# A Comparison of Active Shape Model and Scale Decomposition Based Features for Visual Speech Recognition

Iain Matthews, J. Andrew Bangham, Richard Harvey, and Stephen Cox

School of Information Systems, University of East Anglia, Norwich, NR4 7TJ, UK
Email: {iam,ab,rwh,sjc}@sys.uea.ac.uk

**Abstract.** Two quite different strategies for characterising mouth shapes for visual speech recognition (lipreading) are compared. The first strategy extracts the parameters required to fit an active shape model (ASM) to the outline of the lips. The second uses a feature derived from a one-dimensional multiscale spatial analysis (MSA) of the mouth region using a new processor derived from mathematical morphology and median filtering. With multispeaker trials, using image data only, the accuracy is 45% using MSA and 19% using ASM on a letters database. A digits database is simpler with accuracies of 77% and 77% respectively. These scores are significant since separate work has demonstrated that even quite low recognition accuracies in the vision channel can be combined with the audio system to give improved composite performance [16].

## 1 Introduction

The emerging field known as speechreading is of importance both as a tough problem on which to test generic vision algorithms and also as a problem of considerable value in its own right. It is known that speech recognition systems fail in those poor signal-to-noise conditions that humans manage successful discourse. Furthermore, it is known that a speech reading system with even quite poor performance can provide useful improvements in recognition accuracy under noisy conditions [16].

There is useful information conveyed about speech in the facial movements of a speaker. Hearing-impaired listeners can learn to use lipreading techniques very successfully and are capable of understanding fluently spoken speech. Even for untrained human listeners, being able to see the face of a speaker is known to significantly improve intelligibility particularly under noisy conditions [17, 39]. Likewise the pose of the head affects intelligibility [33]. There is evidence that visual information is used to compensate for those elements the audio signal that are vulnerable in acoustic noise, for example the cues for place of articulation are usually found above 1kHz, and these are easily lost in noise [40]. In practice, some signals which are easily confused in the audio domain (e.g. 'b' and 'e', 'm' and 'n', etc.) are distinct in the visual domain. The intimate relation between the audio and visual sensory domains in human recognition can be seen with audio-visual illusions [30] where the perceiver "hears" something other than what was said acoustically. These effects have even been observed in infants [24].

Early evidence that vision can help speech recognition by computer was presented by Petajan [35]. Using a single talker and custom hardware to quantify mouth opening together with linear and dynamic time warping, he showed that an audio-visual system was better than either alone. Others mapped power spectra from the images [42], or used optic flow [28] and achieved similar results. At around that time a major improvement in audio speech recognition systems emerged with the development of hidden Markov models (HMM's) [25]. HMM's were first applied to visual speech recognition by Goldschen using an extension of Petajan's mouth blob extraction hardware [18]. HMM's were also used for audio-visual recognition with a vector quantised codebook of images and were shown to enhance accuracy in the presence of audio noise [37].

A number of recognition systems which demonstrate improved audio-visual speech recognition compared to audio alone have been reported. As with all recognition systems, the key lies in a good choice of feature space in which to operate. A major problem in generating visual speech features, common to most pattern recognition problems, is that of too much information. Each frame contains thousands of pixels from which a feature vector of between, perhaps, 10 to 100 elements must be extracted. One may categorise ways of reducing the image data to the feature vector ranging from: what might be called a "low level" approach, where features are obtained by direct analysis of the image, for example simple blob analysis, grey scale statistics, etc. and a "high level" approach, where features are obtained by using prior information, such as a model. In practice there is a continuum [21] between these two extremes, but the distinction helps us to show how our approach fits into that framework. Provided the correct model is used then a high level model based system might be expected to be the more robust.

Current high level models either explicitly or implicitly define shape. They take the form of dynamic contours [10,23] deformable templates [13,21] and active shape models [27]. Although there has been considerable success attaching shape models to images of some objects, e.g. [14] and the process looks most attractive, it is not easy to fit them to lips under varying lighting conditions and in real-time. Using blue lipstick chroma-key extraction [1] or small stick-on reflectors [15] makes the process easier but such techniques are useful for research purposes only. As the model tracks the mouth so the parameters required to maintain tracking are used to form the visual feature vector. A particular problem of shape models lies in what exactly to include in the model. There is evidence, for example, that using both the inner and outer lip contours is more effective than just the outer edge [27], but what else should one include? The high level model used in this paper is our implementation of active shape models [14].

Examples of the low level approach include the blob extraction of [35] and the 'eigenlips' approaches of [9,11] in which the greyscale image is subsampled and the principal components accounting for the variance during articulation form the features. A variant we have tried that is designed to reduce the impact of changing lighting conditions is robust blob extraction via an area sieve [4, 8]. A sieve, Sect. 3.2, is used to extract the dark blob representing the mouth aperture using a method analogous to a band-pass filter [19]. The disadvantage is that blob area measurements take little account of shape.

