Extraction of Visual Features for Lipreading

Iain Matthews, *Member*, *IEEE*, Timothy F. Cootes, *Member*, *IEEE*,
J. Andrew Bangham, *Member*, *IEEE*, Stephen Cox, *Member*, *IEEE*, and
Richard Harvey, *Member*, *IEEE*

Abstract—The multimodal nature of speech is often ignored in human-computer interaction, but lip deformations and other body motion, such as those of the head, convey additional information. We integrate speech cues from many sources and this improves intelligibility, especially when the acoustic signal is degraded. This paper shows how this additional, often complementary, visual speech information can be used for speech recognition. Three methods for parameterizing lip image sequences for recognition using hidden Markov models are compared. Two of these are top-down approaches that fit a model of the inner and outer lip contours and derive lipreading features from a principal component analysis of shape or shape and appearance, respectively. The third, bottom-up, method uses a nonlinear scale-space analysis to form features directly from the pixel intensity. All methods are compared on a multitalker visual speech recognition task of isolated letters.

Index Terms—Audio-visual speech recognition, statistical methods, active appearance model, sieve, connected-set morphology.

1 Introduction

It has been documented since the 17th century that there is useful information conveyed about speech in the facial movements of a speaker [19]. Hearing-impaired listeners are able to use lipreading techniques very successfully and many are capable of understanding fluently spoken speech. However, even for those with normal hearing, being able to see the face of a speaker is also known to significantly improve intelligibility, especially under noisy conditions [36], [66], [68], [81]. Some speech sounds which are easily confused in the audio domain (e.g., "b" and "v," "m," and "n") are distinct in the visual domain [82], [85]. In addition, there is evidence that visual information is used to compensate for those elements in the audio signal that are vulnerable in acoustic noise, for example, the cues for place of articulation [82]. The intimate relation between the audio and visual sensory domains in human recognition can be demonstrated with audio-visual illusions such as the McGurk effect [57], [62] where the perceiver "hears" something other than what was said acoustically due to the influence of a conflicting visual stimulus.

These observations provide a motivation for attempting to integrate vision with speech in a computer speech recognition system. Early evidence that vision can improve speech recognition was presented by Petajan [70] who used the then current technology of dynamic time-warping with visual features derived from mouth opening and showed that the audio-visual system was better than either speech or vision alone. Others mapped power spectra from static images [89], or used optic flow [58] as visual features and achieved similar results. In the mid-1980's, the development of hidden Markov models (HMM's) [51] improved speech recognition accuracy and made possible large-vocabulary recognition. HMM's were first applied to visual speech recognition by Goldschen using an extension of Petajan's mouth blob extraction hardware [38]. Many approaches have since been applied to visual and audio-visual speech recognition; recent reviews may be found in [22], [39], [44].

The goal is to combine the acoustic and visual speech cues so that recognition performance follows the human characteristic that bimodal results are always better than those from either modality alone [1], [76]. This problem has three parts:

1. speech recognition from an audio signal,
2. identification and extraction of salient visual features, and
3. optimal integration of the audio and visual signals.

The first of these problems, audio speech recognition, is now "solved" to the extent that speech recognition systems that run on personal computers are widely and cheaply available, although their robustness to such factors as different speakers, accents, microphones, interfering channel, or environmental noise, needs to be improved. The second problem, that of extracting visual features from image sequences [9], [15], [17], [33], [39], [44], [48], [55], [58], [65], [71], [79], [87], [89], is the problem addressed here together with the third problem, integration of audio and visual signals [1], [30], [40], [76], [82]. A preliminary report [61] used the "Tulips" database of 96 utterances as opposed to the 780 used here to provide a better chance of discerning differences between the methods under development.

A major problem in generating visual features is the enormous quantity of data in video sequences, a problem common to all computer vision systems. Each video frame contains thousands of pixels from which a feature vector of between about 10 and 100 elements must be extracted.

- *I. Matthews is with the Robotics Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213. E-mail: iainm@cs.cmu.edu.*
- *T.F. Cootes is with the Imaging Science and Biomedical Engineering, University of Manchester, Manchester M13 9PT, UK. E-mail: T.Cootes@man.ac.uk.*
- *J.A. Bangham, S. Cox, and R. Harvey are with the School of Information Systems, University of East Anglia, Norwich, Norfolk, NR4 7TJ, UK. E-mail: {ab, rwh, sjc}@sys.uea.ac.uk.*

Manuscript received 31 Jan. 2001; accepted 13 Apr. 2001. Recommended for acceptance by M. Shah. For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 113560.

Fig. 1. Example frame from each of the 10 talkers in the AVletters database.

Ideally, these features should be robust to such variables as different talkers, head poses, and lighting conditions. One can categorize along a continuum ways of reducing lip image data to a feature vector. At one end of this continuum is a "bottom-up" approach, where features are estimated directly from the image (for example, statistical analysis of pixel intensities, e.g., "eigenlips"). At the other end is a "top-down" approach, where a priori information and assumptions are encapsulated into a model and the features consist of the model parameter values fitted to the image. We would expect the bottom-up approach to avoid systematic model errors and the top-down approach to be more resistant to noise. However, between these extremes lie many possibilities (see [39], [45], for example).

