Proceedings of the Institute of Acoustics

LIPREADING USING SHAPE, SHADING AND SCALE

Iain Matthews (1), Tim Cootes (2), Stephen Cox (1), Richard Harvey (1) and J. Andrew Bangham (2)

(1) School of Information Systems, University of East Anglia, Norwich, NR4 7TJ
(2) Department of Medical Biophysics, University of Manchester, Manchester M139PT

1 INTRODUCTION

Two clear paradigms for lipreading, or visual speech recognition, are the high-level, model-based, and low-level, pixel-based, approaches. Between these two extremes is a continuum of possibilities and in this paper we present three different methods in the same experimental context. The first method uses active shape models (ASM's) [6] to track the inner and outer lip contours [11, 12]. This uses a statistical, learned, model of valid lip shapes to constrain the tracking process. An extension to our previous ASM tracking results [12] is the addition of coarse to fine image multiscale tracking. The second method is a recent extension to ASM's, the active appearance model (AAM) [5]. This models both shape and greylevel appearance in a single statistical model, unifying ASM tracking and an eigen analysis of the underlying greylevels. The third method is a pixel-based analysis using a nonlinear multiscale spatial analysis (MSA) to transform images into a more robust scale-space domain that is related to pixel values before extracting features.

Here we compare ASM's, AAM's and MSA on the AVletters database of the isolated letters 'A'–'Z' from ten talkers using a standard implementation of hidden Markov model recognition.

We then compare all of these methods using best weighted late integrated audio-visual recognition, updating our previous results [8].

2 DATABASE

The AVletters database consists of three repetitions by each of ten talkers, five male (two with moustaches) and five female, of the letters A–Z. Each utterance was digitised at quarter frame PAL resolution (376 × 288 at 25fps) using a Macintosh Quadra 660AV in ITU-R BT.601 8-bit headroom greyscale. Audio was simultaneously recorded at 22.05kHz, 16-bit resolution[1]. The mouth images were further cropped to 80 × 60 pixels after locating the centre of the mouth in the middle frame of each utterance. Table 1 summarises the database.

Task	'A'–'Z'
Talkers	10
Repetitions	3
Utterances	780
Frames	18,562
Mouth size	80 × 60
Lighting	studio ceiling

Table 1: Summary of AVletters database.

[1]This database is available on CDROM by contacting the authors.

3 METHODS

3.1 Active Shape Model Tracking

Active shape models are a high-level, model-based, method of extracting lip shape information from image sequences. An active shape model (ASM) is a shape constrained iterative fitting process. The constraint comes from a statistical shape model built from labelled training data. The model compactly describes the space of valid lip shapes, in the sense of the training data, and points in this reduced space are representations of lip shape that can be directly used for lipreading.

To form the model a mean shape, $\bar{\mathbf{x}}$, is calculated from points hand located in 1,144 aligned images and principal component analysis (PCA) applied to identify the directions of the variations about this shape. The inner and outer lip contour are defined using 44 landmark points. Any valid shape, $\mathbf{x}$, in the sense of the training data, can then be approximated by adding the weighted sum of a reduced subset, t, of these modes of variation to the mean shape,

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

where $\mathbf{P}$ is a matrix containing the first t eigenvectors and $\mathbf{b}$ is a vector of t weights.

The order of the point distribution model is chosen so that 95% of the variance of the models is represented in the t modes of variation. The first three modes (out of seven) are shown in Figure 1.

Mode 1 Mode 2 Mode 3

Figure 1: First three modes of variation at ± 2 standard deviations about the mean.

In order to iteratively fit a shape model to an example image some cost function is required that can be evaluated to determine the current goodness of fit. A model of the concatenated gray level profiles of the normals of each point of a shape model is used [11]. This model is formed in the same way as the shape model and called a grey level profile distribution model (GLDM).

$$\mathbf{x}_p = \bar{\mathbf{x}}_p + \mathbf{P}_p\mathbf{b}_p \tag{2}$$

The sum of squares error between the concatenated grey level normals vector and the t_p modes of the GLDM is,

$$\mathbf{R}_p^2 = (\mathbf{x}_p - \bar{\mathbf{x}}_p)^T(\mathbf{x}_p - \bar{\mathbf{x}}_p) - \mathbf{b}_p^T\mathbf{b}_p \tag{3}$$

To locate modelled features the model is placed at an initial location on an image and (3) is iteratively minimised using the simplex algorithm for translation, rotation, scale and shape parameters until convergence. During minimisation shape and grey level profile model parameters are constrained to lie within $\pm 3\sigma$ of the mean.

