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

SPEECHREADING USING SHAPE AND INTENSITY INFORMATION

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

We describe a speechreading system that uses both, shape information from the lip contours and intensity information from the mouth area. Shape information is obtained by tracking and parameterising the inner and outer lip boundary in an image sequence. Intensity information is extracted from a grey level model, based on principal component analysis. In comparison to other approaches, the intensity area deforms with the shape model to ensure that similar object features are represented after non-rigid deformation of the lips. We describe speaker independent recognition experiments based on these features and Hidden Markov Models. Preliminary results suggest that similar performance can be achieved by using either shape or intensity information and slightly higher performance by their combined use.

1. INTRODUCTION

Visual information of the speaker's face provides speech information which is often complementary to the acoustic signal and which can improve the performance of speech recognition systems [1][2]. One of the main difficulties in speechreading is the extraction of visual speech features. It is still not well known which features are important for speech recognition and how to represent them. Although it is generally agreed that most visual speech information is contained in the inner and outer lip contour, it has also been shown that information about the visibility of teeth and tongue provide important speech cues [3][4]. Particularly for fricatives, the place of articulation can often be determined visually, i.e. for labiodental (upper teeth on lower lip), interdental (tongue behind front teeth) and alveolar (tongue touching gum ridge) place. Other speech information might be contained in the protrusion and wrinkling of lips.

Speechreading approaches can be classified into image-based and model-based systems. Image-based systems use grey level information from an image region containing the lips either directly or after some processing as speech features. Most image information is therefore retained, but it is left to the recognition system to discriminate speech information from linguistic variability and illumination variability. Model-based systems usually represent the lips by geometric measures, like the height or width of the outer or inner lip boundary or by a parametric contour model which represents the lip boundaries. The extracted features are of low dimension and invariant to illumination. Model-based systems depend on the definition of speech related features by the user. The definition may therefore not include all speech relevant information and features like the visibility of teeth and tongue are difficult to represent.

We have previously described a speechreading system [5] based on shape features which represent the outline of the inner and outer lip contour and their modelling by Hidden Markov Models (HMMs). The system performed well for a speaker independent recognition task, but it did not contain any intensity information which might provide additional speech information. Here we extend this system by augmenting the feature vector with intensity information extracted from the mouth region. We evaluate the contribution of intensity information separately and in combination with shape features.

2. SHAPE MODELLING

For modelling the shape variability of lips, we use an approach based on active shape models [6][7]. These are statistically based deformable models which represent a contour by a set of points. Patterns of characteristic shape variability are learned from a training set, using principal component analysis (PCA). The main modes of shape variation captured in the training set can therefore be described by a small number of parameters. The main advantage of this modelling technique is that heuristic assumptions about legal shape deformation are avoided. Instead, the model is only allowed to deform to shapes similar to the ones seen in the training set. Any shape $\mathbf{x}$ representing the co-ordinates of the contour points can be approximated by

$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{Pb}, \tag{1}$$

where $\overline{\mathbf{x}}$ is the mean shape, $\mathbf{P}$ the matrix of eigenvectors of the covariance matrix and $\mathbf{b}$ a vector containing the weights for each eigenvector. Only the first few eigenvectors corresponding to the

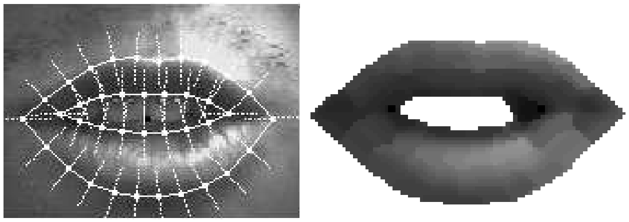

Figure 1: left: shape model for the inner and outer lip contour with profile vectors, perpendicular to the lip contours; right: lip model with mean shape and mean intensity.

largest eigenvalues are needed to describe the main shape variability.

