To appear in Proceedings of the VIII European Signal Processing Conference, Trieste, Italy, 1996.

STATISTICAL LIP MODELLING FOR VISUAL SPEECH RECOGNITION

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

[1]Dept. of Electronic and Electrical Engineering
University of Sheffield
Sheffield S1 3JD, UK
N.Thacker@shef.ac.uk, S.Beet@shef.ac.uk

[2]IDIAP
CP 592, 1920 Martigny,
Switzerland
Luettin@idiap.ch

ABSTRACT

We describe a speechreading (lipreading) system purely based on visual features extracted from grey level image sequences of the speaker's lips. Active shape models are used to track the lip contours while visual speech information is extracted from the shape of the contours. The distribution and temporal dependencies of the shape features are modelled by continuous density Hidden Markov Models. Experiments are reported for speaker independent recognition tests of isolated digits. The analysis of individual feature components suggests that speech relevant information is embedded in a low dimensional space and fairly robust to inter- and intra-speaker variability.

1. INTRODUCTION

Despite rapid progress in automatic speech recognition over the past decades, the performance levels of most systems degrade significantly in the presence of noise. Many potential application areas such as office, car, aircraft and factory usually contain high levels of noise which often hinders the use of speech recognition systems. Much research effort has therefore been directed to systems for noisy environments [1] and the robustness of speech recognition systems has been identified as one of the biggest obstacles to overcome in future research [2].

Most approaches for robust recognition make use of the acoustic speech signal only and ignore the multi-modal nature of human speech. Psychological studies have shown that besides the acoustic signal, visual information of the speaker's face is often involved in the recognition process [3]. In the presence of noise, visual speech signals can provide information complementary to the acoustic signal and thus improve speech perception [4]. Even if the acoustic signal is undistorted, visual information can lead to more accurate speech perception [5]. In the absence of the acoustic speech signal it has been shown, that individuals with hearing impairments achieved high recognition rates using visual information only [6].

In order to reach their full performance potential, speech recognition systems may have to use several modalities such as those naturally used by humans. Whereas acoustic feature extraction methods are well established, it is not well known which visual features carry important speech information and how to represent them. We describe a speechreading system where speech features are extracted from the lip shape and modelled by Hidden Markov Models (HMMs). We demonstrate that high recognition accuracy can be achieved for speaker independent recognition tests using visual information only.

2. FEATURE EXTRACTION

Speechreading approaches can be classified into three categories: model/geometry based [7], image/pixel-based [8] and optical flow based [9]. Image based systems use the image intensities either directly or after some pre-processing as speech features. In the model based approach a model of the visible speech articulators is defined which is normally described by some geometric measures like height and width of the outer or inner lip contour or by a parameterised contour model. Optical flow based systems assume that speech relevant information is contained in the optical flow information of the mouth area.

2.1 Shape Modelling

Most visual speech information is contained in the shape of the inner and outer lip contour and to a minor extent in the visibility of teeth and tongue [5]. We follow a model based approach where a model of the lips is represented by a deformable shape model. The model is used to locate, track and parameterise lip contours in image sequences. We use an active shape model (ASM) [10] to model the deformation of the lip contours. These are deformable models which represent a contour by a set of points. ASMs make no heuristic

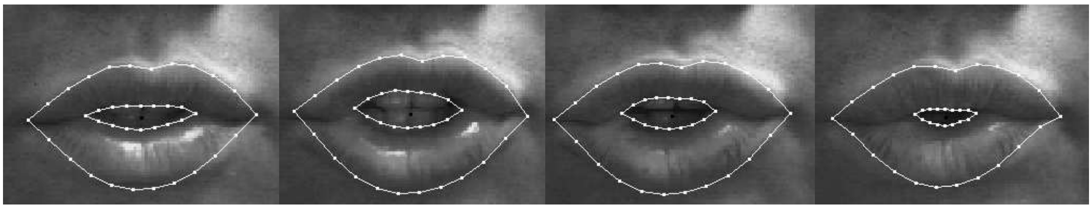

Figure 1: Image sequence of the word *two* with lip tracking results.