A measure of shape may be obtained by applying a one-dimensional sieve along each of the vertical lines of the mouth image [19]. The effect is to measure the vertical lengths and positions of all the image features, such as the opening between the lips, lip width, etc. at all positions across the mouth. This represents a coding of the mouth shape, teeth, tongue, etc. In other words it is a mapping of the original image [3, 4] with no information reduction. However, in this new domain it turns out that even an unsophisticated data reduction method, such as finding the distribution of these lengths still preserves useful information. This can readily be seen in real-time by watching the histogram change as the shape of the mouth is changed (the algorithm has been implemented at 30 frames per second on a Silicon Graphics O2 workstation).

The high and low level approaches are fundamentally different. The shape models are attractive because they instantiate a model that corresponds closely to our understanding of what we think might be important in lipreading. However, there can be significant problems fitting the models to moving lips under varying lighting conditions and there is an open question on exactly what the shape model should include. On the other hand the low level approach generates a simpler length histogram that is very fast to compute and for which there is evidence that it can robustly reject noise [8]. However, it is hard to see how to introduce prior information into the low level model. One might expect a combination of both methods to be the best solution [9]. In this paper we try to get some intuition into how the high and low level methods compare.

## 2 Databases

In the audio-visual speech community there remains the need for a large standard database on which to build statistically sound models and form comparative results. This is being addressed by, for example, BT Labs [12] and AT&T [36].

In the absence of a standard database each research group has collected their own, invariably small, database. Two easily obtained are the *Tulips* database of isolated digits recorded by Javier Movellan at UCSC [32] and our own *AVletters* database of isolated letters [16, 19, 29]. Here we compare both of these.

The AVletters database consists of three repetitions by each of ten talkers, five male (two with moustaches) and five female (none with moustaches), of the letters A–Z, a total of 780 utterances. Recording took place in the campus TV studio under normal studio ceiling lighting conditions. All recording was of the full face and stored on SVHS quality videotape. The output of a studio quality tie-clip microphone was adjusted for each talker through a mixing desk and fed to the video recorder. Talkers were prompted using an autocue that presented each of three repetitions of the alphabet in a non-sequential, non-repeating order. Each talker returns their mouth to the neutral position. No restraint was used but the talkers do not move out of a close-up frame of their mouth.

Each utterance was digitised at quarter frame PAL resolution ($376 \times 288$ at 25fps) using a Macintosh Quadra 600AV in ITU-R BT.601 8-bit headroom greyscale. Audio was simultaneously recorded at 22.05kHz, 16-bit resolution. This database is available on CDROM by contacting the authors. The mouth images were further cropped to $80 \times 60$ pixels after locating the centre of the mouth in the middle frame of each utterance. Each utterance was hand segmented using the visual data such that each utterance began and ended with the talkers mouth in the neutral position.

The Tulips database contains two repetitions of the digits 1–4 by each of 12 talkers, 9 male and 3 female, a total of 96 utterances. This was recorded using office ceiling lights with an additional incandescent lamp at the side to simulate office working conditions. Talkers were not restrained but could view their mouths and asked not to move out of shot.

The database was digitised at $100 \times 75$ resolution at 30fps using a Macintosh Quadra 840AV in ITU-R BT.601 8-bit headroom greyscale. Audio was simultaneously recorded at 11kHz, 8-bit resolution. This database is available from `http://cogsci.ucsd.edu/~movellan/`. Each utterance was hand segmented so that the video and audio channels extended to one frame either side of an interval containing the significant audio energy. If the lips were clearly moving before or after this time up to an additional three extra frames were included.

Table 1 shows the comparison between both databases.

**Table 1.** Comparison of databases

| Database | Task | Talkers | Reps. | Utts. | Frames | Image size | Lighting |
|---|---|---|---|---|---|---|---|
| AVletters | 'A'–'Z' | 10 | 3 | 780 | 18,562 | $80 \times 60$ | ceiling |
| Tulips | '1'–'4' | 12 | 2 | 96 | 934 | $100 \times 75$ | ceiling & side |

## 3 Methods

### 3.1 Active Shape Models

Active shape models are the application of point distribution models (PDM's) [14] to locate image objects. A point distribution model is defined from the statistics of a set of labelled points located in a set of training images. Examples of the positions of the points used for the AVletters database and the Tulips database are shown as crosses in Fig. 2. Notice that the AVletters database includes two talkers with moustaches and has less direct lighting so does not emphasise the lip contour as much as the Tulips database. Active shape models have been successfully used in visual speech recognition by [26,27]. The implementation described here follows that of [27].

