AN IMPROVEMENT OF AUTOMATIC SPEECH READING USING AN INTENSITY TO CONTOUR STOCHASTIC TRANSFORMATION

Simon Lucey, Sridha Sridharan and Vinod Chandran
Speech Research Laboratory, RCSAVT
School of Electrical and Electronic Systems Engineering
Queensland University of Technology
s.lucey@qut.edu.au, s.sridharan@qut.edu.au and
v.chandran@qut.edu.au

ABSTRACT: The extraction of lip contour features is difficult and computationally expensive. In this paper we explore the desirable alternative of estimating the contour from area features (ie. mouth grey-scale image) directly via a non-linear stochastic mapping technique. Results are presented on our own speaker dependent database to demonstrate this method and explain why it performs better than previous techniques.

1. INTRODUCTION

Speech Reading is the process of gaining symbolic (ie. viseme or word) meaning from a visual signal of the mouth. The importance of extracting useful area and contour information of the mouth region in multimodal speech processing (MMSP) and especially automatic speech reading is well documented (Chiou and Hwang, 1997; Luettin et al., 1996). Principal component analysis (PCA) has been successfully used to provide a compact representation of both area and contour features. Area features can be computed efficiently using a PCA based technique called 'Eigenlips' first implemented by Bregler and Konig (1994) which compacts a mouth region of interest (ROI) grey-scale image into a few features suitable for a pattern recognition task such as speech reading. Most image information is preserved in these Eigenlip features, but it is left to the recognition system to discriminate speech information from linguistic, translational and illumination variabilities. Contour features describing the labial contour circumvent this problem as they are invariant to illumination and translation variabilities (Luettin et al., 1996; Harvey et al., 1997). Contour extraction is able to pre-emphasize some of the linguistic variabilities present in the mouth ROI thus easing the speech reading task for the recognition engine.

Unfortunately whether one is is using active shape models (Luettin et al., 1996; Lucey et al., 2000b), active contour models (Chiou and Hwang, 1997) or deformable templates (Yuille et al., 1992), tracking lips is a hard and computationally expensive problem. The fitting of a contour to a mouth ROI image may remove the possibility of learning other visual cues significant for speech reading (Harvey et al., 1997) (ie. oral cavity information). The use of area and contour features in conjunction with each other has been shown to be more useful in automated speech reading than using either feature by itself. Another alternative is to discard the parametric contour model and attempt to extract features from the grey-level data directly (Harvey et al., 1997).

In a recent paper (Lucey et al., 2000a) we demonstrated that an effective initial estimate of the outer labial contour can be formulated using a multivariate linear regression mapping from area to contour features. The results presented demonstrate an improvement in speech reading performance from this mapping. The mapping performed a pre-emphaisis of features important to speech reading. In this paper we are extending the idea presented in (Lucey et al., 2000a) to find a mapping function that can perform a similar pre-emphasis such that a physical contour does not need to be fitted reducing computation in MMSP applications significantly.

The approach in this paper uses a non-linear mapping technique called direct estimation (Chen and Rao, 1998). *In this approach, the best estimate of the contour features is derived directly from the joint statistics of pre-tracked contours and their respective area features.* A depiction of this process can be seen in Figure 1. By modeling the joint distribution with Gaussian mixture models (GMMs) a non-linear mapping function can be formulated for estimating contour features from area features. In Section 2

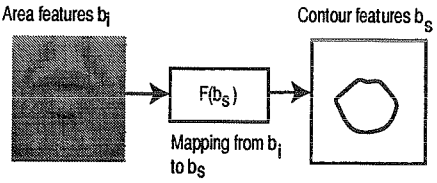


Figure 1: Depiction of contour estimation process from area features.

and 5 background theory behind PCA and direct estimation is given. Sections 3 and 4 describe how the training data for area and contour based features was extracted. Section 6 demonstrates how this new technique compares to conventional area and contour features.

2. PRINCIPAL COMPONENT ANALYSIS

The techniques outlined in this paper rely heavily on PCA to extract features from different feature spaces as well as providing a method for mapping from one feature space to another. So not to confuse between the different feature spaces being used a generalised nomenclature for PCA is defined. A feature space (eg. lip intensity image features, lip contour features) can be approximated by,

$$\mathbf{y} \approx \bar{\mathbf{y}} + \mathbf{P}\mathbf{b} \tag{1}$$

where $\bar{\mathbf{y}}$ is the mean of the training feature vectors, $\mathbf{P}$ 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 vector $\mathbf{b}$ can be used as a compact and decorrelated approximate representation of the original vector $\mathbf{y}$ in which the main modes of variation have been preserved.

