To appear in Proceedings of the International Conference on Spoken Language Processing, Philadelphia, PA, USA, 1996.

# SPEAKER IDENTIFICATION BY LIPREADING

*Juergen Luettin*[1,2], *Neil A. Thacker*[1], *Steve W. Beet*[1]

[1]Dept. of Electronic and Electrical Engineering
University of Sheffield, Sheffield S1 3JD, UK
[2]IDIAP, CP 592, 1920 Martigny, Switzerland

Luettin@idiap.ch, N.Thacker@shef.ac.uk, S.Beet@shef.ac.uk

## ABSTRACT

This paper describes a new approach for speaker identification based on lipreading. Visual features are extracted from image sequences of the talking face and consist of shape parameters which describe the lip boundary and intensity parameters which describe the grey-level distribution of the mouth area. Intensity information is based on principal component analysis using eigenspaces which deform with the shape model. The extracted parameters account for both, speech dependent and speaker dependent information. We built spatio-temporal speaker models based on these features, using HMMs with mixtures of Gaussians. Promising results were obtained for text dependent and text independent speaker identification tests performed on a small video database.

## 1. INTRODUCTION

Whereas current state-of-the-art speaker recognition systems make use of the acoustic speech signal only, in the speech recognition community it is well known that visual information from lip movements provides additional speech information, which can lead to improved speech recognition performance [1 ]. A difficult problem encountered in speechreading is the modelling of visual speech features which can account for both, speech information and speaker information [2 ]. Speaker recognition research has mainly ignored the visual modality of speech, which we will show, contains discriminative information for speaker recognition.

Face recognition research on the other hand is often based on static face images, assuming a neutral facial expression [3 ]. However, the appearance of a face can change considerably during speech and due to facial expressions. In particular, the mouth is subject to fundamental changes but is also a very important source for discriminating faces. Although research on facial expression analysis has shown promising results, the aim of the methods was to recognise facial expressions rather than to model facial deformation for person recognition. An approach for combining face recognition and speaker recognition has been reported in [4 ]. Face recognition was performed on static images with neutral expressions and speaker recognition was performed on acoustic analysis of isolated digits uttered by the subject.

This paper describes a novel approach for person recognition based on spatio-temporal analysis of the image sequence of the talking face. A deformable shape model is used to track and parameterise the lip boundaries during speech production. A grey level model based on principal component analysis deforms with the shape model and provides intensity information from the mouth area. Both, shape and intensity parameters serve as features to build speaker models based on HMMs with mixtures of Gaussians. Preliminary speaker identification tests have shown high performance on a small database of 12 subjects based on these features.

## 2. FEATURE EXTRACTION

Two important questions to consider are: (I) which image features are important for representing a person and (II) how to model a person with these features. We adopted methods from the face recognition community to extract and model spatial information and methods from the speaker recognition community to model the temporal dependencies of these features.

Since we are only interested in visual information which originates from speech production, we limit the image analysis to the mouth area. Common approaches in face recognition are based on geometric and intensity information. We combine both approaches, assuming that much information about the identity of a speaker is contained in the lip contours and the grey-level distribution around the mouth area. During speech production, the lip shape varies smoothly but still contains speaker dependent information. We try to exploit this fact by building a spatio-temporal model for each speaker which describes the mouth of the speaker and its temporal change during speech production.

We use a shape model to describes the outer and inner lip contour and a deformable grey level model to describe intensity values around the lip contours. The profile model was initially developed for the purpose of image search and only represents a

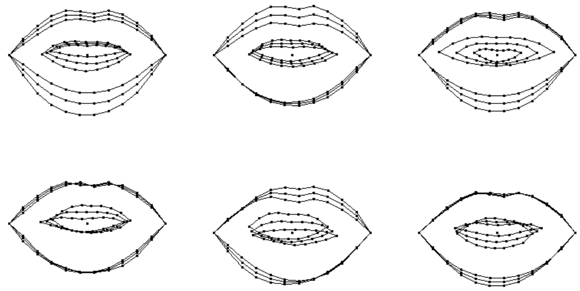

**Figure 1:** First six principal modes of shape variation captured in the training set across all subjects and over all word sequences.