To form the PDM a mean shape, $\overline{\mathbf{x}}$, is calculated from points hand located in 469 images (AVletters) or 223 images (Tulips) and principal component analysis (PCA) applied to identify the directions of the variations about this shape. It is imperative that points are labelled consistently throughout the training set otherwise modes are formed that represent labelling errors. To minimise our labelling error we spline smooth secondary points to be equidistant between a few reliably locatable primary 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 to the mean shape,

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

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 such that 95% of the variance of the models is represented in the first $t$ modes of variation. The first six modes (out of seven) for each of the databases are shown in Fig. 1. The first two modes from AVletters represent the degree of vertical and horizontal mouth opening. These modes are interchanged for the Tulips database. The third mode in both cases alters the 'smile' and the remaining modes account for pose variation and lip shape assymetry. It is gratifying to find the six modes are similar when training independently on two databases.

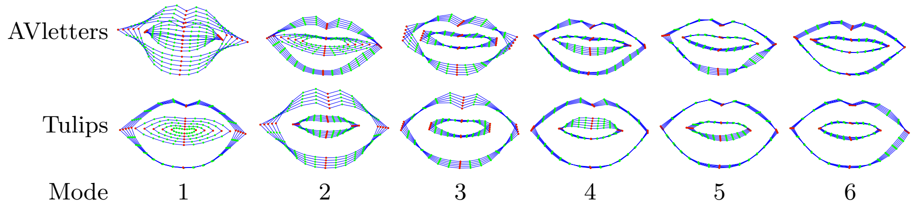


**Fig. 1.** First six modes of variation at $\pm 2$ standard deviations about the mean for both AVletters and Tulips databases. Note, modes 3–6 are the same for both databases and modes 1 and 2 are interchanged.

To actively fit a PDM to any image we require a cost function that can be evaluated in terms of the model weight parameters $\mathbf{b}$ and a rotation, translation and scaling of resulting points. The standard method for ASM's [14] is to build statistical models of the grey levels along the normal of each model point, Fig. 2. In common with [27] we concatenate all the model normals into a single vector and, in analogy to building a PDM, perform PCA to find the mean $\overline{\mathbf{x}}_{\mathbf{g}}$ and $t$ modes of variation of the this concatenated grey level profile vector. This is a grey-level profile distribution model (GLDM).

$$\mathbf{x}_{\mathbf{g}} = \overline{\mathbf{x}}_{\mathbf{g}} + \mathbf{P}_{\mathbf{g}}\mathbf{b}_{\mathbf{g}} \tag{2}$$

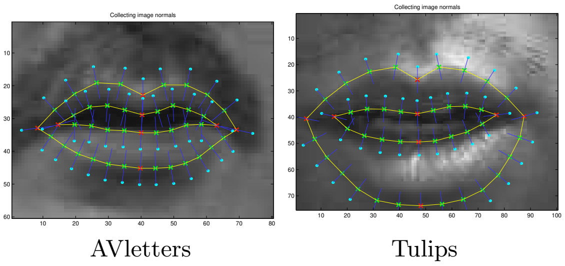


**Fig. 2.** Example point models from AVletters and Tulips with grey level profile normals. The speaker in the AVletters example has a moustache and the lighting does not emphasise the lips.

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

$$\mathbf{R_g^2} = (\mathbf{x_g} - \overline{\mathbf{x}}_\mathbf{g})^\mathbf{T}(\mathbf{x_g} - \overline{\mathbf{x}}_\mathbf{g}) - \mathbf{b_g^T b_g} \tag{3}$$

To locate modelled features the model is placed at an initial location on an image and 3 is iteratively minimised using the simplex algorithm [34] for translation, rotation, scale and model parameters until convergence. During minimisation shape and grey level profile model parameters are constrained to lie within $\pm 3\sigma$ of the mean. In the majority of utterances the converged models fitted well, in a few cases they fitted poorly and in far fewer the fit was bad. Fig. 3 shows some examples.

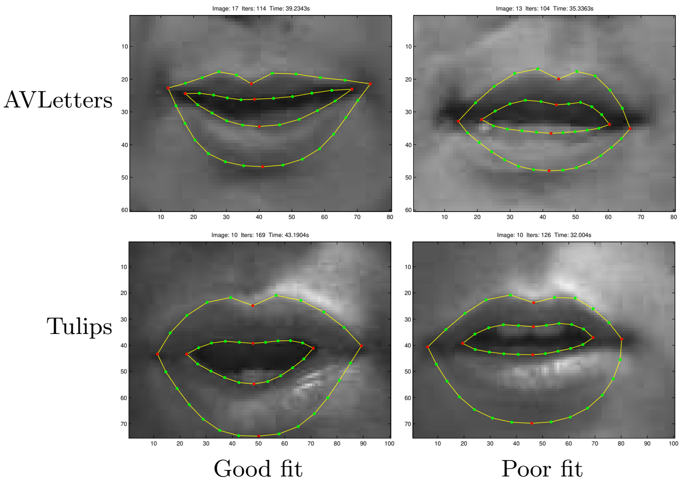


**Fig. 3.** Examples of good and poor ASM fits for AVletters and Tulips.

Observation vectors are formed using the shape model weight parameters **b** obtained after the ASM has converged for each frame. To speed up the fitting process the model is initialised at its previous position for the next frame. Running on an SGI O2 workstation each simplex iteration takes approximately 0.5ms. This can accurately track video at 15 frames per second or faster given lower termination tolerances.

