# A SYNTACTIC APPROACH TO AUTOMATIC LIP FEATURE EXTRACTION FOR SPEAKER IDENTIFICATION

*T. Wark and S. Sridharan*

Speech Research Laboratory
Signal Processing Research Centre
Queensland University of Technology
GPO Box 2434, Brisbane QLD 4001, Australia
twark@markov.eese.qut.edu.au

## ABSTRACT

This paper presents a novel technique for the tracking and extraction of features from lips for the purpose of speaker identification. In noisy or other adverse conditions, identification performance via the speech signal can significantly reduce, hence additional information which can complement the speech signal is of particular interest. In our system, syntactic information is derived from chromatic information in the lip region. A model of the lip contour is formed directly from the syntactic information, with no minimization procedure required to refine estimates. Colour features are then extracted from the lips via profiles taken around the lip contour. Further improvement in lip features is obtained via linear discriminant analysis (LDA). Speaker models are built from the lip features based on the Gaussian Mixture Model (GMM). Identification experiments are performed on the M2VTS database [1], with encouraging results.

## 1. INTRODUCTION

There is an increasing requirement for robust and reliable person authentication systems in areas of high security or secure access. The majority of current techniques for person authentication focus on either static facial information [2] or speaker recognition via the speech signal [3]. Whilst in clean conditions, the speech signal has proved to be a valuable source of speaker dependent information, problems occur in noisy or channel mis-match conditions.

Whilst lip information presents predominantly speech dependent information, valuable speaker dependent information is also contained within the static and dynamic features of the lips [4]. Current work aims at developing a person authentication system using lip information alone.

Lip tracking has attracted a lot of recent interest, due to the complementary nature of its information to the speech signal. The task however is very difficult, particularly where systems must cope with the facial movement or poor lighting conditions. Gradient based techniques for edge detection of lips are often not successful due to the poor contrast of lips to the surrounding skin region. Successful lip tracking via intensity information has been presented using active shape models [5], however this technique requires *a priori* knowledge of the lip shape via manual labelling.

Recent work [6] has used B-splines to track the outer lip contour using chromatic information around the lips. Other similar techniques [7] also use colour information to build a parametric deformable model for the lip contour. These techniques require optimization or iterative techniques to refine estimates of the contour model to the lips.

We present a new technique for the tracking and extraction of lip information for the purpose of speaker identification. The tracking scheme extracts syntactic information from the lip region. A lip contour model is derived directly from the syntactic information, with no minimization procedure required to refine estimates. A chromatic lip model is then built from profiles taken around the lip contour.

Classification of lip information for speaker identification is achieved using the Gaussian Mixture Model (GMM). Identification experiments are performed on the M2VTS database [1], with encouraging results.

## 2. LIP LOCALISATION

In order to obtain lip information, a region of interest (ROI) for the lips must be obtained from each facial image. We present a scheme for locating lips based on the use of chromatic information.

### 2.1. Segmentation of Facial Skin Area

Previous work [8] has shown that lip shape can be approximated by locating those pixels which conform with:

$$L_{lim} \leq \frac{R}{G} \leq U_{lim} \quad (1)$$

where $R$ and $G$ are the *red* and *green* colour components respectively. $L_{lim}$ and $U_{lim}$ are dependent on the particular lighting used over the range of facial images. As many other facial regions exhibit this same property, such as pixels lying around the nostrils and eyes, we cannot simply locate the lips from the complete facial image using this technique.

We can make a very general assumption that lips will be located in the lower portion of the facial skin region. Each image in the M2VTS database, also contained a fair amount of information outside of the facial skin region such

This work was supported by a CSIRO research contract.

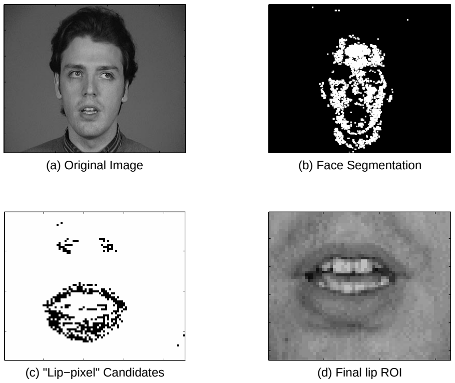


Figure 1: Location of Lips in Facial Image

as hair, neck and shoulders, hence processing was first required to extract just the facial skin region. We chose to use an adaption of the $L_{lim}$ and $U_{lim}$ parameters to segment the facial skin region from the surrounding data. We found suitable limits to be $L_{lim} = 1.2$ and $U_{lim} = 1.45$ for the database used.