In this paper, we use the same task to evaluate both top-down and bottom-up strategies. The first is an Active Shape Model (ASM) lip contour tracker which uses a (top-down) model of lip shape to constrain the tracker. It has previously been claimed [14], [16] and refuted [49] that shape alone is an insufficient feature for lipreading. Therefore, we extend the ASM to an Active Appearance Model (AAM), which combines the same (top-down) shape analysis used for ASM tracking with a bottom-up statistical model of gray-level appearance. The experiment reported here is run on identical data under the same conditions. It shows that the addition of appearance modeling to shape models significantly improves lipreading performance.

We also present a novel bottom-up approach that uses a nonlinear, scale-space analysis to transform images into a domain where scale, amplitude, and position information are separated. This multiscale spatial analysis (MSA) technique is a fast and robust method of deriving visual features that are not dependent on the absolute amplitude or position of image intensities. This method performs as well as the AAM method and, interestingly, the results suggest that combining the two techniques would lead to an overall performance gain.

2 Databases

A number of "standard" speech databases (e.g., TIMIT, DARPA Resource Management, Wall Street Journal, Switchboard, etc.) have provided important benchmarking information which has been invaluable in the development of audio speech recognition. In the AV community, no such databases exist. Some commercially-funded audio-visual databases have recently been recorded [23], [73] but remain largely untested. The problems of storing and distributing an audio-visual database are significant.

For this work, we recorded our own aligned audio-visual database of isolated letters called *AVletters*. The AVletters database consists of three repetitions by each of 10 talkers, five male (two with moustaches) and five female, of the isolated letters A-Z, a total of 780 utterances.[1]

Talkers were prompted using an autocue that presented each of three repetitions of the alphabet in nonsequential, nonrepeating order. Each talker was requested to begin and end each letter utterance with their mouth in the closed position. No head restraint was used but talkers were provided with a close-up view of their mouth and asked not to move out of frame. Each utterance was digitized at quarter frame of 625 line video (376×288 at 25fps) using a Macintosh Quadra 660AV in 8-bit "grayscale" mode recording only the luma information. As defined in the ITU-R BT.601-4 standard, luma is coded using headroom range [74] which for 8-bit data has the range [16-235]. Audio was simultaneously recorded at 22.05kHz, 16-bit resolution.

The full face images were further cropped to a region of 80×60 pixels after manually locating the center of the mouth in the middle frame of each utterance. The task of automatically finding the region of interest containing the mouth has been discussed elsewhere [63], [69]. Each utterance was temporally segmented by hand using the visual data so that each utterance began and ended with the talkers mouth in the closed position. The audio signal was further manually segmented into periods of silence-utterance-silence. Fig. 1 shows example frames from each of the 10 talkers.

All of the parameterization methods presented here concern feature extraction from these roughly hand-located mouth images from the AVletters database. The mouth image sequences used often show significant motion as they were not accurately tracked and the talkers were not restrained. It is not unreasonable to demand equivalent performance from an automatic system used as a front end.

3 Top-Down Analysis

The top-down, or model-based, approach to lip feature extraction requires some prior assumptions of what are important visual speech features. The shape of the lip contours is often used because the lips are the most prominent features [9], [48], [55]. Here, the first method

1. This database is available, contact J.A.B. at UEA.

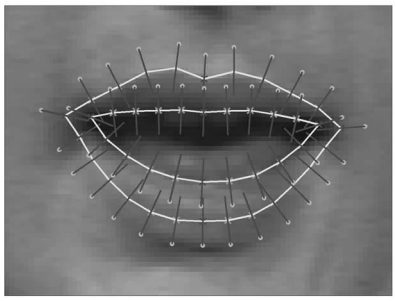

Fig. 2. Inner and outer contour lip model. The white dots represent both the primary and secondary landmarks. Lines indicate normals through each point.

reported uses Active Shape Models (ASM's) to track the inner and outer lip contour and provides a set of control results. ASM's were first formulated in [27], [28] and applied to lipreading by [54], [55]. The second top-down method we use exploits a recent extension of ASM's [26] that combines statistical shape and gray-level appearance in a single, unified Active Appearance Model (AAM).

3.1 Active Shape Models

An active shape model (ASM) is a shape-constrained iterative fitting algorithm [28]. The shape constraint comes from the use of a statistical shape model, also known as a point distribution model (PDM), that is obtained from the statistics of hand-labeled training data. In this application, the PDM describes a reduced space of valid lip shapes, in the sense of the training data, and points in this space are compact representations of lip shape that can be directly used as features for lipreading.

A point distribution model is calculated from a set of training images in which landmark points have been located. Here, landmark points were located by eye, but this may be automated [46]. Each example shape model is represented by the (x, y) coordinates of its landmark points, which have the same meaning in all training examples. The inner and outer lip contour model used is shown in Fig. 2 and has 44 points (24 points on the outer and 20 on the inner contour).