### 3.2 Multiscale Spatial Analysis

The low level method we use has its theoretical roots in mathematical morphology and is similar to granulometry. The system used here is related to alternating sequential filters (formed from openings and closings) and multiscale recursive median filters known as *sieves*. Sieves preserve scale-space causality [4–6] and, like certain wavelets, they can transform the signal to another domain, called granularity, and such a transformation is invertible [3]. The granularity domain can be useful for pattern recognition [2]. Another feature of sieves that is important for lip-reading, lies in the observation that sieves preserve edges well by robustly rejecting random and clutter noise [8].

The sieve may be defined in any number of dimensions by defining the image as a set of connected pixels with their connectivity represented as a graph [20], $G = (V, E)$ where the set of vertices, $V$, are pixel labels and $E$, the set of edges, represent the adjacencies. Defining $C_r(G)$ as the set of connected subsets of $G$ with $r$ elements allows the definition of $C_r(G, x)$ as those elements of $C_r(G)$ that contain $x$.

$$C_r(G, x) = \{\xi \in C_r(G) | x \in \xi\} \tag{4}$$

Morphological openings and closings, over a graph, may be defined as

$$\psi_r f(x) = \max_{\xi \in C_r(G,x)} \min_{u \in \xi} f(u) \tag{5}$$

$$\gamma_r f(x) = \min_{\xi \in C_r(G,x)} \max_{u \in \xi} f(u) \tag{6}$$

The effect of an opening of size one, $\psi_2$, is to remove all *maxima* of area one when working in 2D. In 1D it would remove all maxima of length one. $\gamma_2$ would remove *minima* of scale one. Applying $\psi_3$ to $\psi_2 f(x)$ will now remove all maxima of scale two and so on. The $\mathcal{M}$ and $\mathcal{N}$ operators are defined as $\mathcal{M}^r = \gamma_r \psi_r$ and $\mathcal{N}^r = \psi_r \gamma_r$. Sieves, and filters in their class such as alternating sequential filters with flat structuring elements, depend on repeated application of such operators at increasing scale. This cascade structure is key, since each stage removes maxima and/or minima of a particular scale. The output at scale $r$ is denoted by $f_r(x)$ with

$$f_1 = \mathcal{Q}^1 f = f \text{ and } f_{r+1} = \mathcal{Q}^{r+1} f_r \tag{7}$$

where $\mathcal{Q}$ is one of the $\gamma$, $\psi$, $\mathcal{M}$ or $\mathcal{N}$ operators. Illustrations of sieves and formal proofs of their properties appear elsewhere [4]. The differences between successive stages of a sieve, called *granule functions*, $d_r = f_r - f_{r+1}$, contain non-zero regions, called *granules*, of only that scale.

In one-dimension the graph, (4), becomes an interval

$$C_r(x) = \{[x, x + r - 1] \,|x \in \boldsymbol{Z}\} \tag{8}$$

where $\boldsymbol{Z}$ is the set of integers and $C_r$ is the set of intervals in $\boldsymbol{Z}$ with $r$ elements and the sieves so formed give decompositions by length. It is this that is of importance to lip-reading. The 1D sieve is used to measure the lengths of features seen vertically down the face in the mouth region and these vary as the mouth opens and shuts.

The sieves used in this paper differ in the order in which they process extrema. In 1D the effect of applying an opening of size one, $\psi_2$, is to remove all maxima of length one, an *o*-sieve. Likewise a $\gamma_2$ would remove minima of length one, a *c*-sieve. A 1D alternating sequential filter would remove either maxima and then minima at each, increasing scale, an $N$-sieve, or remove minima and then maxima at each scale an $M$-sieve.

For this lip-reading work, we use a novel variant in which the maxima and minima are removed in a single pass. This is equivalent to applying a recursive median filter at each scale [6]. The sieve so formed is called an $m$-sieve. It inherits the ability to robustly reject noise in the manner of medians and furthermore is much quicker to compute than conventional scale-space preserving schemes.

A granularity is obtained for each image of an utterance, in turn, by applying a one-dimensional sieve along each vertical line in the region of the mouth. A large number granules are obtained and the problem is how to reduce the number of values to manageable proportions. Here, we take the simple step of creating a histogram of granule scales. This is a rough measure of the shape of the mouth. It provides a simple method of substantially reducing the dimensionality from that of the raw image data to the maximum scale used in the sieve. In these examples between 60 and 100 scales are used. The observation vector for the HMM classification is formed by further processing each "scale-histogram".

The simplest form of scale-histogram is obtained by counting the number of granules found at each scale, from 1 to maximum scale and plotting this as a histogram, $sh$. An alternative is to calculate "granule energy" by summing the squared amplitudes, $a^2$. Other alternatives include summing the raw amplitudes, $a$ and the absolute amplitudes, $|a|$, noting that granules can have negative amplitude. Examples of these are shown in Fig. 4. The number of granules of around scale 8 is associated with the mouth being open.