Following facial segmentation, spurious pixels in the background were removed via morphological opening. The results of this process are shown in Figure 1(b).

### 2.2. Extraction of Lip Area

Once the facial skin had been segmented from the rest of the image, the lip search was restricted to the lower half of the facial area. We used the same technique to locate lip pixels, this time with limits set to $L_{lim} = 1.7$ and $U_{lim} = 2.0$. The outcome of this process is shown in Figure 1(c).

The final stage of the process was to pass a sampling window over the evaluation area of Figure 1(c) in coarse jumps. The window position with the highest number of points contained with the sampling box was chosen to be the lip position. A final larger window, centred around the sampling window, was chosen as the final ROI for the lips, as shown in Figure 1(d).

## 3. LIP TRACKING

This section describes a new procedure for the automatic extraction of lip information for tracking. We define an automatic lip tracking system as one where no prior manual labelling from a training set is required, in order to gain lip information.

### 3.1. Extraction of Lip Shape

It is assumed that a ROI for the lips has already been identified, as outlined in Section 2. Once again we used the same approach for the extraction of lip pixels as outlined in Equation 1. This approach assumes that the ratio of *red* to *green* colour component pixels are somewhat constant over the lip area. On this occasion we applied slightly broader limits in order to detect as much of the lip shape as possible, and to cater for variations in skin colour over a range of subjects. Suitable limits were found to be $L_{lim} = 1.5$ and $U_{lim} = 2.2$. The outcome of applying this process to the lip image in Figure 1(d), is shown in Figure 2(a).

In order to have a simplified description of the lip shape, we need to extract the basic structure from the clutter of points of Figure 2(a). This was achieved by applying three main steps: 1. Clean image of spurious "stand-alone" pixels, 2. Morphologically *open* the image, 3. Morphologically *close* the image. By morphologically opening the image, we remove any spurious clusters of points not removed by the initial cleaning stage, whilst the closing process fills in any small gaps within the lip shape. The results of these steps are shown in Figure 2(b).

Due to the frequently poor visibility around the inner contour of the mouth, it was decided to simply locate the outer lip contour, and ignore the inner. To allow an edge-detection algorithm to only locate points on the outer lip contour, it was necessary to fill in any gap left in the processed image by the oral cavity. This gap was filled by placing a small box at centroid of all the points and filling any remaining gaps via morphological closure. The resulting image is shown in Figure 2(c).

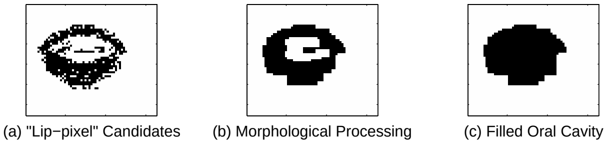


Figure 2: Extraction of Lip Shape

### 3.2. Modelling of Outer Lip Contour

Once the lip shape structure has been formed, we wish to extract the outer edge information only. A standard edge-detection algorithm was applied to the structure of Figure 2(c) to locate the outer edge. This resulted in the set of points shown in Figure 3(a).

To further simplify the description of the outer lip contour, a model was devised consisting of two polynomials. It was found that points on the upper lip could be adequately represented by a $4th$ order polynomial of the form:

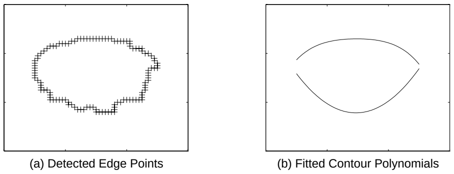
$$y_{upper} = a_4x^4 + a_3x^3 + a_2x^2 + a_1x + a_0 \qquad (2)$$

whilst points on the lower lip could be represented by a $2nd$ order polynomial of the form:

$$y_{lower} = a_2x^2 + a_1x + a_0 \qquad (3)$$

In each case the polynomials were fitted to the data using a *least-squares* approximation algorithm. The resulting data-fit is outlined in Figure 3(b).

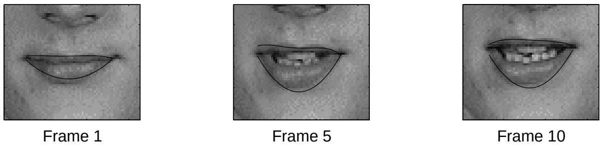
An example of the tracking performance of the system is shown over a number of frames in Figure 4. It can be seen that the polynomial model can cater for a wide range of lip poses, providing a good representation of the outer lip contour.