assumptions about legal shape deformation. Instead, a priori knowledge about typical deformation is obtained from a training set of labelled lips. The mean shape of the training set is calculated and principal component analysis (PCA) is performed to obtain the principal modes of shape variation. Any shape $\mathbf{x}$, where $\mathbf{x}$ contains the co-ordinates of the model points, can then be approximated with

$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{P}\mathbf{b} \tag{1}$$

where $\overline{\mathbf{x}}$ is the mean shape, $\mathbf{P}$ the matrix of the first few column eigenvectors of the covariance matrix which correspond to the largest eigenvalues and $\mathbf{b}$ a vector containing the weights for each eigenvector. The weights thus describe the shape of the model and are used as input features for the recognition system. This method enables detailed shape description with a small number of parameters. The parameters are linearly independent although non-linear dependencies might be present between the modes.

2.2 Locating and Tracking Lips

Locating and tracking lips in image sequences is a difficult object recognition problem due to the lack of dominant image features representing the lip contours. The contrast at the outer lip contour is often very small and the contrast at the inner lip contour is highly variable due to mouth opening and visibility of teeth and tongue. The problem becomes more difficult if the algorithm is used for several subjects or for different lighting conditions such as those expected for potential applications.

While most previous approaches have used gradients for representing the lip contour, we try to avoid heuristic assumptions and build a model which learns typical intensity information from a training set [11]. We sample one-dimensional intensity profiles perpendicular to the contour at each model point and for each training image. The profiles of all model points of a training image are concatenated to form a global profile vector $\mathbf{h}$. The principal modes of intensity variation captured in the training set are obtained by PCA. In analogy to shape modelling, any intensity profile seen in the training set can now be approximated by

$$\mathbf{h} = \overline{\mathbf{h}} + \mathbf{P}_g\mathbf{b}_g \tag{2}$$

where $\overline{\mathbf{h}}$ is the mean profile vector, $\mathbf{P}_g$ the matrix of the first few eigenvectors of the profile covariance matrix and $\mathbf{b}_g$ a vector containing the weights for the eigenvectors. This method assumes that the deviations of profiles at different profile points are correlated with each other. The motivation for this approach is to build a model which describes the mean intensity profile of the training set and its main modes of variation which originate from different speakers, different lighting conditions and different "mouth states".

The intensity model is used for image search to measure the fit between the model and the image. The model is first placed at an initial position in the image, then the optimal weight vector $\mathbf{b}_g$ is calculated to align the mean profile as closely as possible to the image profile using the first few profile modes. To obtain the cost, we calculate the mean square error between the aligned model profile and the image profile. For locating and tracking the lip contours in an image, the Downhill Simplex Method [12] is used to find a minimum cost. To restrict the model to only deform to shapes similar to the ones in the training set, we constrain each shape and profile weight to stay within +/- 3 standard deviation.

3. VISUAL SPEECH MODELLING

We built two models of the lips, one representing the outer lip contour (Model 1) and one representing the inner and outer lip contour (Model 2). This allows us to evaluate the information of each contour towards recognition performance.

The parameters describing the shape of the lips are extracted at each time frame and used as visual speech features. An advantage of these features compared to image/pixel based methods is their invariance to illumination, scale, rotation and translation. Much speech information is contained in the dynamics of the lip movements rather than the actual shape. Furthermore dynamics of lip movements might be less sensitive to linguistic variability. We therefore performed some recognition tests by including temporal differences of the shape features (delta features).

Similar to acoustic speech modelling, we model visual speech by representing each utterance as a sequence of visual speech vectors. Their emission probabilities are modelled by continuous Gaussian

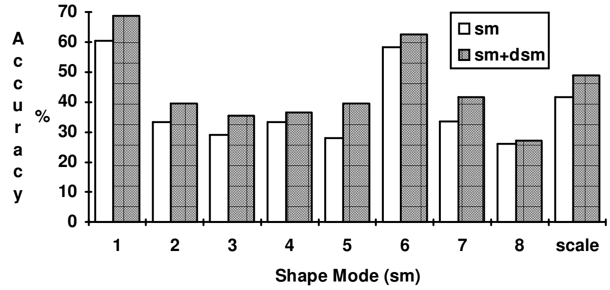

Table 1: Recognition performance of Model 1 for each shape mode (sm) and delta shape mode (dsm).