For this paper we have also used a coarse to fine multiscale image search, initially used for ASM's in [7] but using a point-wise iterative fit rather than a simplex minimisation over parameter space. For multiscale

fitting each training image is successively Gaussian filtered and subsampled a number of times and a set of GLDM's are built, one for each scale. Each example image is likewise subsampled and the search begins at the most coarse scale with the corresponding GLDM and scaled shape model. When converged the next scale image and GLDM are selected until a fit is obtained in the original image. This allows much greater tolerance in the initial parameters: for example, at the coarse scale, a displacement of five pixels is much more significant than at fine scale.

We also examine the use of separate shape models for each talker. When tracking the lips for a given talker their shape model is used in the ASM. This always has fewer modes of variation than the seven of the talker independent shape model so the tracking is a minimisation in a dimensionally smaller search space and should be improved.

However, to avoid training a separate HMM for each talker (which is difficult with a small training set) we attempt to map the low number of talker dependent modes into the talker independent space. This is achieved by minimising the talker dependent mean shape difference and mapping through the 88 dimensional point coordinate space into the seven dimensional talker independent shape space. The limitations of this approach are discussed later.

3.2 Active Appearance Model Tracking

An Active Appearance Model (AAM) fits a statistical model of appearance to a new image using a fast iterative technique [5]. The appearance model is an extension of the statistical shape models described in Section 3.1. It combines a model describing the shape variation of a set of landmark points with a statistical model of the greylevels in the region bounded by the points.

The greylevel texture model is built by warping each training image so that the landmark points lie on the mean shape positions, $\bar{\mathbf{x}}$. This effectively normalises for shape. If we then sample the intensity values from each normalised image into a vector, $\mathbf{g}$, we can compute the mean, $\bar{\mathbf{g}}$, and main modes of variation, $\mathbf{P}_g$, giving a model

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g \mathbf{b}_g \tag{4}$$

The combined appearance model is generated by concatenating the shape parameters, $\mathbf{b}$ and the texture parameters, $\mathbf{b}_g$, and building a similar model of the result, (with modes $\mathbf{Q}$)

$$\begin{pmatrix} \mathbf{b} \\ \mathbf{b}_g \end{pmatrix} = \mathbf{Q}\mathbf{c} \tag{5}$$

Figure 2 shows the effect of varying the first three parameters of an appearance model trained on the lip data.

To match such a model to the image we use the iterative Active Appearance Model algorithm. Given a pose and a set of parameters $\mathbf{c}$, we can project a model estimate into a target image. If we compute the difference, $d\mathbf{g}$, between the model and the image (measured in the shape normalised reference frame) we can use this to update our estimates of the parameters using

$$\mathbf{c} \rightarrow \mathbf{c} - \mathbf{R}d\mathbf{g} \tag{6}$$

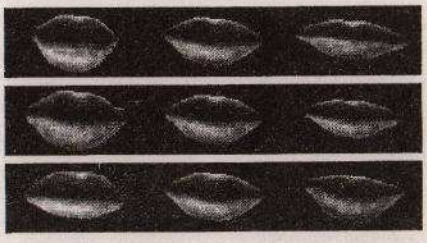
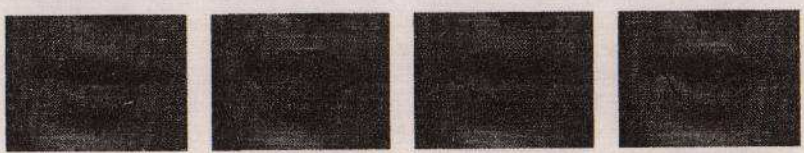


Figure 2: First three modes of variation at ± 2 standard deviations about the mean.

where **R** is a matrix describing the relationship between displacements of the parameters and the difference vector. This can be estimated from the training set (see [5] for details). A similar process can update the estimate of the pose.