The changes in scale-histogram can be followed over time in Fig. 5 where the $|a|$ histogram is plotted over time. The scale-histogram is plotted as intensity, white represents a large number of granules. The top row is the smallest scale and the bottom the largest. There is a clear association between each word and the pattern formed by the scale-histogram over time. There is a strong analogy between these patterns and spectrograms formed from an audio signal.

Figure 6 shows another example scale-histogram as it evolves over time. The top panel shows four frames from the image sequence of the utterance "D-G-M", the first is a neutral mouth position and the others taken from the centre of

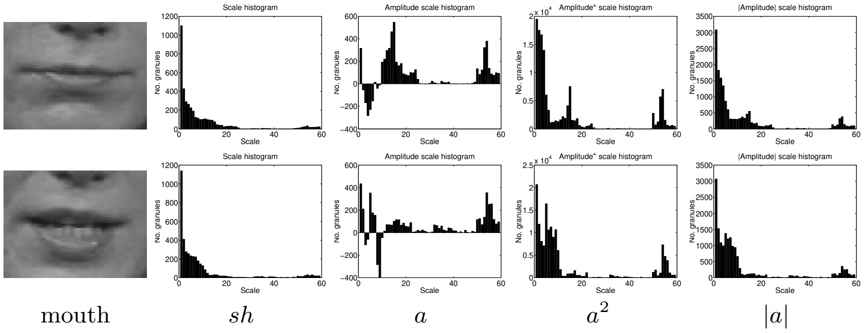


**Fig. 4.** Comparison of scale-histograms for closed, top panel and open, bottom panel, mouths. Abscissa runs from scale 0 to scale 60 and the ordinate shows a function of the number of granules.

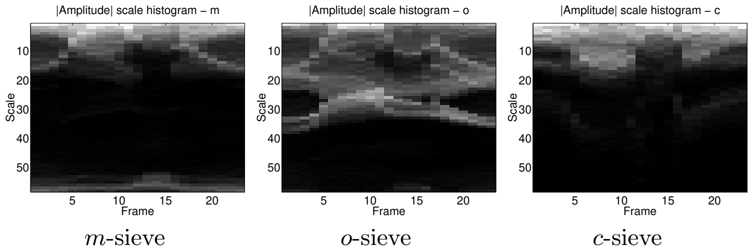


**Fig. 5.** The changes in three different $|a|$ histograms over time observed for the utterance *M*. Intensity is a function of absolute amplitude, the abscissa is time and the ordinate scale with small scale granules shown at the top. Left panel, $m$-sieve, middle panel $o$- and right panel $c$-sieve.

each of the utterances. The scale-histogram clearly changes during articulation and remains stationary between utterances. As expected, motion is present just before and after the acoustic utterance which confirms that visual features can be used to provide audio cues. The dimensionality of the scale-histograms is further reduced to 5, 10, 15 or 20 features by principal component analysis.

## 4 Results

For the AVletters database 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). For the Tulips database recognition was performed using the first utterance from each of the twelve talkers as a training set (12 examples per utterance) and the second utterance from each talker as a test set (12 examples per utterance).

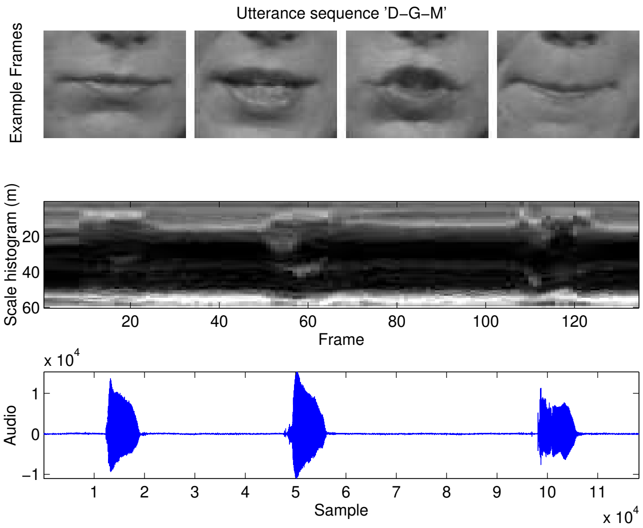


**Fig. 6.** Showing the temporal relationship between the visual information in a scale-histogram (middle panel) and the audio signal (bottom panel). The utterances were *D-G-M*.

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 [41].