(a) Detected Edge Points

(b) Fitted Contour Polynomials

Figure 3: Lip Polynomial Model



Frame 1

Frame 5

Frame 10

Figure 4: Lip Tracking Performance

## 4. LIP FEATURE EXTRACTION

We can extract two main sources of information from the lips being shape information and colour information in the surrounding lip areas. Previous work [4], has shown that the majority of speaker dependent information comes from lip intensity information, rather than lip shape. As the shape information we have is limited to the outer lip contour, we decided to base our lip features on colour information only.

### 4.1. Extraction of Colour Information

To derive colour information from the lips in a consistent manner, we chose to take profiles along normals to points placed around the outer lip contour. The profiles were chosen to be of sufficient length so as to cover areas inside the oral cavity and in the skin surrounding the lips. This process is outlined in Figure 5.

### 4.2. Formulation of Colour Feature Vectors

For each point along each profile vector, the *red*, *green* and *blue* colour-component values were stored. All colour values for all profile vectors were then concatenated to form one grand profile vector (GPV) for the image. This step assumes a certain amount of correlation between the colours in each profile vector.

We extracted lower dimensional feature vectors using Principle Component Analysis (PCA) [5] over a set of GPV's from a representative training set covering all speakers. This is performed by finding the eigenvectors corresponding to the largest eigenvalues of the covariance matrix from the training set. Thus if our initial GPV's are of order $n$, and we require our new feature vectors to be of order $t$, then we evaluate:

$$X_{new} = P^T(X_{old} - \bar{X}_{old}) \tag{4}$$

where $\bar{X}_{old}$ is the mean GPV from the training set, and $P$ is an $n \times t$ matrix whose $t$ columns are the eigenvectors corresponding to the largest $t$ eigenvalues.

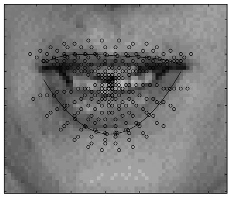


Figure 5: Colour Profile Vectors

### 4.3. Improving Speaker Discrimination in Features

Rather than just letting the classifier determine which features were most speaker dependent, we investigated the use of further processing *prior* to classification to assist in this task. To do this, we employed Linear Discriminant Analysis (LDA) [9]. In LDA we determine a linear mapping function, so as to minimise the intraclass dispersion whilst maximising the interclass distance.

LDA uses two main information sources being the *within-class scatter matrix* and the *between-class scatter matrix.* The within-class scatter matrix $S_w$, describes the scatter of samples about their own class mean, and is evaluated by:

$$S_w = \frac{1}{N_c}\sum_{i=1}^{N_c}\Sigma_i \tag{5}$$

where $\Sigma_i$ is the covariance matrix for class $i$, and $N_c$ is the number of classes. On the other hand, the between-class scatter matrix $S_b$, describes the scatter of class means about the expected vector of the overall distribution, being evaluated by:

$$S_b = \frac{1}{N_c}\sum_{i=1}^{N_c}(M_i - M_0)(M_i - M_0)^T \tag{6}$$

where $M_i$ is the expected vector for each class and $M_0$ is the expected vector of the overall distribution.

It has been proven [9] that the mapping to optimally represent each class, whilst maximising the interclass distance, can be found as the $N_c - 1$ eigenvectors corresponding to the $N_c - 1$ greatest eigenvalues of the matrix:

$$S_w^{-1}S_b \tag{7}$$

In this way lip features are extracted which are most speaker dependent, thus maximizing the resulting discrimination between speaker classes.

## 5. SPEAKER MODELS

Speaker models are based around the use of Gaussian Mixture Models (GMM), or single-stage Hidden Markov Models (HMM). The GMM is a classifier which has received a lot of recent interest for use in speaker identification using the speech signal [10]. The multi-modal nature of the classifier allows it to model a wide variation in voice characteristics for a particular speaker. We choose to use this property of the GMM to allow it to model the wide variation in colour characteristics of a speaker's mouth.