In Fig. 2, the primary landmarks are those that the operator can position reliably and the secondary landmarks are , subsequently, equispaced between the primary points. To reduce positioning errors, the secondary points are smoothed using spline interpolation.

If the ith shape model is

$$\mathbf{x}_i = (x_{i1}, y_{i1}, x_{i2}, y_{i2}, \ldots, x_{i44}, y_{i44})^T,$$

then two similar shapes $\mathbf{x}_1$ and $\mathbf{x}_2$ are aligned by minimizing,

$$E = (\mathbf{x}_1 - M(s,\theta)[\mathbf{x}_2] - \mathbf{t})^T \mathbf{W} (\mathbf{x}_1 - M(s,\theta)[\mathbf{x}_2] - \mathbf{t}), \quad (1)$$

where the pose transform for scale, s, rotation, θ, and translation in x and $y(t_x, t_y)$ is,

$$M(s,\theta)\begin{bmatrix} x_{jk} \\ y_{jk} \end{bmatrix} = \begin{pmatrix} (s\cos\theta)x_{jk} - (s\sin\theta)y_{jk} \\ (s\sin\theta)x_{jk} + (s\cos\theta)y_{jk} \end{pmatrix} \quad (2)$$

$$\mathbf{t} = (t_{x1}, t_{y1}, \ldots, t_{xN}, t_{yN}) \quad (3)$$

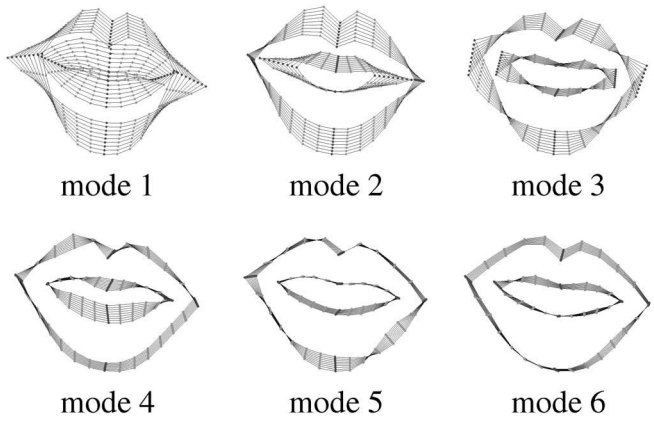

Fig. 3. Point distribution model. Each mode is plotted at $\pm 2\sigma$ about the mean. Seven modes of variation described 95 percent of the variance of the training set of letters A, E, F, M, and O for all 10 talkers.

and $\mathbf{W}$ is a diagonal weight matrix for each point with weights that are inversely proportional to the variance of each point.

To align the set of training models, the conventional iterative algorithm is used [28]. Given the set of aligned shape models, the mean shape, $\overline{\mathbf{x}}_s$, can be calculated and the axes that describe most variance about the mean shape can be determined using a principal component analysis (PCA). Any valid shape can then be approximated by adding a reduced subset, t, of these modes to the mean shape,

$$\mathbf{x}_s = \overline{\mathbf{x}}_s + \mathbf{P}_s \mathbf{b}_s, \quad (4)$$

where $\mathbf{P}_s$ is the matrix of the first t eigenvectors,

$$\mathbf{P}_s = (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_t)$$

and $\mathbf{b}_s$ is a vector of t weights, $\mathbf{b}_s = (b_1, b_2, \ldots, b_t)^T$. As the eigenvectors are orthogonal, the shape parameters $\mathbf{b}_s$ can also be calculated from an example set of points, $\mathbf{x}_s$,

$$\mathbf{b}_s = \mathbf{P}_s^T (\mathbf{x}_s - \overline{\mathbf{x}}_s). \quad (5)$$

This allows valid lip shapes to be represented in a compact, statistically derived shape space. The number of modes of variation is many fewer than the number of landmark points used because the number of landmark points is chosen to clearly define lip shape and they are highly correlated. There are no PCA scaling problems as all variables are either x or y values in square image coordinate axes. The order of the PDM is chosen so that the first t eigenvalues of the covariance matrix describe 95 percent of the total variance.

The top six (out of seven) modes of the PDM calculated from 1,144 hand labeled training images of the AVletters database are shown in Fig. 3. All frames of the first utterances of A, E, F, M, and O for all 10 talkers were used as training data. Each mode is plotted at plus and minus two standard deviations from the mean on the same axes.