### 4.1 Active Shape Models

Building the PDM is laborious since the model needs to be trained by hand placing example points on images until the PDM has converged. In both ASM cases (Tulips and AVletters) the PDM's have converged, adding new data does not significantly alter any of the modes. In both cases two complete utterances from each talker were used (for Tulips it was '1' and '3', 223 images (9812 points) and for AVletters it was 'A' and 'O', 469 images (20636 points) which represents a significant amount of manual input). The grey level distribution models always have a great many modes; for example, about 15 account for 95% of the variance for an individual and 48 modes for a whole-database model on AVletters. This suggests that whole database GLDM's are too general for accurate tracking. In our recognition system, separate GLDM's are used for each speaker to improve the chances of tracking reliably (not practical in a real situation). Equivalent data for the Tulips database is 12 modes for a greyscale model of a single person and 44 modes for all speakers.

Having fitted ASM's to all images we use the shape model weight vector as the observation vector for a HMM. Table 2 shows the recognition accuracy obtained using various HMM model parameters. The best recognition on the Tulips database was obtained using a 9 state HMM with a single Gaussian mode.

This result differs slightly from that shown in [27]. However, here we have used multi-talker training and testing and not the ‘leave-one-out’ or jacknife method. It should be emphasised that although the methods are the same we have used an our own MATLAB implementation of ASM’s and we have reduced the grey level profile lengths to 10 points in order to decrease convergence time. Also, although the number of points in the model is comparable, the actual positions defined during training are different. Given this, results from [27] appear consistent with those shown here.

**Table 2.** ASM recognition accuracies, %, for Tulips and AVletters with variations in the HMM parameters: no. states and no. Gaussian modes per state. Dashes indicate that models could not be trained

| States | 3 | | | 5 | | | 7 | | | 9 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Modes | 1 | 3 | 5 | 1 | 3 | 5 | 1 | 3 | 5 | 1 | 3 | 5 |
| AVletters | 10.8 | 15.0 | 13.5 | 15.8 | 17.7 | 14.2 | 17.3 | **18.8** | 17.3 | 17.7 | - | - |
| Tulips | 56.2 | 56.2 | 47.9 | 58.3 | 56.2 | - | 75.0 | - | - | **76.7** | - | - |

### 4.2 Multiscale Spatial Analysis

There is little collective experience of how one might use either granulometry or granularity to characterise a gesture sequence such as the mouth movements during speaking, nor is there any readily accessible analysis to steer by. We therefore find some ground rules by exploring *all* combinations of the following variables:

1. Figure 5 shows that the type of MSA could affect the result. Test: $m$-sieve, $o$-sieve, $c$-sieve;
2. It is observed that the DC (baseline) component of the raw image affects the result obtained using MSA. Test: preserve DC, ignore DC;
3. In acoustic speech recognition features are typically evaluated faster than video frame rate. Others have found that using temporally interpolated visual features can improve performance. Test: interpolated, non-interpolated;
4. Test: the number of principal components;
5. PCA can be calculated in a square or non-square pattern space. Test: using covariance and correlation matrices;
6. Test: the number of states in the HMM;
7. Test: the number of Gaussian modes per state.

We form features using principle component analysis (PCA) so all that needs to be determined are the eigenvectors of the covariance or correlation matrix. Exploring all the above variables was a lengthy computational task, however, the results show several trends that allow us to dispense with a number of the options and present the interesting results. For example the experiments show that it is generally better to ignore the DC component when using MSA and to use the covariance matrix when calculating the PCA.

It would be expected that most of the information would be associated with the boundary of the dark interior of the mouth. This is most effectively distinguished by a closing granulometry, and very badly characterised by an opening granulometry. We therefore concentrate on results from the $c$ and $m$-sieve, which is bipolar and more robust [8].

The remaining results are summarised in Fig. 7. Using nine states and three Gaussian modes per state are preferred. There also seems to be a slight advantage in using interpolated data. The best results are obtained using the $|a|$ histograms from a $c$-sieve, followed closely by the $m$-sieve. The best results, 44.6% and 40.8%, are obtained with interpolated $|a|$ histograms for $c$ and $m$-sieves respectively.

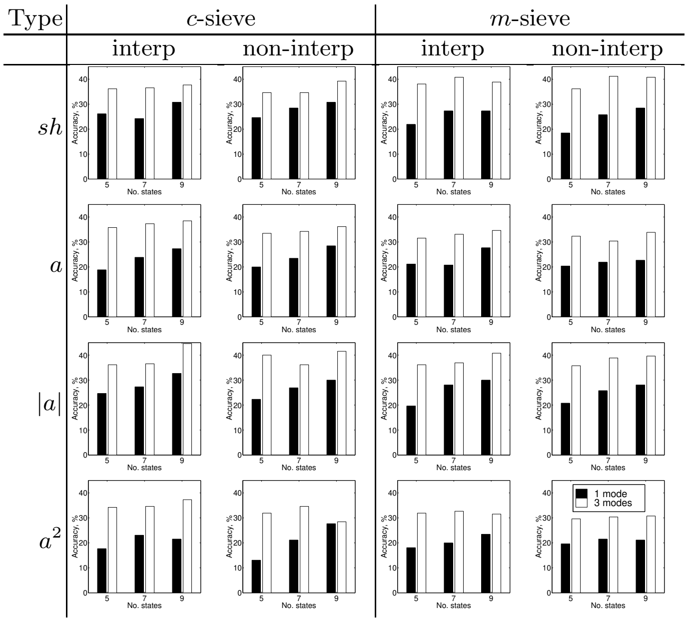


**Fig. 7.** Left $c$-sieve. Right $m$-sieve. Shows how varying the HMM parameters: number of states (abscissa) and Gaussian modes (white columns, 3, black, 1) effects recognition accuracy (ordinate) for interpolated and non-interpolated AVletters data.