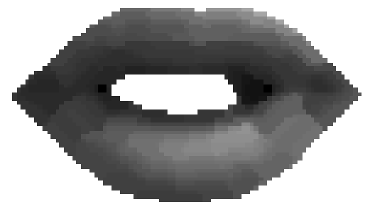

**Figure 2:** Reconstructed lips, using mean shape and mean profile parameters. The profile model also covers the area inside the mouth and a region about half the lip thickness wide extending the outer lip boundary which are not shown here.

small part of the image area. The information contained in the profile vectors might therefore not cover all speaker dependent information, i.g. the sample space does not span the whole mouth area and information about teeth, which might be important, is limited.

An approach based on active shape models [5 ] is used to locate, track and parameterise the lips over an image sequence. These are deformable contour models which represent an object by a set of labelled points. The principal modes of deformation are obtained by performing principal component analysis (PCA) on a labelled training set. Any shape can be approximated by a linear combination of the mean shape and the first few principal modes of deformation.

To model the intensity around the mouth area we use a grey level model which describes intensity vectors perpendicular to the contour at each model point. The concatenated vectors of all model points represent a profile example. In analogy to shape deformation, PCA is performed on all profile examples of a training set to reduce the feature space and to obtain the principal modes of profile variation. Any profile of the training set can be approximated by a linear combination of the mean profile and the first few modes of profile variation. The profile model deforms with the contour model and therefore always represents the same object features.

The profile model is first used to enable robust tracking of the inner and outer lip contour by modelling intensity variation due to different persons and different lighting conditions. The weights for the principal shape and profile modes are recovered from the tracking result and serve as features for the speaker recognition system. These parameters have also been shown to provide important information for visual speech recognition (lipreading) [2]. We have described the detailed feature extraction method elsewhere [6 ].

The first few modes of shape variation are shown in Figure 1. The modes account for variation between speakers and variation due to speech production. Figure 2 shows reconstructed lips with mean shape and mean profile (the lip region was filled in by interpolating the grey levels between the profile vectors). The actual profile model also describes the inside of the mouth and a region extending the outer lip contour.

## 3. SPEAKER MODELLING

Speaker modelling is based on HMMs with mixtures of Gaussians. We use the extracted visual parameters and their temporal dependencies during speech production for representing a particular talking person. The shape and intensity parameters are extracted at each time frame to form a frame dependent feature vector. Although scale might contain important speaker dependent information, we did not use it as feature since it was not possible to estimate absolute scale values from the database we used, which contained only the mouth area. The shape and intensity parameters contain speech dependent information and speaker dependent information, some of the intensity parameters will also contain lighting information.

Speaker recognition tests can be classified into text dependent (TD) and text independent (TI) tasks. For text dependent tasks the test utterance is known while for text independent tasks it is not known. We performed experiments for both, TI and TD mode, where TI mode here is restricted by the size of the vocabulary. The database we used consisted of isolated words, we therefore trained whole word speaker models. For the TD mode, we built one HMM per word class and speaker while for TI mode, one HMM was built per speaker representing all word classes. In the text dependent mode the spoken word is known, therefore only HMMs of the same word class are used for recognition. In TI mode the spoken word is not known and non-specific HMMs representing all word classes are used for recognition.

We trained HMMs which only allow self-loops and sequential transitions between the current and the next state. The initial state probabilities are set to zero for all states but the first. The remaining parameters are estimated from the extracted model parameters of the training set. Each HMM is initialised by linear segmentation of the training vectors onto the HMM states followed by iterative segmental k-means clustering and Viterbi