To iteratively fit a PDM to an example image, a goodness-of-fit or cost function is needed. Here, a statistical model of the concatenated gray-level profiles from the normals of each point of a shape model is used [43], [54], [55]. This allows PCA to represent all the gray-level normals

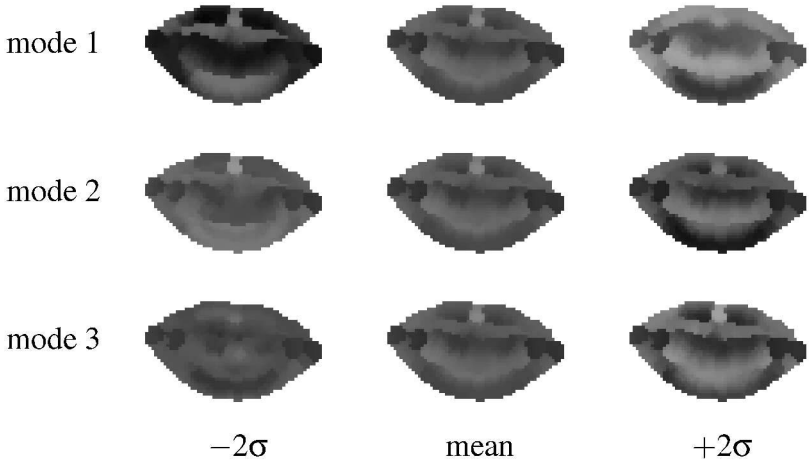

Fig. 4. First three modes of the GLDM. Each mode is plotted at $\pm 2\sigma$ about the mean.

of the shape model with a single statistical model, and so account for correlation between the gray-level profiles at different points. Fig. 2 plots normals in 11 pixels length about each model point. In this example, the concatenated gray-level profiles form a $44 \times 11 = 484$ length vector. In the same way that PCA was used to calculate the PDM, a gray-level distribution model (GLDM) can be calculated,

$$\mathbf{x}_p = \overline{\mathbf{x}}_p + \mathbf{P}_p \mathbf{b}_p. \tag{6}$$

The order of the GLDM is also chosen such that t modes describe 95 percent of the variance. For the AVletters database, the GLDM has 71 modes.

The first three modes of the GLDM are shown in Fig. 4 at ± 2 standard deviations about the mean. The GLDM models only single pixel width normals at each landmark point. To aid visualization, the profiles have been widened and the image smoothed to give the appearance of a full mouth image (see Fig. 4).

As with the PDM (5), the vector of model weights for a given concatenated gray-level profile vector may be calculated,

$$\mathbf{b}_p = \mathbf{P}_p^T (\mathbf{x}_p - \overline{\mathbf{x}}_p). \tag{7}$$

The original ASM algorithm [27], [28] models gray-level profiles for each individual landmark point and iteratively fits a particular image by calculating the model update on a point-wise basis. Here, a simpler fitting algorithm is used. The combined pose and shape parameters

$$(t_x, t_y, s, \theta, b_1, b_2, \ldots, b_t)$$

form the variables for a downhill simplex function minimization [54], [55]. The simplex algorithm [67] does not require calculation of the gradient of the error surface but may require many iterations to converge to a local minimum.

The cost function used in this algorithm calculates sum of squares error of the GLDM and is a measure of how well the gray-level profiles about the current model points match those seen in the training set of hand-located points. The cost function is evaluated for each point in the simplex at each iteration and ideally has a minimum only at the correct model shape and position.

The weight parameters $\mathbf{b}_p$ can be calculated for a particular concatenated profile vector $\mathbf{x}_p$ using (7) to find the best approximation to the current concatenated gray-level profile given the GLDM (6). There is some error introduced due to the approximation, $\hat{\mathbf{x}}_p$, using only t_p modes of the GLDM,

$$\mathbf{e} = \mathbf{x}_p - \hat{\mathbf{x}}_p = (\mathbf{x}_p - \overline{\mathbf{x}}_p) - \mathbf{P}_p \mathbf{b}_p \tag{8}$$

and the sum of squares error between the model and the profile is

$$E^2 = (\mathbf{x}_p - \overline{\mathbf{x}}_p)^T (\mathbf{x}_p - \overline{\mathbf{x}}_p) - \mathbf{b}_p^T \mathbf{b}_p. \tag{9}$$

The fit process was initialized using the mean shape in the center of the image with zero rotation and unity scale. The simplex was initialized as a perturbation from this position by a translation of five pixels in both x and y directions, rotationally by 0.1 radians, with a 10 percent scale increase and by 0.5 of a standard deviation for each of the seven modes of variation of the PDM. Convergence is obtained when the ratio of the cost function at the maximum and minimum points in the simplex is less than 0.01. Only the shape parameters of the simplex minimized pose and shape vector are used as lipreading features. Fig. 5 plots the directions in each of the seven modes of variation of the PDM for the tracking results on the letter sequence D-G-M.

3.1.1 Per Talker Modeling

The large variation between the appearance of the talkers in the database, Fig. 1, means the GLDM's trained over the entire database have a great many modes. The cost function (9) evaluated in such a high-dimensional space is unlikely to have a clear minimum and the simplex local search will be unable to find the correct pose and shape of the talkers lips.