The trends in the results shown for the AVletters database are reflected in the results obtained with the Tulips database. Table 3 shows a direct comparison of results obtained using the best MSA options ($|a|$ histogram using a $c$-sieve, ignoring DC, PCA with covariance matrix).

The MSA Tulips result, 77% correct is identical to that obtained using ASM's. However, for the larger and more complex AVletters database the MSA result, 45%, is much higher than the ASM result, 19%.

**Table 3.** MSA recognition accuracies, %, for Tulips and AVletters with variations in the HMM parameters: no. states and no. Gaussian modes per state. Top panel shows results for 10 PCA coefficients, bottom panel for 20 PCA coefficients.

| States | 5 | | 7 | | 9 | |
|---|---|---|---|---|---|---|
| Modes | 1 | 3 | 1 | 3 | 1 | 3 |
| AVletters 10 | 16.5 | 30.8 | 25.4 | 37.7 | 30.0 | 37.3 |
| AVletters 20 | 24.6 | 36.1 | 27.3 | 36.5 | 32.7 | **44.6** |
| Tulips 10 | 66.7 | 54.2 | **77.1** | 58.3 | 75.0 | 72.9 |
| Tulips 20 | 62.5 | 52.1 | 66.7 | 58.3 | 64.6 | 68.7 |

## 5 Conclusion

The results presented here compare two different methods for visual speech recognition. The results suggest that multiscale spatial analysis (MSA) scales better to a larger task than active shape models (ASM's). This might be due to the ASM incorporating inaccurate prejudice as well as good priors or that the lip contour is simply to diffuse to accurately track. Another problem is that as a proportion of the database ASM's are better trained on the smaller Tulips database. It is impracticable to train over a quarter of the (still unrealistically small) AVletters database by hand placing points. Methods to help automate this process are being developed [22].

Results show that the MSA based method is more robust, quicker and more accurate. With multispeaker trials, using image data only, the accuracy is 45% using MSA and 19% using ASM on the letters database. The digits database is simpler with accuracies of 77% and 77% respectively. This is the first time a mathematical morphology based low level method has been compared directly with a high level model based method for the same task. The results show that a low level approach can be very effective, especially when scaling to the more complex letters database. It also has the advantage that it can run in real-time using existing hardware, without consuming all system resources.

A significant omission from the MSA system is a method for normalising the scale. This might be solved when an automatic head tracker is included in the system, such an approach has been implemented elsewhere [9,31]. This suggests significant improvement might be obtained by combining the two methods presented here.

## References


1. A. Adjoudani and C. Benoît. *On the Integration of Auditory and Visual Parameters in an HMM-based ASR*, pages 461–471. In Stork and Hennecke [38], 1996.
2. J. A. Bangham, T. G. Campbell, and R. V. Aldridge. Multiscale median and morphological filters used for 2d pattern recognition. *Signal Processing*, 38:387–415, 1994.
3. J. A. Bangham, P. Chardaire, C. J. Pye, and P. Ling. Mulitscale nonlinear decomposition: The sieve decomposition theorem. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 18(5):529–539, 1996.