The probability of an observation sequence $o$, belonging to a speaker model $\lambda_i$ is evaluated by:

$$P(o(t)|\lambda_i) = \sum_{n=1}^{N} p_{in},\ (o, \mu_{in}, \Sigma_{in}) \qquad (8)$$

where $p_{in}$ is the mixture weight for mixture $n$ of speaker $i$, and , $(o, \mu, \Sigma)$ is a multivariate Gaussian function with mean $\mu$ and covariance matrix $\Sigma$.

## 6. EXPERIMENTS

We trained and tested the visual recognition system using the M2VTS multi-modal database. This consists of over 27000 colour images of 37 subjects counting from *zero* to *neuf* on five different occasions. We used the first three recording sets as training data, and the fourth set as test data.

One of the key aims of the experiments was to test the performance improvement using LDA features as opposed to features obtained from PCA only. Overall best results were obtained using a GMM with three mixture components. Identification results are presented in Figure 6.

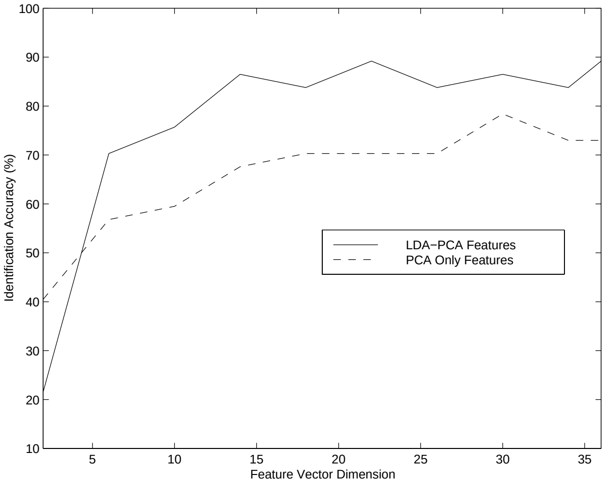


Figure 6: Identification Results with Lip Features

It can be seen that except for the case of very low feature dimensions, the linear discriminant analysis selects features which significantly outperform the identification performance with standard PCA features.

## 7. CONCLUSIONS

We have implemented a speaker recognition system using lip information alone. A novel technique is presented to allow automatic location and tracking of lips using chromatic information. A contour model is derived directly from syntactic information, with no re-estimation procedure necessary.

Colour information in and around the lips is extracted as the basis for speaker dependent feature vectors. Feature vectors are formed through the use of principle component analysis, followed by linear discriminant analysis. We demonstrate in our experiments that discriminant analysis can select features which significantly improve identification performance.

The results from experiments on the M2VTS database are encouraging, and show the importance of lip information for speaker identification. Future research will look at fusing this visual information with speech information, to improve speaker recognition performance in adverse conditions.

## 8. ACKNOWLEDGEMENT

This work was carried out in support of the European Commission ACTS Project M2VTS.

## 9. REFERENCES

[1] S. Pigeon, "The m2vts database," technical report, Laboratoire de Telecommunications et Teledetection, Place du Levant, 2-B-1348 Louvain-La-Neuve, Belgium, (http://www.tele.ucl.ac.be/M2VTS), 1996.

[2] R. Chellappa, C. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proceedings of the IEEE*, vol. 83, pp. 705–740, May 1995.

[3] R. Mammone, X. Zhang, and R. Ramachandran, "Robust speaker recognition - a feature based approach," *IEEE Signal Processing Magazine*, pp. 58–71, Sept. 1996.

[4] J. Luettin, N. Thacker, and S. Beet, "Learning to recognise talking faces," in *Int. Conf. on Pattern Recognition*, (Vienna), 1996.

[5] J. Luettin, N. Thacker, and S. Beet, "Locating and tracking facial speech features," in *Proc. Int. Conf on Pattern Recognition*, vol. I, pp. 652–656, 1996.

[6] M. U. R. Sanchez, J. Matas, and J. Kittler, "Statistical chromacity models for lip-tracking with b-splines," in *Int. Conf. on Audio and Video-based Biometric Person Authtication*, 1997.

[7] T. Coianiz, L. Torresani, and B. Caprile, "2d deformable models for visual speech analysis," in *Speechreading by Humans and Machines*, Springer-Verlag, 1995.

[8] S. Igawa, A. Ogihara, A. Shintani, and S. Takamatsu, "Speech recognition based on fusion of visual and auditory information using full-frame color image," *IEICE Trans. Fundamentals*, pp. 1836–1840, Nov. 1996.

[9] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York: Academic Press, 1972.

[10] D. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, pp. 91–108, 1995.