A solution is to build separate GLDM's for each talker. These are required to model the variance of the speech of only a single talker and are much more compact. A set of GLDM's can be built, one for each talker, k,

$$\mathbf{x}_p^k = \overline{\mathbf{x}}_p^k + \mathbf{P}_p^k \mathbf{b}_p^k. \tag{10}$$

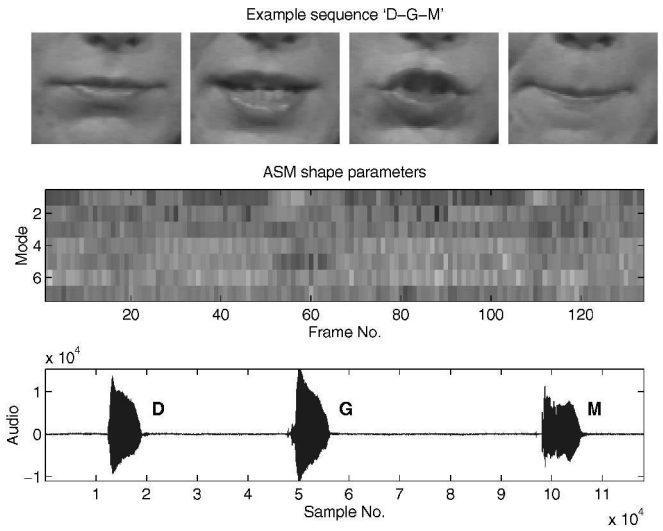

Fig. 5. ASM tracked sequence. The shape parameters, the directions in each of the seven modes of the PDM, are plotted as intensity for the sequence of isloated letters, "D-G-M." The top row shows example images and the bottom row is the aligned audio waveform.

When fitting to an image, the correct GLDM is chosen for the talker, which requires a priori knowledge of the identity of the talker. In practice, it might be possible to automatically select a GLDM by evaluating several and finding which has the lowest cost function. For all experiments using these *local* GLDM's, the identity of the talker was known. The whole database GLDM is referred to as the *global* GLDM.

The simplex minimization over all pose and shape space can be simplified by reducing t, the number of modes of variation of the PDM, and, hence, the dimensionality of the space, but this would result in a poorer fit as the shape model would represent less of the variance seen in the training set. The number of modes can be reduced without sacrificing variability if a PDM is also built for each talker. By removing the intertalker variation, the per-talker models have fewer or the same number of modes as the global PDM. This gives a set of PDM's, one for each talker, k, in the database,

$$\mathbf{x}_s^k = \overline{\mathbf{x}}_s^k + \mathbf{P}_s^k \mathbf{b}_s^k. \tag{11}$$

Fig. 6 shows the PDM modes at ± 2 standard deviations about the mean for each talker of the AVletters database plotted on the same axes. There are clearly large scale and mean shape differences between talkers. Only talkers two and seven have more than three modes. These are the two talkers with moustaches and this may be due in part to the difficulty that poses when locating the landmark points.

The shape parameters obtained by running an ASM with a *local* PDM cannot be related to the parameters obtained by fitting using the relevant PDM for a different talker. For multitalker speech recognition (trained and tested using examples from all talkers), to avoid training a separate hidden Markov model for each talker (which would be difficult for either of the small databases), the fit parameters must be mapped into a talker independent shape space. This is possible by transforming the fit parameters through the 88 point image coordinate space and into talker independent shape space using the talker independent, global PDM. First, the translation, rotation, and scaling pose differences between the mean shapes of the talker dependent and talker independent models must be removed. This aligns the mean 88 landmark points of both models as closely as possible, the remaining difference is described by the shape parameters in talker independent shape space, using (5), giving the multitalker shape parameters required.

The use of a coarse to fine multiresolution image pyramid to improve the performance of ASM's was demonstrated in [29] for point-wise iterative fitting. The image is Gaussian filtered and subsampled by two at each stage of the pyramid to form smaller versions of the original image. For each stage of the pyramid, a new GLDM must be built to learn the gray-level profiles about the model points. The mouth images of the AVletters databases are small, so usually only two resolutions are used—the original and the half sized image.

3.2 Active Appearance Models

An active appearance model (AAM) is a statistical model of both shape and gray-level appearance [26]. In the lipreading context, it combines the gray-level analysis approaches of [13], [15], [17], [33], [52], [65], [79], [89] with the shape analysis of [9], [24], [48], [75], [78], [84], [90].

There are some examples of using both gray-levels and shape. Luettin [54], [55], [56] used the GLDM fit parameters as well as the PDM shape parameters from an ASM fit, and Bregler [13], [14], [15], [16] used nonlinearly shape-constrained snakes to find the lips for an eigenanalysis. However, neither combine gray-level and shape in a *single* statistically learned model. An active appearance model is an extension of both of these techniques, it unifies eigenanalysis of the gray-levels and ASM lip tracking.

