation are defined. Once the visual parameters are defined in the video sub-space, the dimensionality of the speech recognition system is increased to have the visual dimensions in addition to the audio feature vector dimensions. Classification experiments for the combined feature vectors are performed to see the discriminating capability between the different phonemes. Section 4 provides the details of these classification experiments. The last section compares the result of audio only and audio-video classification using different ASMs.

## 2. Active Shape Models

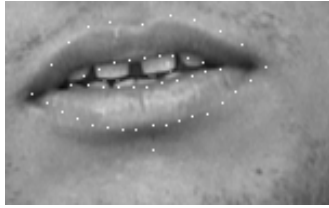
Active shape models were first proposed by Cootes and Taylor [3] to locate deformable shapes in medical images. Subsequently they have also been used in lip feature extraction for audio-visual speech recognition [6][7]. Active shape models are flexible statistical models which represent salient features of the object through a set of labeled points. The practice is to use a Point Distribution Model to represent these labeled set of points.

ASM uses a priori knowledge from statistics over a training set of images. For training, a set of lip images is selected which provide enough variation in lip shapes and appearance across different persons. Points on these training images are hand-marked such that the exact

position of the contour over the object is known to the training system for each of these images. The implementation in this paper uses 46 points in total which have been obtained by marking the object at specific intervals (as shown in Figure 1 below). Each image is represented by a point-set given by

$$p_i = (x_1, y_1, x_2, y_2, \dots, x_N, y_N)^T$$

where N is the number of points used to represent the contour. Here  $(x_1, y_1, x_2, y_2, \dots)$  represent the (x,y) coordinates of the point 1,2,...,N marked on the image. All the images are normalized and aligned with respect to scaling, rotation and translation.



**Figure 1. The Point Distribution Model (46 points are hand marked in white)**

A Principal Component Analysis (PCA) of the training data gives the mean position of the points, which gives the average shape, and modes of variation describing the ways in which the shapes tend to deform from the average shape. The mean shape of the training image database is given by

$$p_{mean} = \frac{1}{M} \sum_{i=1}^M p_i$$

The deviation of different lip shapes from this mean shape is captured by the other dimensions of the PCA of the point-sets. So we can represent any contour in the training data by the mean and the weighted sum of these modes of variation, i.e.,

$$p = p_{mean} + E_s b_s$$

where  $b_s$  is a vector representing the weights of the different dimensions and  $E_s$  is a matrix containing the eigenvectors of the PCA. This set, comprising the mean and the modes of variations, represents the ASM system. Each shape parameter is constrained to lie within  $\pm 3$  standard deviation of the training set which accounts for 97% of variation.

After initialization, assuming that all shapes are equally probable, a cost function  $e$  is associated with the contour and is evaluated for every set of  $b_s$ . The cost function used here is

$$e = (g - g_{mean})^T (g - g_{mean}) - b_i^T b_i$$

where  $g$  is the grey level profile vector,  $g_{mean}$  is the average grey level profile vector, and  $b_i$  is a set of parameters that describe the profile model [3].  $b_i$  is obtained from the PCA of the grey level profiles using

$$b_i = E_s^T (g - g_{mean})$$

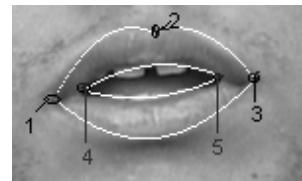
This  $e$  is a measure of how well the grey level profiles around the current model points match those seen in the training set. The contour that minimizes this cost function is chosen as the final contour representing the shape of the lip in the given image.

While working in the specific domain of lip images, a better fit can be obtained by incorporating color information in ASM. Lips have a definite continuous color within the image, this color information is exploited by taking color profiles along the PDM. Three separate models for each Red, Blue and Green color are formed. The total cost function  $e_{total}$  is given by

$$e_{total} = w_{red} \times e_{red} + w_{green} \times e_{green} + w_{blue} \times e_{blue}$$

where,  $w_{red}$ ,  $w_{green}$  and  $w_{blue}$  are the respective weights for each of the three colors.

The regularity of lip shapes can be further exploited using curve based ASM. Here the lip contours are represented by curves. Hence in the fitting process, it is not the points but the parameters of the curve that are to be iterated. In the curve based model regularity of the lip shape is maintained by constraining it to the shape as given by the set of curves used to represent the lips.