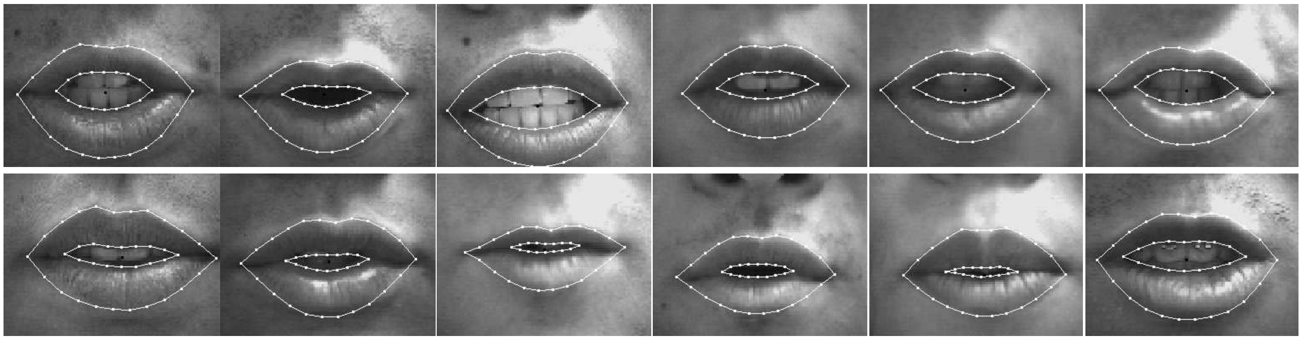

**Figure 3**: Example images of all 12 speakers used for the experiments with lip tracking results.

alignment. The models are further re-estimated using the Baum-Welch procedure. Scores for each subject are calculated using the Viterbi algorithm and classification is performed by estimating the maximum *a posteriori* probability (MAP) and choosing the subject with the highest likelihood as the identified person. All experiments were performed using the HMM toolkit HTK V1.5.

## 4. EXPERIMENTS

Due to the novelty of the approach we were not able to perform tests on a large database. Instead we used the audio-visual Tulips 1 database [7 ] which was recorded for speechreading research. It consists of 96 grey-level image sequences of 12 speakers (9 male, 3 female) each uttering the first four digits in English twice. The images contain only the mouth area of the speakers and are digitised at 30 frames/sec., 100 x 75 pixels and 8 bits/pixel. We used the first utterance of each word and each speaker as the training set for the HMMs and the second instances as the test set. Example images of all speakers are shown in Figure 3.

The performance of acoustic speaker recognition systems is often evaluated as a function of training duration and test duration and typically performance increases if either of those periods is increased. Typical periods used in speaker recognition are 10-30 seconds for training and 3-10 seconds for testing. In our experiments, average training duration was 0.3 seconds for TD tests and 1.3 seconds for TI tests. Average test duration was 0.3 seconds for both tasks. Both periods were therefore considerably shorter than typical periods used in the acoustic domain. Particularly for TD mode, the small size of the database, providing only one example of each word for a given speaker, caused difficulties in estimating the HMM parameters. We tried to alleviated these problems by applying different training and parameter tying methods.

We performed experiments for HMMs with different numbers of states and mixtures. Best results were obtained with HMMs of 4 states and 3 mixture components. The performance generally increased with the number of mixture components but the database only permitted to train a maximum of 3 components. To evaluate the contribution of shape and intensity information towards identification performance, we performed several tests where the feature vector either contained shape parameters or intensity parameters or both. 10 parameters were used for representing the shape and 20 parameters for representing the profile.

### 4.1 Text Dependent Test (TD)

For TD recognition we built a separate HMM for each subject and each word class, resulting in a total of 48 models. Due to the small size of the database we used the following training procedure:

1. Estimating variances for one *global model* using the whole training set.
2. Re-estimating means, mixture weights and transition probabilities for *subject independent word models*.
3. Re-estimating the mean and mixture weights for *subject dependent word models*.

All HMMs have therefore the same variances and the transition probabilities of any word class are tied for all subjects. Only the means and mixture weights are estimated individually for each class of each subject.

Identification is based on the speaker models of the spoken word. Results for TD tests are summarised in Table 1. Best performance was achieved by using both, shape and intensity parameters.

### 4.2 Text Independent Test (TI)

For text independent person recognition we built one HMM for each subject, representing all utterances. The motivation behind this approach is to construct one model which represents different word classes by different mixture components. Parameter estimation was not as critical as in text dependent mode and was performed as follows:

| | Shape | Intensity | Shape + Intensity |
|---|---|---|---|
| TD | 72.9 % | 89.6 % | 91.7 % |
| TI | 83.3 % | 95.8 % | 97.9 % |

**Table 1:** Accuracy for text dependent (TD) and text independent (TI) person identification tests using shape and intensity parameters.