The active appearance model is trained from the same set of landmark point labeled images of the AVletters databases that were used for the PDM in Section 3.1. The shape part of an AAM is the PDM (4). The gray-level appearance model is built by warping each training image so the landmark points lie on the mean shape, $\overline{\mathbf{x}}$, normalising each image for shape. The gray-level values, $\mathbf{g}_{raw}$ are sampled within the landmark points of this shape normalized image. These are normalized over the training set for lighting variation using an iterative approach to find the best scaling, α, and offset, β,

$$\mathbf{g} = (\mathbf{g}_{raw} - \beta \mathbf{1})/\alpha, \tag{12}$$

where α and β are chosen to best match the normalized mean gray-level appearance, $\overline{\mathbf{g}}$. The mean appearance is scaled and offset for zero mean and unity variance, so the values of α and β are calculated using

$$\alpha = \mathbf{g}_{raw} \cdot \overline{\mathbf{g}} \tag{13}$$

$$\beta = (\mathbf{g}_{raw} \cdot \mathbf{1})/n, \tag{14}$$

where n is the number of elements in the gray-level appearance vector $\mathbf{g}$. As when aligning the shape models of the PDM, a stable normalized appearance model is

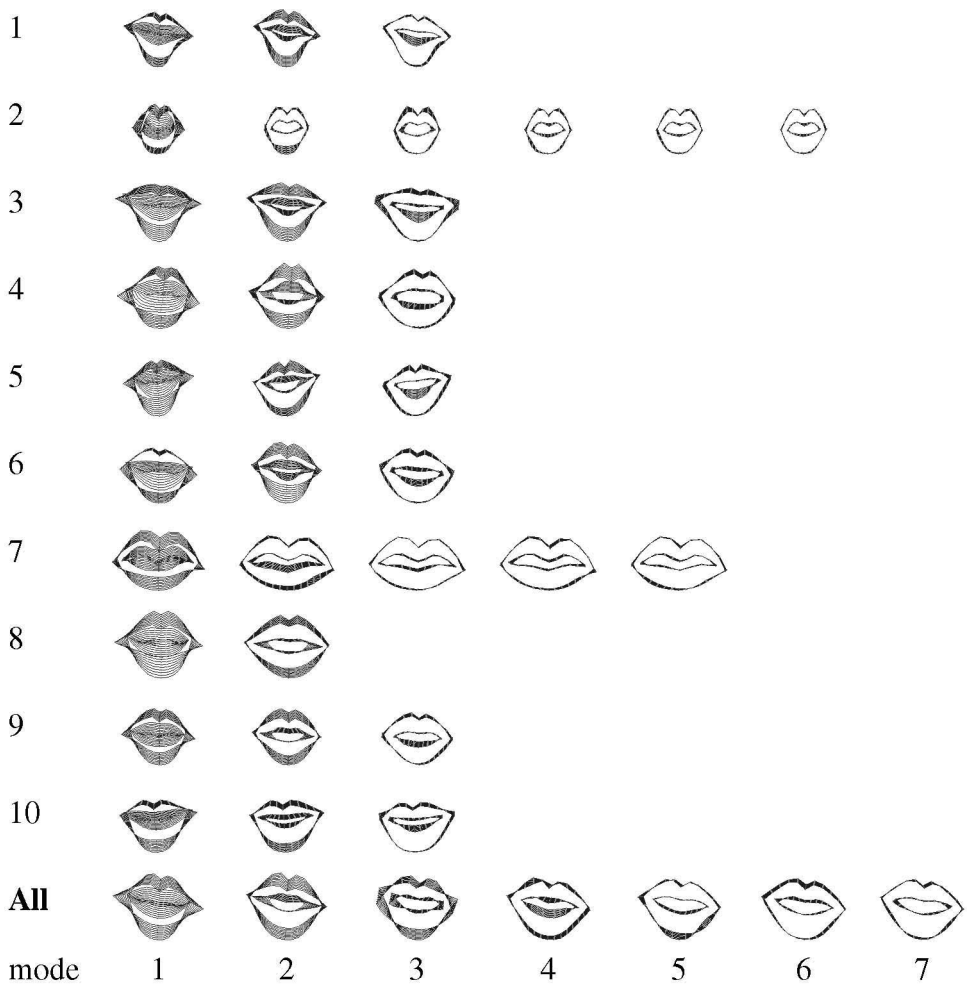

Fig. 6. Per talker PDM's showing modes that account for 95 percent of the variance. Talkers two and seven have moustaches—the extra modes for these talkers may be due to mislabelling the landmark points, which are much harder to place when the lip contours are obscured.

obtained by aligning to the first model, reestimating the mean, transforming, and reiterating.

The gray-level appearance model is calculated using PCA on the normalized gray-level data to identify the major modes of variation about the mean,

$$\mathbf{g} = \overline{\mathbf{g}} + \mathbf{P}_g \mathbf{b}_g, \tag{15}$$

where $\mathbf{P}_g$ is the set of t orthogonal modes of variation of the gray-level appearance and $\mathbf{b}_g$ a vector of t weights.