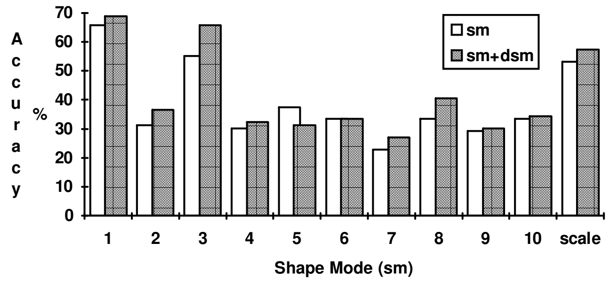

Table 2: Recognition performance of Model 2 for each shape mode (sm) and delta shape mode (dsm).

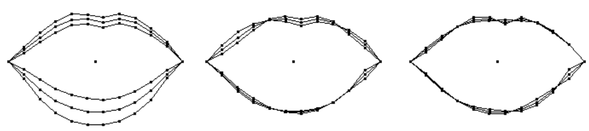

Figure 2: First 3 shape modes (1, 6, 7) of Model 1 with the highest recognition performance (not counting scale).

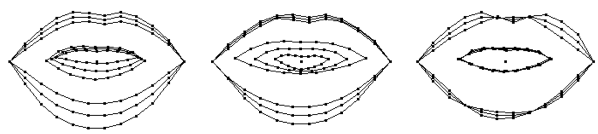

Figure 3: First 3 shape modes (1, 3, 8) of Model 2 with the highest recognition performance (not counting scale).

distributions and temporal changes are modelled by Hidden Markov Models. We used whole-word HMMs and trained one HMM for each word class to be recognised. The models are trained using the Baum-Welch re-estimation algorithm and recognition is performed using the Viterbi algorithm.

The shape features contain some information which contributes to class discriminability and some information which describes inter- and intra-speaker variability (linguistic variability). If we have sufficient training data and if we perform discriminative training, we assume that the recognition network will learn which features contribute to class discriminability and which do not. The database we used was however very small and therefore unlikely to provide enough training material. In order to analyse the contribution of each shape mode to recognition performance, we performed a set of tests where each shape mode was used separately as feature for the recognition system.

4. EXPERIMENTS

We used the Tulips 1 database [13] for our experiments which consists of grey level image sequences of the first four digits. Each digit was spoken twice by 12 individuals (9 male, 3 female). The database reflects a broad variety of speakers and illumination conditions.

One major difficulty in speechreading is caused by the large inter-speaker variability due to different lip shapes and different lip movements during speech production. To see how well the recognition system generalises for new speakers we performed all tests for speaker independent isolated word recognition, using different subjects for training and testing. Because of the small size of the database, recognition tests were performed using the 'jack-knife' or 'leave-one-out' method, i.e. 11 subjects were used for training and the 12th subject for testing. The whole procedure was repeated 12 times, each time leaving a different subject out for testing and the results were averaged over all speakers.

We used HMMs with 6 states and one Gaussian with diagonal variance vector per state, which have previously lead to high recognition performance using this database [14]. We performed separate tests for Model 1 and Model 2. The feature vector was constructed of either all shape features or single shape features or a combination of the first few features with the highest recognition score. All experiments were repeated with delta features added in the feature vector.

Table 1 shows recognition performance for Model 1 where each shape mode and scale were used separately. Most speech information seems to be contained in only about 2 shape modes and scale for the given experiment. The shape modes with the highest performance are shown in Figure 2. The first mode mainly accounts for the shape of the lower lip whereas the 6th and 7th mode seem to account for shape deformation due to lip protrusion.

Similar results were obtained for Model 2, where about 3 shape modes and scale seem to contain most speech information (Table 2). The 3 shape modes are

Coefficients	Single Contour Model	Double Contour Model
sm	61.46 %	64.6 %
sm + dsm	81.25 %	77.1 %

Table 3: Recognition performance using all shape modes (sm) and delta shape modes (dsm).