**Figure 2. Representation of lip contour by five parabolic curves**

As seen in Figure 2, the five parabolic curves very closely represent the inner and outer lip contours. The object of interest can then be modeled by training the ASM using these five second-order curves and the five end points marked in Figure 2. The parabola is represented by three coefficients  $[c_0, c_1, c_2]$ , where the coordinates are related by

$$y = c_0 x^2 + c_1 x + c_2$$

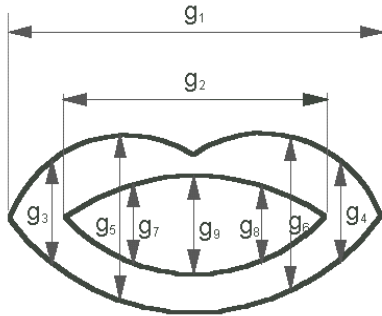
The outer contour of the lip can be very closely approximated by a set of three parabolas and the inner contour by a set of two parabolas as shown above. Even though the curves represent the whole contour, the

number of parameters required to do this are very few (25 for this representation as opposed to 92 for the PDM). It also restricts the final contour shape to one which is more likely and defined, thus ensuring that the lower bound on the quality of fit goes up. Regularity in the final contour increases owing to the curve shape as compared to the points which could go to out of place.

### 3. Extraction of Visual Features

In the last section we presented the technique of ASM to fit a contour over the lip boundaries to capture its shape. Here we detail the various geometrical parameters that we extract from these contours to represent the shape for use in bimodal speech recognition. Parameters should be so chosen that they maximize the cluster differences when projected on to the phonetic space. We choose 12 different parameters to represent the shape of the lip in the video vector. The audio is represented by 24 cepstral coefficients. These are then combined to perform an audio-visual classification. The visual parameters chosen for the experiment are the nine geometrical parameters representing the lip shape through distances as shown in Figure 3. Also the weight of top twelve eigen vectors are taken as parameters to represent the shape of the lip region. The idea here is that the information conveyed should be of help in deciding the phoneme being uttered with this mouth shape. The nine geometrical parameters chosen are such that they vary in magnitude when the shape change. The width and height of the outer and inner contour of the lip is captured along with the height at other portions along the lip shape. These are the directions having maximum variation. So the final visual feature vector is represented by

$$\vec{v} = [b_1, b_2, \dots, b_{12}, g_1, g_2, \dots, g_9]^T$$



**Figure 3. Geometrical parameters of a given lip shape**

Where,  $\vec{v}_i$  is the video vector representation of the image frame  $i$ ,  $b_1, \dots, b_{12}$  are the weights of the eigen

vectors as in equation 3 and  $g_1, g_2, \dots, g_9$  are the geometrical parameters as marked in Figure 3.

### 4. Classification Experiments

After extracting the shape of the lip, the ability of the shape to represent the audio is measured. Classification is performed over the phonetic space using the feature vectors from the visual space. This is then compared with other approaches of representing articulatory video [8].

The data we have used for performing the experiments is from an MPEG2 video stream. Audio is demultiplexed from this signal and is decoded to perform the signal processing. A 24-dimensional mel-cepstral coefficient feature vector is formed for each audio frame which is of a duration of 10 ms. The procedure for processing of the audio stream is presented in [8]. For the video, each frame (decoded at 30 frames per sec) is extracted from the MPEG sequence and the pyramid based face tracking algorithm is used to track the face [9]. A mouth image of size 140x90 is then extracted from the face image centered around the lips. Then the visual feature vector is formed using the techniques mentioned in the last section. We now have a visual feature vector which is of duration 1/30 sec. These feature vectors are then interpolated to have a visual feature for each 10 ms so that the frame rate for the audio and video vectors is the same.

Standard speech recognition can be applied to the classification and subsequent recognition of visual features. In order to compare the effect of the visual features to improve the recognition accuracy, a technique of integrating the audio and visual data for the purpose of recognition is required. We use the early integration approach involving the extraction of audio features and the visual features and then concatenating the two to generate a single audio-visual feature vector [10]. The combined vector is then projected on the phonetic space and a classification is performed.