To match a model to an image, we simply repeatedly update the current estimate of the pose and parameters until no significant change occurs. For instance, Figure 3 shows frames from an AAM search matching the lip model to an image, given an initial estimate in the centre of the image. In this case only 15 iterations were required to get a good match. The search completes in less than one second on a 166MHz Pentium machine.



Original Initial Model After 4 it.s After 15 it.s

Figure 3: Example of AAM Search.

3.3 Multiscale Spatial Analysis

Multiscale spatial analysis is a low-level, pixel based, method of image analysis. The image is decomposed using a nonlinear scale-space decomposition algorithm called a *sieve* [1]. This is a mathematical morphology serial filter structure that progressively removes features from the input signal by increasing scale, Figure 4 shows this structure. At each stage the filtering element ϕ removes extrema of only that scale. The first stage, ϕ_1, removes extrema of scale 1, ϕ_2 removes extrema of scale 2 and so on until the maximum scale m. The extrema removed are called *granules* and a decomposition into a *granularity* domain is invertible.

A full sieve decomposition of an image retains all of the information in the original image, since the granule amplitudes at a particular position form a partition of the intensity at that position.

We have previously investigated using 2D area sieves to extract the 'blob' associated with the open mouth [9] in a similar approach to the original lipreading work [13]. A more successful approach has been to build *scale histograms* using a 1D length decomposition to find the distribution of image features over scale. A sieve can be applied in 1D to a 2D image by raster scanning vertically.

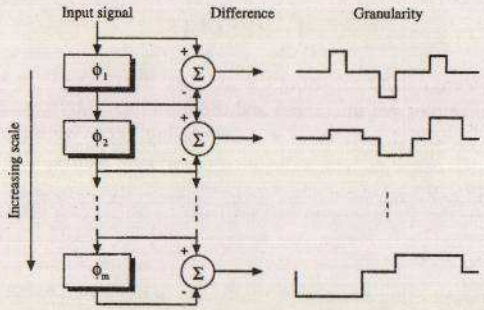


Figure 4: Sieve structure.

A full analysis of the methods used for different types of sieve under different conditions is given in [12]. We find that best results are obtained using a closing sieve, one that processes only negative extrema in the signal, and using the top twenty coefficients of a PCA analysis of the resulting magnitude scale histogram. A magnitude scale histogram is formed by summing the absolute values of the granules at each scale rather than simply counting them. Figure 5 illustrates this process.

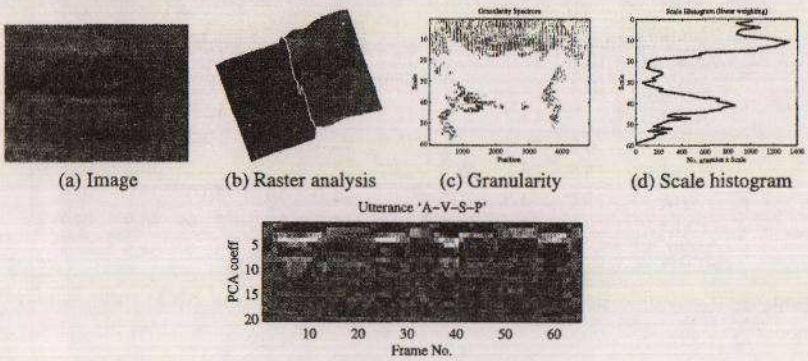


Figure 5: Multiscale spatial analysis. The image (a) is vertically raster scanned, an example cutaway (b) highlights a single slice. The entire image is decomposed by scale (vertical length) into granularity (c) and a scale histogram formed by summing or counting over position at each scale (d). PCA transforms the 60 dimensional scale histogram to form feature vectors of the top 20 directions, (e) shows the result for several frames in sequence.

4 RESULTS

Recognition experiments were performed using the first two utterances from each of the ten talkers as a training set (20 training examples per utterance) and the third utterance from each talker as a test set (10 test examples per utterance). All classification was done using left to right HMM's, each state associated with a one or more Gaussian densities with a diagonal covariance matrix. All HMM's were implemented using the HMM Toolkit HTK V2.1.