3. EIGENLIPS

Eigenlips were first presented by Bregler and Konig (1994) in which cropped lip images were decomposed using PCA into lower dimensional subspace for the purposes of speech recognition. The term *Eigenlips* refers to the first n principal components of a gray-level matrix centered and scaled around the lips and is an extension of the *Eigenfaces* work first developed in (Turk and Pentland, 1991) which dealt with cropped intensity images of the face.

Bregler and Konig (1994) demonstrated that the lip ROI intensity image has important lip information contained in it. They showed that since the window around the lip ROI doesn't deform with the lips, the principle modes of variation are mainly attributed to lighting, skin shade and shape variations. In their experiments they demonstrated that the Eigenlip features used directly in lip reading applications performed well but were highly affected by lighting differences and never quite out performed a contour model of the lips.

Instead of using the lip intensity image directly our work has concentrated on a variant by trying to use the lip intensity image to directly estimate the outer lip contour. In our technique the scale of the face and thus to a certain extent the lips is computed a priori by tracking, the speakers eyes being tracked a priori. Our own speaker dependent database was used for testing due to its completeness. PCA was performed on the set of lip training images after normalising each image as recommended in (Bregler and Konig, 1994).

The localised M by M lip (ROI) can be expressed as a M^2 vector $\mathbf{i}$. Any lip intensity image $\mathbf{i}$, where $\mathbf{i}$ contains the intensity information of the lips, can then be decomposed using PCA so as to extract the principal modes of intensity variation resulting in a feature vector of weights $\mathbf{b_i}$ as modeled in Equation 1.

These modes of variation can be attributed to a number of characteristics such as lighting and more importantly lip shape.

4. EIGEN-DECOMPOSITION OF LIP CONTOURS

To train an effective mapping function from area to contour features we have to first accurately track the labial contour of our own speaker dependent database. To obtain our contours for training we employed our own lip tracking algorithm (Lucey et al., 2000b). This algorithm uses adaptive chromatic thresholding to generate a potential image of the outer labial contour. An active shape model (ASM) is then fitted via a potential force field created through a technique called gradient vector flow (GVF). The use of an ASM parametric contour model fitted via a GVF field makes the lip tracking task robust to lighting and camera noise and variabilities due to colour constancy. PCA was performed on the fitted contours so as to extract the principal modes of contour shape variation resulting in the feature vector of weights $\mathbf{b_s}$ as modeled in Equation 1. Any translational variabilities were removed from the tracked contours before the PCA process such that the centre of the labial contour was always at the origin.

5. DIRECT ESTIMATION

Direct estimation is able to provide a mapping $F(x)$ that gives the best estimate in a least squares sense of contour features from area features from the joint statistics of pre-tracked contour features and their respective area features. This technique has been used previously in bimodal speech analysis and synthesis in the establishment of a mapping between acoustic and mouth shape features Chen and Rao (1998) where it was shown to be superior to other non-linear mapping techniques such as classification based conversion and neural networks. Kain and Stylianou (2000) give a good explanation of the technique when being applied to speech synthesis for the estimation of pitch from the spectral envelope of a speaker.

Direct estimation and multivariate (Lucey et al., 2000a) are mathematically equivalent for a unimodal Gaussian joint density distribution (Kain and Stylianou, 2000). Due to the non-linear nature of the contour fitting process and the small translational variabilities in the area features with respect to the speakers head movement, a linear relationship between area and contor features cannot be assumed. Since direct estimation is a regression technique based on a statistical framework, we can employ a Gaussian Mixture Model (Kain and Stylianou, 2000) (GMM) to model the multimodal joint density distribution of the contour and area features.

A GMM models the probability distribution of a statistical variable z as the sum of Q multivariate Gaussian functions,

$$p(z) = \sum_{i=1}^{Q} \alpha_i N(z; \mu_i, \Sigma_i), \tag{2}$$

where $N(z; \mu, \Sigma)$ denotes a normal distribution with mean vector μ, covariance matrix Σ and α denoting the priori probability of class i. The parameters of the model (α, μ, Σ) can be estimated using the Expectation Maximization (EM) algorithm Dempster et al. (1977).

To model the joint density, we vertically join x, our known vector, and y, the vector we want to estimate, to form

$$z = \begin{bmatrix} x \\ y \end{bmatrix} \tag{3}$$

and estimate the GMM parameters $N(z; \mu, \Sigma)$ for the density $p(z)$, which is the joint density $p(x, y)$.