This extends the gray-level profile modeling used for ASM tracking to model the entire gray-level appearance within the landmark points rather than just profiles taken at the normal of each point. It is a principal component analysis of the shape and gray-level normalized pixel intensities within the shape defined by the landmark points of the hand labeled training images. The appearance model is built by applying a further PCA to identify the correlation between the shape parameters $\mathbf{b}_s$ and gray-level appearance parameters $\mathbf{b}_g$. A concatenated shape and gray-level appearance vector is formed for each example,

$$\mathbf{b} = \begin{pmatrix} \mathbf{W}_s \mathbf{b}_s \\ \mathbf{b}_g \end{pmatrix} = \begin{pmatrix} \mathbf{W}_s \mathbf{P}_s^T (\mathbf{x}_s - \overline{\mathbf{x}}_s) \\ \mathbf{P}_g^T (\mathbf{g} - \overline{\mathbf{g}}) \end{pmatrix}, \tag{16}$$

where $\mathbf{W}$ is a diagonal weight matrix for each shape parameter chosen to normalize the difference in units between the shape and gray-level appearance parameters and remove PCA scaling problems [26].

This gives a combined shape and gray-level appearance model,

$$\mathbf{b} = \mathbf{Q}\mathbf{c}, \tag{17}$$

where $\mathbf{Q}$ is the matrix of t eigenvectors and $\mathbf{c}$ the vector of t *appearance* parameters. Since the shape and gray-level appearance parameters have zero mean weights, $\mathbf{c}$ is also zero mean.

As the model is linear, shape and appearance can be expressed independently in terms of $\mathbf{c}$,

$$\mathbf{x}_s = \overline{\mathbf{x}}_s + \mathbf{P}_s \mathbf{W}_s \mathbf{Q}_s \mathbf{c} \tag{18}$$

$$\mathbf{g} = \overline{\mathbf{g}} + \mathbf{P}_g \mathbf{Q}_g \mathbf{c}, \tag{19}$$

where

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_s \\ \mathbf{Q}_g \end{pmatrix}. \tag{20}$$

Fig. 7 shows the first three modes at ± 2 standard deviations about the mean of the combined appearance

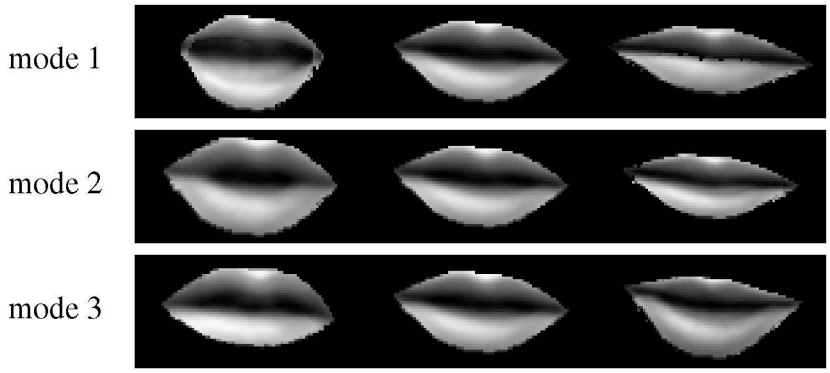

Fig. 7. Combined shape and gray-level appearance model. First three modes of variation of at ± 2 standard deviations about the mean.

model trained on the AVletters database. The full model has 37 modes of variation.

To fit an appearance model to an image, the Active Appearance Model algorithm [26] is used to find the best pose and appearance parameters. The fitting process minimizes the difference between the example image and that synthesized by the current model parameters. If the normalized gray-level appearance parameters of the image are $\mathbf{g}_i$ and the model synthesised values, from (19), $\mathbf{g}_m$, the difference is

$$\delta \mathbf{g} = \mathbf{g}_i - \mathbf{g}_m. \tag{21}$$

The AAM algorithm simplifies this high-dimensional optimization problem by learning, in advance, how to update the model parameters given the current difference image. Over a limited range of displacements, a linear model can accurately predict the correct model update from the difference image. The update model, $\mathbf{R}$, is calculated from the statistics obtained by systematically displacing the model pose and appearance parameters in the training images. To iteratively fit an AAM to an image, the model parameters are updated at each iteration using the update model

$$\mathbf{c} \mapsto \mathbf{c} - \mathbf{R}\delta \mathbf{g} \tag{22}$$

until no significant change occurs. A similar procedure is used for the pose parameters. The accurate prediction range of the linear update model is increased by using a multiresolution fitting approach.

Example iterations from a fit are shown in Fig. 8. The model is initialized in the center of images with the mean appearance parameters at the coarse resolution. After 15iterations, the model has converged at the fine scale; this took less than one second on a 166MHz Pentium. The converged appearance parameters form 37-dimensional lipreading feature vectors.

4 Bottom-up Multiscale Analysis

This section describes a bottom-up, pixel-based method that uses a multiscale spatial analysis (MSA) technique based on sieves [3], [4], [5], [7], [8]. Bottom-up statistical methods operating on pixels have the potential to reduce errors made in model-based approaches that are due to inaccurate prior assumptions about the salient image features for lipreading. Several previous low-level methods used principal component analysis to extract "eigenlip" features from the image gray levels [14], [17], [18], [33], [52]. Here, we also use principal component analysis, but not on gray-level intensity. Rather, in an attempt to make the system more robust, features are derived after mapping them into a nonlinear scale-space. The remapping, using a sieve transform, has the effect of decoupling spatial information from pixel intensities.