Coefficients	Single Contour Model	Double Contour Model
sm	69.8 %	69.8 %
sm + dsm	83.3 %	82.3 %

Table 4: Recognition performance using the first four shape mode (sm) with the highest individual performance and corresponding delta shape modes (dsm).

displayed in Figure 3 and seem to describe mainly the shape of the lower lip, the horizontal mouth opening coupled with the lower lip shape and the deformation caused by protrusion.

Recognition performance was higher using the 4 best modes compared to using all shape modes (Table 3, Table 4). Best overall results were obtained with Model 1 rather than Model 2. The slightly lower performance of Model 2 might be due to the small training set, causing inaccurate HMM parameter estimation. Including temporal difference parameters in the feature vector improved recognition results in almost all experiments which confirms the importance of dynamic lip information. Best recognition scores obtained for the given configuration were 83.3 % for Model 1 and 82.3 % for Model 2.

5. CONCLUSIONS AND FUTURE WORK

An approach for visual speech recognition has been described based on lip shape information and their modelling with HMMs. Speaker independent word recognition experiments have demonstrated high recognition performance and analysis of individual shape modes indicate that speech relevant information is contained in only a few modes which are quite robust to inter- and intra-speaker variability.

The main constraint on the experiments was posed by the small size of the database. Although the four word classes were relatively easy to confuse visually, they represent only a small sub-set of all phoneme classes. More experiments on a larger database are desirable to gain more insight into the discrimination ability of shape features for all phonemes.

We are currently working on the extraction of additional speech information from the profile weight vector which might provide information about protrusion and visibility of teeth and tongue. The fact that the extracted features are person dependent is exploited in another experiment in progress, where we evaluate the extracted features for their discrimination ability for the purpose of person verification.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] Y. Gong, “Speech recognition in noisy environments: A survey”, Speech Communication, Vol. 16, No. 3, pp. 261-291, 1995.

[2] R. Cole, L. Hirschman, L. Atlas et al., “The Challenge of Spoken Language Systems: Research Directions for the Nineties”, IEEE Trans. on Speech and Audio Processing, Vol. 3, No. 1, 1995.

[3] H. McGurk and J. MacDonald, “Hearing lips and seeing voices”, Nature, Vol. 264, pp. 746-748, 1976.

[4] K. W. Grant and L. D. Braida, “Evaluating the articulation index for audio-visual input”, J. Acoustic Soc. Am., Vol. 89, pp. 2952-60, 1991.

[5] B. Dodd and R. Campbell (eds.), “Hearing by eye: The psychology of lip-reading”, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1987.

[6] A. MacLeod and Q. Summerfield, “Quantifying the contribution of vision to speech perception in noise”, British J. Audiology, Vol. 21, pp. 131-141, 1987.

[7] E. Petajan, “Automatic Lip Reading to Enhance Speech Recognition”, Proc. IEEE Computer Vision and Pattern Recognition, pp. 44-47, 1985.

[8] C. Bregler and S. M. Omohundro, “Nonlinear Manifold Learning for Visual Speech Recognition”, Pro. Int. Conf. Computer Vision, 1995.

[9] K. Mase and A. Pentland, “Automatic Lipreading by Optical-Flow Analysis”, MIT Media Lab, Perceptual Computing Group, Technical Report #117, 1991.

[10] 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.

[11] J. Luettin, N. A. Thacker and S. W. Beet, “Active Shape Models for Visual Speech Feature Extraction”, D. G. Stork and M. E. Hennecke (eds.), Speechreading by Humans and Machine, NATO ASI Series, Springer Verlag, Berlin, 1995.

[12] W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, “Numerical Recipes in C”, Cambridge University Press, Cambridge, 1992.

[13] J. R. Movellan, “Visual Speech Recognition with Stochastic Networks”, G. Tesauro, D. Touretzky, T. Leen (eds.), Advances in Neural Information Processing Systems 7, MIT Press, Cambridge, 1995.

[14] J. Luettin, N. A. Thacker and S. W. Beet, “Visual Speech Recognition Using Active Shape Models and Hidden Markov Models”, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 1996.