4. 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.
5. J. A. Bangham, R. Harvey, P. Ling, and R. V. Aldridge. Nonlinear scale-space from $n$-dimensional sieves. *Proc. European Conference on Computer Vision*, 1:189–198, 1996.
6. J. A. Bangham, P. Ling, and R. Young. Mulitscale recursive medians, scale-space and transforms with applications to image processing. *IEEE Trans. Image Processing*, 5(6):1043–1048, 1996.
7. C. Benoît and R. Campbell, editors. *Proceedings of the ESCA Workshop on Audio-Visual Speech Processing*, Rhodes, Sept. 1997.
8. A. Bosson, R. Harvey, and J. A. Bangham. Robustness of scale space filters. In *BMVC*, volume 1, pages 11–21, 1997.
9. C. Bregler and S. M. Omohundro. Learning visual models for lipreading. In M. Shah and R. Jain, editors, *Motion-Based Recognition*, volume 9 of *Computational Imaging and Vision*, chapter 13, pages 301–320. Kluwer Academic, 1997.
10. C. Bregler, S. M. Omohundro, and J. Shi. *Towards a Robust Speechreading Dialog System*, pages 409–423. In Stork and Hennecke [38], 1996.
11. N. M. Brooke, M. J. Tomlinson, and R. K. Moore. Automatic speech recognition that includes visual speech cues. *Proc. Institute of Acoustics*, 16(5):15–22, 1994.
12. C. C. Chibelushi, S. Gandon, J. S. D. Mason, F. Deravi, and R. D. Johnston. Desing issues for a digital audio-visual integrated database. In *IEE Colloquium on Integrated Audio-Visual Processing*, number 1996/213, pages 7/1–7/7, Savoy Place, London, Nov. 1996.
13. T. Coianiz, L. Torresani, and B. Caprile. *2D Deformable Models for Visual Speech Analysis*, pages 391–398. In Stork and Hennecke [38], 1996.
14. T. F. Cootes, A. Hill, C. J. Taylor, and J. Haslam. The use of active shape models for locating structures in medical images. *Image and Vision Computing*, 12(6):355–366, 1994.
15. P. Cosi and E. M. Caldognetto. *Lips and Jaw Movements for Vowels and Consonants: Spatio-Temporal Characteristics and Bimodal Recognition Applications*, pages 291–313. In Stork and Hennecke [38], 1996.
16. S. Cox, I. Matthews, and A. Bangham. Combining noise compensation with visual information in speech recognition. In Benoît and Campbell [7], pages 53–56.
17. N. P. Erber. Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research*, 12:423–425, 1969.
18. A. J. Goldschen. *Continuous Automatic Speech Recognition by Lipreading.* PhD thesis, George Washington University, 1993.
19. 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.
20. H. J. A. M. Heijmans, P. Nacken, A. Toet, and L.Vincent. Graph morphology. *Journal of Visual Computing and Image Representation*, 3(1):24–38, March 1992.
21. M. E. Hennecke, D. G. Stork, and K. V. Prasad. *Visionary Speech: Looking Ahead to Practical Speechreading Systems*, pages 331–349. In Stork and Hennecke [38], 1996.
22. A. Hill and C. J. Taylor. Automatic landmark generation for point distribution models. In *Proc. British Machine Vision Conference*, 1994.
23. 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, 1996.

24. P. K. Kuhl and A. N. Meltzoff. The bimodal perception of speech in infancy. *Science*, 218:1138–1141, Dec. 1982.
25. S. E. Levinson, L. R. Rabiner, and M. M. Sondhi. An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition. *The Bell System Technical Journal*, 62(4):1035–1074, Apr. 1983.
26. J. Luettin. Towards speaker independent continuous speechreading. In *Proc. of the European Conference on Speech Communication and Technology*, 1997.
27. J. Luettin. *Visual Speech and Speaker Recognition*. PhD thesis, University of Sheffield, May 1997.
28. K. Mase and A. Pentland. Automatic lipreading by optical-flow analysis. *Systems and Computers in Japan*, 22(6):67–75, 1991.
29. I. Matthews, J. A. Bangham, and S. Cox. Scale based features for audiovisual speech recognition. In *IEE Colloquium on Integrated Audio-Visual Processing*, number 1996/213, pages 8/1–8/7, Savoy Place, London, Nov. 1996.
30. H. McGurk and J. McDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, Dec. 1976.
31. U. Meier, R. Stiefelhagen, and J. Yang. Preprocessing of visual speech under real world conditions. In Benoît and Campbell [7], pages 113–116.
32. J. R. Movellan. Visual speech recognition with stochastic networks. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, 1995.
33. K. K. Neely. Effect of visual factors on the intelligibility of speech. *Journal of the Acoustical Society of America*, 28(6):1275–1277, Nov. 1956.
34. J. A. Nelder and R. Mead. A simplex method for function minimisation. *Computing Journal*, 7(4):308–313, 1965.
35. E. D. Petajan. *Automatic Lipreading to Enhance Speech Recognition*. PhD thesis, University of Illinois, Urbana-Champaign, 1984.
36. G. Potamianos, Cosatto, H. P. Graf, and D. B. Roe. Speaker independent audio-visual database for bimodal ASR. In Benoît and Campbell [7], pages 65–68.
37. P. L. Silsbee. *Computer Lipreading for Improved Accuracy in Automatic Speech Recognition*. PhD thesis, The University of Texas, Austin, Dec. 1993.
38. D. G. Stork and M. E. Hennecke, editors. *Speechreading by Humans and Machines: Models, Systems and Applications*. NATO ASI Series F: Computer and Systems Sciences. Springer-Verlag, Berlin, 1996.
39. W. H. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26(2):212–215, Mar. 1954.
40. Q. Summerfield. Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd and R. Campbell, editors, *Hearing by Eye: The Psychology of Lip-reading*, pages 3–51. Lawrence Erlbaum Associates, London, 1987.
41. S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland. *The HTK Book*. Cambridge University, 1996.
42. B. P. Yuhas, M. H. Goldstein, Jr., and T. J. Sejnowski. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, 27:65–71, 1989.