We carry out the classification experiments first on the audio-only data, then on the video-only data. Then the two are combined and an audio-visual classification experiment is performed. Since the technique of ASM involves iterations, it is computationally intensive. So the amount of data being used for performing such a classification is bounded by the amount of visual feature vectors generated. The data used is about 10 minutes of continuous speech with words from a large vocabulary consisting of around 60,000 words. For classification, each phonetic class is represented by a mixture of 5 Gaussian components. Such Gaussian Mixture components are computed in the audio and video feature space. The audio and video model likelihoods are

computed from the respective Gaussian models and then a joint likelihood is computed by adding the two likelihoods.

Visual feature vectors are obtained by first fitting the contour over the lip shape. This is done by using (a) grey level distribution ASM, (b) color level distribution ASM, and, (c) a curve based ASM. The features obtained by the three methods are compared by repeating the classification on the three feature sets.

## 5. Results and Conclusions

Table 1 shows the comparison for classification rates of the three different methods. It is to be noted that the results are based on classification at the phonetic level.

**Table 1. Phonetic classification rates using the three ASMs**

	Audio-only (24-dim)	Video-only (21-dim)	Audio-video (45-dim)
Grey-level	55.29%	25.87%	53.99%
Color-level	55.29%	29.03%	56.67%
Curve-based	55.29%	31.91%	56.81%

Recognition rates for the speech recognition system will be much higher as they are calculated by combining the classification rates with the language model contexts. It is seen that the video classification rate is obviously less than the audio classification rate as the classification is performed in the phonetic space which is more representative of the audio signal than the video signal. However it is clearly seen that using the color information in the ASM gives a relative improvement of 12% over the grey level based ASM for representing the visual features. Similarly the curve based ASM shows an improvement of around 10% over the color based ASM and 23.35% over grey level based ASM. Also the results are shown with the audio being corrupted with a 15dB speech noise to see the improvement achieved by including the visual features in the recognition.

In [6] and [7], results have been presented for isolated word/digit recognition using the point distribution model based ASM. In this paper we have extended these experiments to continuous speech recognition using ASM to compute the visual feature vectors. The results

presented here show that the addition of more information in the form of color improves the audio-visual classification accuracy. Further improvement in classification accuracy is obtained by the use of restrictive models defined using a priori knowledge of the lip shape. Therefore, curve based ASMs, which restrict the shape of the model for fitting over the lips, when combined with color information offer good improvement over existing PDMs.

## References

- [1] E. D Petajan, "Automatic lipreading to enhance speech recognition," *Proc. IEEE Global Telecommunication Conf., Atlanta*, 1984
- [2] G. Potamianos and H. P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," *International Conference on Acoustic Speech and Signal Processing*, 1998.
- [3] A. L. Yuille, P. W. Hallinan, D. S. Cohen, "Feature Extraction from Faces Using Deformable Templates," *International Journal of Computer Vision*, 1992, pp 99-111.
- [4] M. Kass, A. Witkin, D. Terzopoulos, "Snakes: Active Contour Models," *International Journal Computer Vision*, vol. 1, pp. 321-331, 1988.
- [5] T. F. Cootes, C. J. Taylor, "Active Shape Models - 'Smart Snakes'," in *Proc. British Machine Vision Conference*. Springer-Verlag, 1992, pp.266-275.
- [6] J. Luettin, N. A. Thacker, S. W. Beet, "Visual speech recognition using active shape models and hidden Markov models," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1996.
- [7] I. Matthews, "Features for Audio-Visual Speech Recognition," Ph. D. Thesis, School of Information Systems, University of East Anglia, September 1998.
- [8] S. Basu, C. Neti, A. Senior, N. Rajput, L. Subramaniam, A. Verma, "Audio-Visual Large Vocabulary Continuous Speech Recognition in the broadcast domain," *IEEE Workshop on Multimedia Signal Processing*, Sep 13-15, Copenhagen 1999.
- [9] Andrew Senior, "Face and feature finding for face recognition system," *2nd Int. Conf. on Audio-Video based Biometric Person Authentication*, Washington DC, March 1999.
- [10] A. Verma, T. Faruque, C. Neti, S. Basu, A. Senior, "Late integration in audio-visual continuous speech recognition," *Automatic Speech Recognition and Understanding*, 1999.
- [11] T.F.Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. "Training models of shape from sets of examples," *Proc. British Machine Vision Conference*, pp 9-18. BMVA press 1992.