We built and tested two models of the lips: Model 1, which represents the outer lip boundary only and Model 2, which represents the outer and inner lip boundary. The models are used to locate, track and parameterise lip movements in image sequences. The weights for the shape modes are recovered from the tracking results and serve as features for the recognition system.

3. INTENSITY MODELLING

Several approaches for speechreading, based on intensity information have been developed [8 , 9 , 10]. Our approach for extracting intensity information is based on principal component analysis and is related to the 'eigenlips' approach described by Bregler et al. [9] and to the 'local grey-level models described by Lanitis et al. [11]. Bregler et al. placed a window around the mouth area on which PCA was performed. Since the window does not deform with the lips, the eigenvectors of the PCA mainly account for intensity variation due to different lip shape and mouth opening. We already obtain detailed information of the lip shape from our shape model by a small number of parameters and are therefore mainly interested in intensity information which is independent of lip shape. We therefore follow an approach similar to the one described in [11], where one dimensional profiles are sampled perpendicular to the contour at each model point as shown in Figure 1. But instead of using local grey level models we construct a global grey-level model by concatenating the vectors of all model points to form a global intensity vector **h**. We then estimate the covariance matrix of the global profile vectors over the training set and perform PCA to obtain the principal modes of profile variation. Any profile **h** 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, $\mathbf{P}_g$ the matrix of the first column eigenvectors, corresponding to the largest eigenvalues and $\mathbf{b}_g$ a vector containing the weights for each eigenvector. The mean shape and profile of Model 2 is shown in Figure 1.

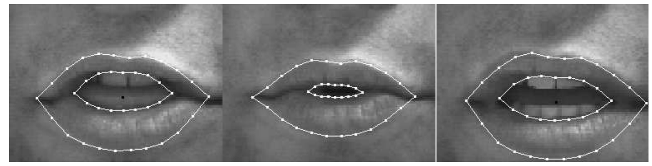

Figure 2: Example images of a person saying the word "three" with tracking results using Model 2.

4. LIP TRACKING: MATCHING THE INTENSITY MODEL TO THE IMAGE

The profile model was initially designed and tailored to enable robust tracking of the lips rather than to extract speech information from the profile vectors. The profile model is used to describe the fit between the image and the model. During image search the model is aligned to the image as closely as possible by calculating the optimal weights for the first few eigenvectors. The mean square error (MSE) between the aligned profile and the image is used as cost and a minimisation algorithm deforms the shape model to find a minimum cost. The profile weight vector for aligning the model is found using

$$\mathbf{b}_g = \mathbf{P}_g{}^T (\mathbf{h} - \overline{\mathbf{h}}) \tag{3}$$

and the cost E is obtained using

$$E = (\mathbf{h} - \overline{\mathbf{h}})^T (\mathbf{h} - \overline{\mathbf{h}}) - \mathbf{b}_g{}^T \mathbf{b}_g. \tag{4}$$

The profile vectors deform with the shape model and therefore always represent the same object features. The weight vector $\mathbf{b}_g$ provides information about the principal modes needed to align to the image. We recover the weights from the tracking results and use them as speech features. A tracking sequence is shown in Figure 2.

5. SPEECH MODELLING

The weights for the shape model and the intensity model are extracted at each image frame to form frame dependent feature vectors for the recognition system. We use either the shape parameters or the intensity parameters or both parameter sets as feature vector for the recognition system. Assuming accurate tracking performance, the shape and intensity parameters are invariant to translation, rotation and scale. The shape parameters are also invariant to illumination. The intensity modes account for both, illumination differences and differences due to the visibility of teeth and tongue and protrusion.