A locally linear mapping function that attempts to minimize the mean squared between predicted and target vectors is the regression

$$\begin{aligned} F(x) &= E[y|x] = \int yp(y|x)dy \\ &= \textstyle\sum_{i=1}^{Q} h_i(x)[\mu_i^y + \Sigma_i^{yx}(\Sigma_i^{xx})^{-1}(x - \mu_i^x)] \end{aligned} \tag{4}$$

where

$$h_i(x) = \frac{\alpha_i N(x; \mu_i^x, \Sigma_i^{xx})}{\sum_{j=1}^{Q} \alpha_j N(x; \mu_j^x, \Sigma_j^{xx})} \tag{5}$$

Features	Number of features	Recognition (%)
$\mathbf{b_i}$	10	95.30
$\mathbf{b_i}$	20	95.50
$\mathbf{b_s}$	6	97.20
$\mathbf{b_s}$ & $\mathbf{b_i}$	16	97.80
DE for 1 mixture	16	97.89
DE for 2 mixtures	16	98.90
DE for 4 mixtures	16	97.89
DE for 8 mixtures	16	96.34

Table 1: Recognition results for the task for speech reading DE stands for direct estimation with 6 contour features being estimated from 10 area coefficients.

with

$$\Sigma_i = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix}, \mu_i = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix} \quad (6)$$

The mathematical framework described above can be used to estimate the Eigencontour $\mathbf{b_s}$ from Eigenlip features $\mathbf{b_i}$ by substituting for x and y respectively. This mapping technique can then be tested for a number of GMM topologies as will be demonstrated in Section 6.

6. SPEECH READING RESULTS.

We tested this mapping technique on our own speaker dependent database to evaluate its effectiveness. We chose ten digits for the recognition task where the data was collected from a single speaker using SGI 02 workstation and Panasonic VSK0537 digital camera. The recordings had the following characteristics:

- ten digits from 'one' to 'ten';
- each word has 19 examples;
- video captured at 25 fps;
- captured at standard 720x576 PAL resolution;

Using HTK ver 2.2 (Young et al., 1999) the continuous video sequences were automatically transcribed into there respective digits using a HMM recognition technique on the synchronous audio with all silence segments being removed. Each video sequence then had the eyes and mouth manually located with the distance between the eyes being used as a measurement of scale. The mouth region of interest (ROI) for each speaker was extracted with the eye distance being used to normalise for scale thus giving a 140x120 intensity image of the mouth.

Due to the small size of the training/testing set recognition tests were performed using the 'leave-one-out' method ie. Eighteen utterances were used for training and one for testing for each individual digit. The whole procedure was repeated nineteen times. The training of the joint density distributions for the direct estimation mapping also employed the 'leave-one-out' method with a separate mapping function being used for each of the nineteen digit sequences. A left to right two mixture five state HMM was used for the word models as this topology gave best performance for area and joint area, contour speech reading. The recognition results can be seen in Table 1. These recognition results were achieved using static and delta features.

7. DISCUSSION

The counter intuitive result of direct estimation pre-emphasis out performing joint contour and area features can be explained if one inspects Figure 3. This figure presents a plot of the actual and various predicted first principal components of Eigencontour features $\mathbf{b_s}$ for a sample digit sequence. Upon closer scrutiny of Figure 3(b) one can see that the actual contour is quite noisy when being compared to the predicted contours for one and two mixture direct estimation topologies. The smoothing effect apparent in the direct estimated contour features can be attributed to the fact that the noise present in

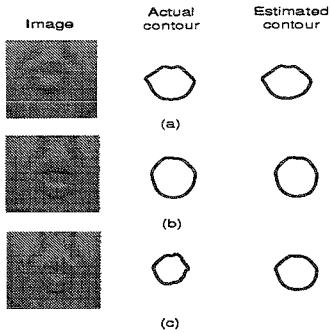
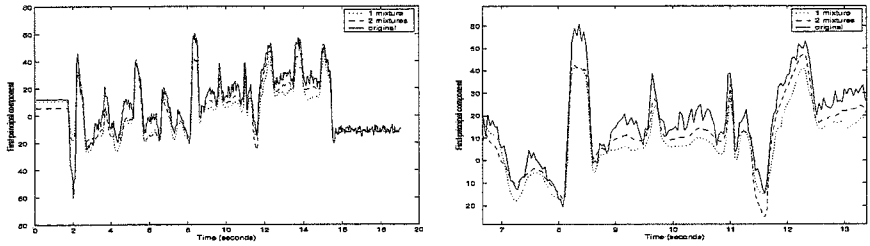


Figure 2: Direct estimation predicts the contour quite accurately as shown in these examples. The predicted contour uses a two mixture joint density topology.