4.1 Lipreading Results

All ASM results were obtained using a two stage multiscale fit initialised to the mean shape in the centre of the coarse 40×30 image for each frame. Three model fitting conditions were tested, talker dependent shape models and GLDM's (DD in Table 2), talker independent shape model with talker dependent GLDM's (ID in Table 2) and talker independent shape model and GLDM (II). The best results are obtained using the talker independent shape model with per talker GLDM's. The performance of the talker dependent ASM's mapped to the talker independent space was poor. We attribute this to the large variation between talkers. Mapping low dimensional talker dependent axes is only sensible if the rotations map the same sort of modes of variation onto the same global axes and this cannot be guaranteed with talkers whose lip shapes vary greatly.

States	5			7			9		
Modes	1	3	5	1	3	5	1	3	5
ASM DD	10.8	15.0	20.4	12.7	15.8	21.2	12.3	16.9	23.5
ASM ID	10.4	19.2	21.2	15.8	25.8	24.6	18.5	22.7	**26.9**
ASM II	7.7	13.9	8.9	12.3	13.1	10.0	12.3	11.2	-
AAM 5	16.2	25.4	-	18.9	32.7	31.2	19.2	28.9	-
AAM 10	16.5	28.1	35.4	23.1	33.1	37.3	23.1	36.2	38.1
AAM 20	23.8	33.8	41.5	27.3	35.0	40.8	30.0	36.9	39.6
AAM 37	23.1	32.3	**41.9**	30.0	38.5	39.2	31.9	36.9	38.9
MSA 20	24.6	36.1	41.5	27.3	36.5	40.4	32.7	**44.6**	41.2

Table 2: Recognition accuracy, % correct, for varying number of HMM states and Gaussian modes per state.

The AAM results were obtained by initialising to the mean appearance in the centre of each frame. The four AAM rows of Table 2 differ in the number of model parameters used for recognition. Best results are obtained using all 37 appearance modes but little accuracy is lost by taking only the most significant 20 or 10 modes. The best MSA result from [12] is shown for comparison. The maximum accuracy is very close to that obtained using AAM's.

4.2 Integration Results

We have previously published late integration results using MSA on the AVletters database [8]. Here we update those results with better MSA lipreading performance and compare them with ASM and AAM results.

If we assume that the output of each recogniser is a set of probabilities, one for each of the V vocabulary words, the recognition decision is to choose word w^* where

$$w^* = \arg\max_{i=1,2,\ldots,V} \{\alpha \Pr(w_i|A) + (1-\alpha)\Pr(w_i|V)\} \tag{7}$$

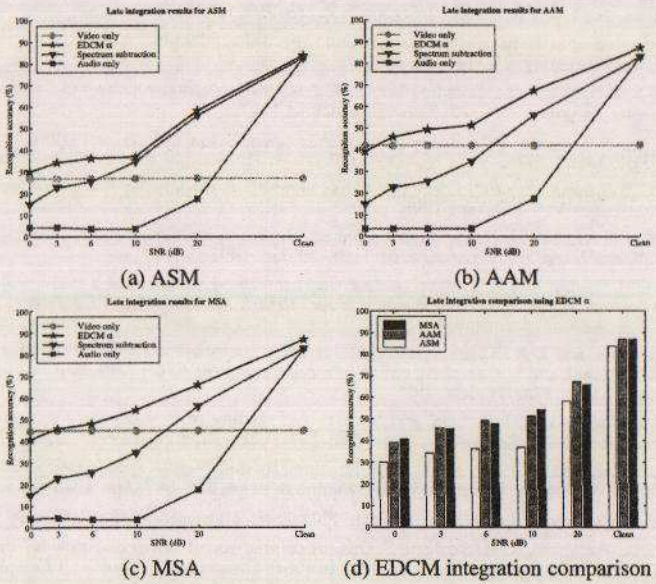
where $\Pr(w_i|A)$ and $\Pr(w_i|V)$ are the respective probabilities of the i'th word from the audio and video recognisers and α is a weighting factor.