Dynamic speech information is important and often less sensitive to inter speaker variability, i.e. intensity values of the lips will remain fairly constant during speech while intensity values of the mouth opening will change during speech. The intensity values of the lips will vary between speakers but the

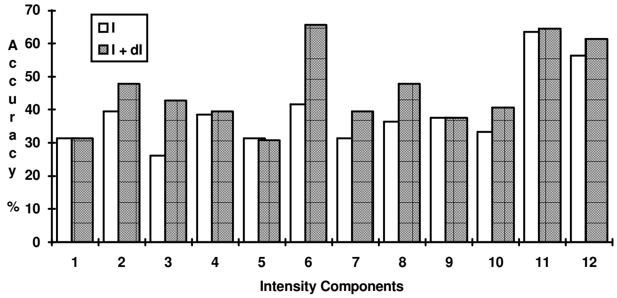

Figure 3: Recognition accuracy for Model 1, using individual intensity features (I) and delta features (dI) for the first 12 principal profile modes.

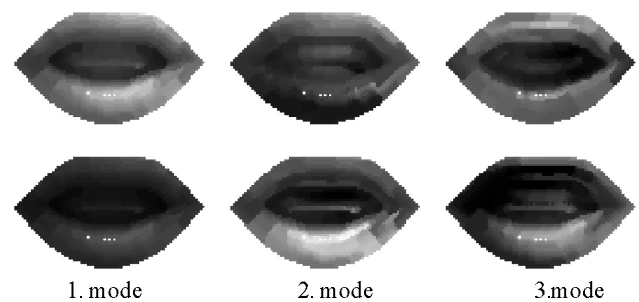

Figure 5: First three principal modes of grey-level variation for Model 2.

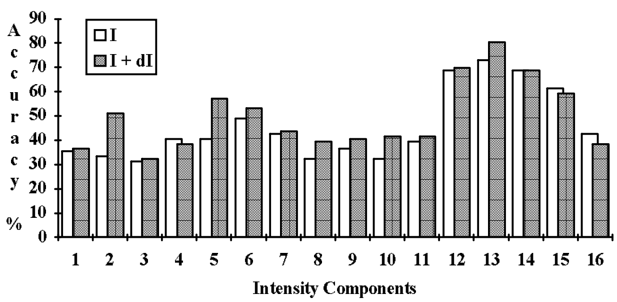

Figure 4: Recognition accuracy for Model 2, using individual intensity features (I) and delta features (dI) for the first 16 principal profile modes.

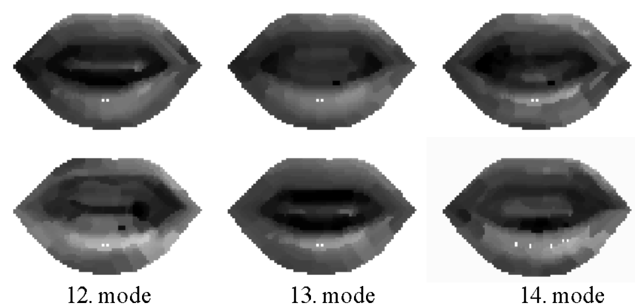

Figure 6: The three profile modes of Model 2 with the highest individual recognition accuracy.

temporal changes of intensity might be similar for different speakers. Dynamic features will therefore be more robust to different illumination and different speakers. We performed one set of tests by including temporal differences of the parameters in the feature vector (delta parameters).

In analogy to acoustic speech recognition we represent an utterance as a sequence of speech vectors. We model the feature vectors with Gaussian distributions and their temporal characteristics with Hidden Markov Models (HMMs). We used whole-word HMMs and trained one HMM for each word to be recognised.

6. EXPERIMENTS

We performed visual speech recognition experiments using the Tulips 1 database [10]. The database consists of grey level image sequences of the first four digits, each spoken twice by 12 subjects, 9 male and 3 female. The images contain the mouth area only and are digitised at 30 fps, 100x75 pixels, 8 bits per pixel. We performed speaker independent tests to see how well the system generalises for new speakers. Due to the small size of the database, the leave-one-out method was used for the tests, 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. Individual results were averaged over all speakers. We trained one HMM per word class with 6 states and one mixture component with diagonal variance vector. The Baum-Welch algorithm was used for testing and the Viterbi algorithm for recognition. All experiments were performed using the HMM toolkit HTK V1.5.