(a) Entire digit sequence (b) Enlarged portion of digit sequence

Figure 3: The first principal component of Eigencontours correlates well with the actual and predicted contours for two direct estimation topologies, as shown above. The frame rate is 25 per second.

the actual contour is due to the contour fitting process itself and not the input image. This contour noise unlike the contour itself is not correlated to the input image so by training a joint density distribution function from a large amount of area and their respective contour features the noise from the contour fitting process is removed. This gives the relatively smooth contour fluctuations from the direct estimation mappings present in Figure 3. Direct estimation has another inherent benefit over normal contour fitting as it is less prone to contour fitting errors as these are seen again by the mapping function as uncorrelated noise. A demonstration of this effect can be seen in Figure 2(c) where the contour fitting algorithm has obviously encountered a tracking error, but the mapping function is able to circumvent the problem.

Finally from Table 1 it was shown that the direct estimation mapping achieved the best results for a system with a two mixture topology. This can be confirmed if one again inspects Figure 3. This result confirms our initial hypothesis that the joint density distribution for area and their respective contour features is not unimodal. It has been demonstrated by Chalmond and Girard Chalmond and Girard (1999) that set geometric shapes (ie. curves) that undergo translational variations and then undergo PCA do not give a unimodal distribution but form what is known in statistical literature 'the horseshoe effect'. Slight translational fluctuations in overall mouth position can be attributed to the non-linearity required for superior performance from the mapping function. Speech reading performance in Table 1 decreases for higher mixtures due to the system over fitting the data and lack of training data for higher order models.

8. CONCLUSIONS

We have described a system for pre-emphasizing mouth ROI area features for speech reading without having to physically fit a contour. Results demonstrated that the non-linear framework of direct estimation was able to account for translational variabilities in area features which can degrade the performance of simpler mapping functions such as multivariate linear regression (Lucey et al., 2000a). The techniques described in this paper provides the framework for a system that is able to outperform traditional MMSP applications which rely on computationally costly contour fitting algorithms.

9. REFERENCES

Bregler, C. and Konig, Y. (1994), Eigenlips for robust speech recognition, *Proc. of IEEE Int. Conf. on Accoustics, Speech and Signal Processing*, Adelaide, Australia, pp. 669–672.

Chalmond, B. and Girard, S. (1999), Nonlinear Modeling of Scattered Multivariate Data and Its Application to Shape Change, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **21**(5), 422–432.

Chen, T. and Rao, R. (1998), Audio-Visual Integration in Multimodal Communication, *Proceedings of the IEEE* **86**(5), 837–852.

Chiou, G. and Hwang, J. (1997), Lipreading from Color Video, *IEEE Transactions on Image Processing* **6**(8), 1192–1195.

Dempster, A., Laird, N. and D. Rubin (1977), Maximum likelihood from incomplete data via the EM algorithm, *Royal Statistical Society* **39**, 1–38.

Harvey, R., Matthews, I., Bangham, J. and Cox, S. (1997), Lip reading from scale-space measurements, *Computer Vision and Pattern Recognition*, Puerto Rico, pp. 582–587.

Kain, A. and Stylianou, Y. (2000), Stochastic Modeling of Spectral Adjustment for High Quality Pitch Modification, *ICASSP'00*, Vol. 2, Istanbul, Turkey, pp. 949–952.

Lucey, S., Sridharan, S. and Chandran, V. (2000a), Initialised Eigenlip Estimator for Fast Lip Tracking Using Linear Regression, *ICPR'2000*, Barcelona, Spain. To appear.

Lucey, S., Sridharan, S. and Chandran, V. (2000b), Robust Lip Tracking using Active Shape Models and Gradient Vector Flow. To appear in Australian Journal of Intelligent Information Processing Systems.

Luettin, J., Thacker, N. and Beet, S. (1996), Speechreading Using Shape and Intensity Information, *International Conference on Spoken Language Processing*, Vol. 1, Philadelphia, USA, pp. 58–61.

Turk, M. and Pentland, A. (1991), Eigenfaces for Recognition, *J. Cognitive Neuroscience* **3**(1), 71–86.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V. and Woodland, P. (1999), *The HTK Book (for HTK version 2.2)*, Entropic Ltd.

Yuille, A., Hallinan, P. and Cohen, D. (1992), Feature Extraction from Faces Using Deformable Templates, *International Journal of Computer Vision* **8**(2), 99–111.