To choose α we used a confidence measure based on the uncertainty of the audio recogniser about a word at a given SNR. If the set of legal input words is denoted X and the recognised words Y a possible entropy derived confidence measure (EDCM) estimate for α is,

$$\alpha = 1 - \frac{H(X|Y)}{H(X|Y)_{max}} \tag{8}$$

Previous results [8] suggest that using this estimate of α gives results that are close to those obtained using an exhaustive search to find the best possible value of α at each SNR.

Figure 6(a) plots recognition accuracy over a range of SNR's. Spectrum subtraction [2] is used to improve the audio-only results and as the noise level increases the benefit of adding the best ASM visual recognition can be seen.



(a) ASM (b) AAM

(c) MSA (d) EDCM integration comparison

Figure 6: Late integration results for all methods.

Figure 6(b) plots the same using the best AAM visual information and Figure 6(c) likewise for the best

MSA results. A comparison between ASM, AAM and MSA is shown in Figure 6(d). The results obtained using AAM and MSA are remarkably similar.

5 DISCUSSION

This paper compares three methods of lipreading for visual and audio-visual speech recognition under the same experimental conditions. Pure shape information obtained using an ASM's is not as effective as modelling the combined shape and greylevel surface using AAM's. Active Appearance Models extend previous 'eigenlip' lipreading approaches [3,4] and the results support the assertation that shape information alone does not allow accurate lipreading [3].

In both ASM and AAM methods the errors may be due to a combination of tracking error and modelling error. The use of a predictive temporal tracking framework can be expected to improve tracking performance [10]. Reducing modelling error may require further labelled training data.

The low-level MSA approach requires neither training nor, in this crudely hand aligned application, accurate tracking to deliver identical performance. However, we would expect to improve accuracy by combining with lip tracking to identify the mouth area and normalise for image scale.

References

[1] J. A. Bangham, R. Harvey, P. Ling, and R. V. Aldridge. Morphological scale-space preserving transforms in many dimensions. *Journal of Electronic Imaging*, 5(3):283–299, July 1996.

[2] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27:113–120, 1979.

[3] C. Bregler and S. M. Omohundro. *Learning Visual Models for Lipreading*, volume 9 of *Computational Imaging and Vision*, chapter 13, pages 301–320. Kluwer Academic, 1997.

[4] N. M. Brooke and S. D. Scott. PCA image coding schemes and visual speech intelligibility. *Proc. Institute of Acoustics*, 16(5):123–129, 1994.

[5] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *Proc. European Conference on Computer Vision*, pages 484–498, June 1998.

[6] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models – their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, Jan. 1995.

[7] T. F. Cootes, C. J. Taylor, and A. Lanitis. Active shape models: Evaluation of a multiresolution method for improving image search. In E. Hancock, editor, *Proc. British Machine Vision Conference*, pages 327–336, 1994.

[8] S. Cox, I. Matthews, and J. A. Bangham. Combining noise compensation with visual information in speech recognition. In C. Benoît and R. Campbell, editors, *Proceedings of the ESCA Workshop on Audio-Visual Speech Processing*, pages 53–56, Rhodes, Sept. 1997.

[9] R. Harvey, I. Matthews, J. A. Bangham, and S. Cox. Lip reading from scale-space measurements. In *Proc. Computer Vision and Pattern Recognition*, pages 582–587, Puerto Rico, June 1997. IEEE.

[10] R. Kaucic, B. Dalton, and A. Blake. Real-time lip tracking for audio-visual speech recognition applications. In *Proc. European Conference on Computer Vision*, volume II, pages 376–387, Apr. 1996.

[11] J. Luettin. *Visual Speech and Speaker Recognition*. PhD thesis, University of Sheffield, May 1997.

[12] I. Matthews, J. A. Bangham, R. Harvey, and S. Cox. A comparison of active shape model and scale decomposition based features for visual speech recognition. In *Proc. European Conference on Computer Vision*, pages 514–528, June 1998.

[13] E. D. Petajan. *Automatic Lipreading to Enhance Speech Recognition*. PhD thesis, University of Illinois, Urbana-Champaign, 1984.