1. Estimating variances, means and mixture weights for one *global model*.
2. Re-estimating mean, mixture weights and transition probabilities for a *text independent speaker model*.

Only the variances are therefore tied for all models. Identification is based on text independent models. Table 1 shows the results for text independent person identification tests. Best performance was also obtained by using both, shape and intensity parameters. Although in acoustic speaker recognition, the performance for TI mode is generally worse than for TD mode, our system performed better for TI mode. This seems to be due to the small training set for TD mode, consisting of only one example per model. For both tasks, performance was higher for intensity parameters than for shape parameters. However, the results for intensity parameters might be misleading, since some of the intensity modes probably account for different illumination. The training procedure might therefore associate a lighting condition with a certain person. The influence of illumination should be investigated on a larger database. One way to increase the robustness to illumination could be to include training images with different illumination or to exclude those intensity modes from the feature vector which are known to account for illumination.

## 5. THE M2VTS PROJECT

The method described in this paper represents a sub-system of the M2VTS project [8]. This is a multi modal person verification project for teleservice and security applications. The aim is to provide a complete solution of secured access to local and centralised services in multi-media environments. The combination of different modalities is expected to improve the performance of unimodal systems and to reduce the acceptance of impostors. The following issues are addressed: face localisation and tracking, facial feature localisation, lip tracking, face recognition, recognition by structured light, profile recognition, speaker recognition and lip motion analysis.

A database was recorded for this project which contains acoustic and visual data. It currently consists of 37 subjects and provides 5 recordings per person. The recordings were taken at one week intervals to account for speech and facial changes. The audio-visual data consists of the first 10 digits continuously spoken. The visual data also contains a sequence of the head rotated by +/- 90 degrees. One of the recordings represents a difficult shot to recognise due to tilted head, closed eyes, different hairstyle etc..

## 6. CONCLUSIONS

A novel approach for speaker identification has been described, based on spatial and temporal analysis of the mouth. Facial features are extracted from image sequences which represent the shape and intensity of the lips. The features are of low dimension and invariant to scale, translation and rotation.

Considering the small training and test duration, results are encouraging and demonstrate that lip information is an important cue for person identification. Further experiments are necessary to evaluate the performance of the method for a large number of subjects and to investigate the benefit of combining it with other approaches like speaker recognition and face recognition.


## ACKNOWLEDGEMENTS

This work has been funded by the University of Sheffield, the German Academic Exchange Service (DAAD) and the European ACTS-M2VTS project.


## REFERENCES


1. E. Petajan, “Automatic Lip Reading to Enhance Speech Recognition”, Proc. IEEE Computer Vision and Pattern Recognition, pp. 44-47, 1985.
2. J. Luettin, N. A. Thacker and S. W. Beet, “Speechreading Using Shape and Intensity Information”, Proc. Int. Conf. on Spoken Language Processing, 1996.
3. R. Chellappa, C. L. Wilson and S Sirohey, “Human and Machine Recognition of Faces: A Survey”, Proc IEEE, vol. 83, no. 5, pp. 705-740, 1995.
4. R. Brunelli and D. Falavigna, “Person Identification Using Multiple Cues” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 17, no. 10, pp. 955-966, 1995.
5. T.F.Cootes, A.Hill, C.J.Taylor and J.Haslam, “Use of active shape models for locating structures in medical images”, Image and Vision Computing, Vol. 12, No. 6, pp. 355-365, 1994.
6. J. Luettin, N. A. Thacker and S. W. Beet, “Locating and Tracking Facial Speech Features”, Proc. Int. Conf. on Pattern Recognition”, 1996.
7. J.R.Movellan, “Visual Speech Recognition with Stochastic Networks”, G.Tesauro, D.Touretzky, T.Leen (eds.) Advances in Neural Information Processing Systems. Volume 7, MIT Press Cambridge, 1995.
8. M. Acheroy et al., “Multi-modal person verification tools using speech and images”, Proc. Europ. Conf. on Multimedia Applications, Services and Techniques, 1996.