In order to evaluate which components contribute most towards recognition performance, we performed recognition tests by using (i) all parameters, (ii) each parameter individually and (iii) the first few parameters obtained from (ii) with the highest accuracy. Model 1 consisted of 8 shape modes and 10 profile modes, Model 2 consisted of 10 shape modes and 20 profile modes.

7. RESULTS

Recognition rates using all intensity parameters with delta parameters were 78.1 % for Model 1 and 85.4 % for Model 2. These results indicate that the system is quite robust to

Model	Intensity	Shape	Shape + Intensity
Single	82.29	83.33	86.46
Double	89.58	81.25	90.62

Table 1: Recognition results using shape and intensity parameters.

illumination differences which are accounted for by some of the intensity parameters. Figure 3 and Figure 4 display the recognition contribution of each individual component of Model 1 and Model 2, respectively. Results are given for static features (I) and for static and dynamic features (I + dI). It is interesting to note that the single contour model achieved high recognition performance, although it only describes grey-level information near the outer lip boundary. Including delta parameters improved the performance in almost all cases. The first few modes corresponding to the largest eigenvalues make very little contribution towards recognition accuracy for both models.

To visualise the principal modes of grey level variation we simply interpolated the grey levels between the profile vectors to fill in the lip area. The first three principal modes of profile variation for Model 2 are shown in Figure 5. The first mode accounts for global illumination, while the second and third mode mainly seem to account for lighting direction. The second mode also describes the intensity inside the mouth. Figure 6 shows the three modes with the highest single recognition contribution. All three modes seem to account for different illumination inside the mouth.

Table 3 summarises the results using only the first few *best* features for either shape parameters or intensity parameters or both. Delta parameters were included in all experiments. Recognition accuracy for Model 2 is considerably higher for intensity parameters than for shape parameters. This might be due to the additional information captured from the mouth opening. The combination of both feature sets has lead to the best overall recognition accuracy of 90.6 %. This is about equivalent to the performance achieved by humans with no lipreading knowledge which were asked to lipread on the same database [10].

8. CONCLUSIONS

We have described a speechreading system that uses both, shape and intensity information. An important property of the intensity model is that it deforms with the lip contour model in order to represent the same object features after lip movements.

Recognition tests using only intensity parameters indicate that much visual speech information is contained in grey level information which might account for protrusion or visibility of teeth and tongue. Recognition performance was slightly higher for intensity features than for shape features and their combined use outperformed both feature sets.

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. C. Bregler, H. Hild, S. Manke and A. Waibel, “Improved Connected Letter Recognition by Lipreading”, Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, pp. 557-560, 1993.
3. A. A. Montgomery, P. L. Jackson, “Physical characteristics of the lips underlying vowel lipreading performance”, J. Acoust. Soc. Am., Vol. 73, pp. 2134-2144, 1983.
4. Q. Summerfield, “Lipreading and audio-visual speech perception”, Phil. Trans. R. Soc. Lond. B 335, pp. 71-78, 1992.
5. 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.
6. 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.
7. J. Luettin, N. A. Thacker and S. W. Beet, “Locating and Tracking Facial Speech Features”, Proc. Int. Conf. on Pattern Recognition”, 1996.
8. B. P. Yuhas, M. H. Goldstein and T. J. Sejnowski, “Integration of Acoustic and Visual Speech Signals using Neural Networks”, IEEE Communications Magazine, pp. 75-81, 1989.
9. C. Bregler and Yochai Konig, “‘Eigenlips’ for Robust Speech Recognition”, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 669-672, 1994.
10. 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.
11. A. Lanitis, C. J. Taylor and T. F. Cootes, “An Automatic Face Identification system Using Flexible Appearance Models”, British Machine Vision Conf., pp. 65-74, 1994.