A *sieve* [3], [4], [5], [7], [8] is a serial filter structure based in mathematical morphology that progressively removes features from the input signal by increasing scale; Fig. 9 shows this structure. At each stage, the filtering element, ϕ, removes the extrema of only that scale. The first stage, ϕ_1, removes extrema of scale 1, ϕ_2 removes the 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 and preserves scale-space causality [41].

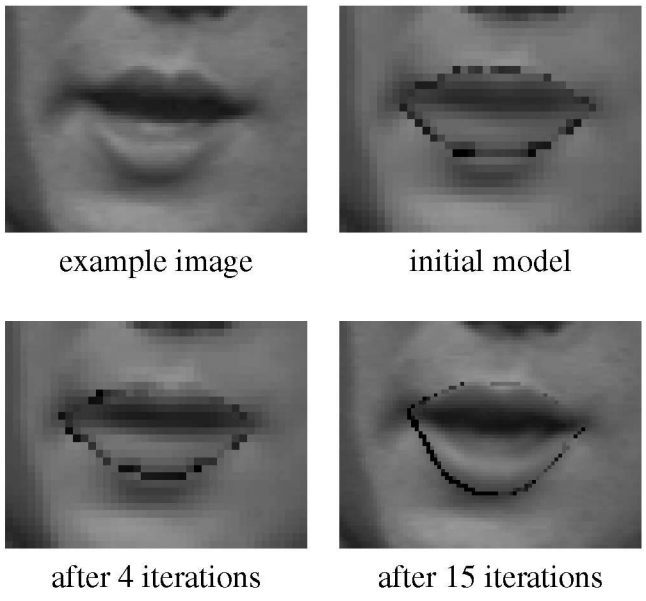

Fig. 8. Example of an AAM search. The model is initialized in the center of the image at coarse scale. Convergence using the AAM algorithm took 15 iterations.

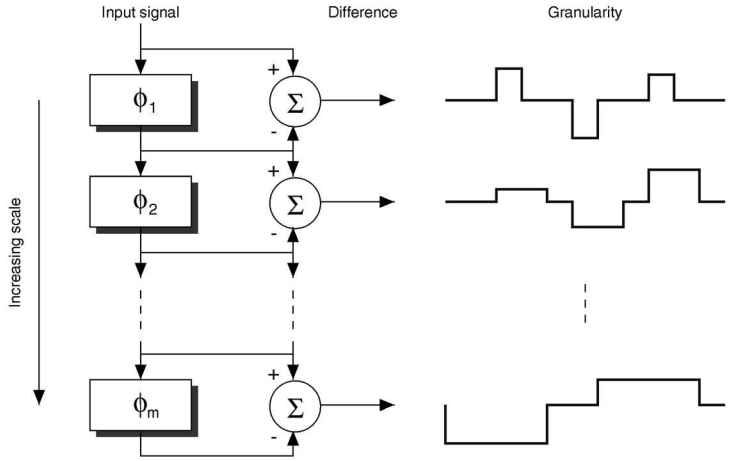

Fig. 9. Sieve structure.

A sieve can be defined in any number of dimensions by considering an image to be a set of connected pixels with their connectivity represented as a graph, $G = (V, E)$, where the set of vertices, V, are the pixel labels and the set of edges, E, represent the adjacencies. If the set of connected subsets of G containing r elements is $C_r(G)$, then the set of connected subsets of r elements containing the vertex x can be defined as

$$C_r(G, x) = \{\xi \in C_r(G) \mid x \in \xi\}. \tag{23}$$

In any number of dimensions, for each integer $r \geq 1$, an operator $\mathcal{Q}_r : \mathbf{Z}^V \mapsto \mathbf{Z}^V$ can be defined over the graph where $\mathcal{Q}_r$ is one of

$$\Psi_r f(x) = \max_{\xi \in C_r(G,x)} \min_{u \in \xi} f(u), \tag{24}$$

$$\gamma_r f(x) = \min_{\xi \in C_r(G,x)} \max_{u \in \xi} f(u), \tag{25}$$

$$\mathcal{M}_r f(x) = \gamma_r(\Psi_r f(x)), \tag{26}$$

$$\mathcal{N}_r f(x) = \Psi_r(\gamma_r f(x)). \tag{27}$$

For example, Ψ_2 is an opening of scale one (Ψ_1 operates on individual pixels so it has no effect on the signal) and removes all maxima of length one in 1D, area one in 2D, and so on for higher-dimensional signals. Likewise, for closing γ and alternating sequential $\mathcal{M}$- and $\mathcal{N}$-filters. Applying Ψ_3 to $\Psi_2(f(x))$ would further remove all maxima of scale two; length two for 1D and area two for 2D. This is the serial structure of a sieve, each stage removes the extrema (maxima and/or minima) of a particular scale. The output at a scale r, f_r, is the current scale filtered version of the previous signal

$$f_{r+1} = \mathcal{Q}_{r+1} f_r, \tag{28}$$

where the initial signal (unaffected by an $r = 1$ morphological filter) is $f_1 = \mathcal{Q}_1 f = f$. The differences between successive stages are the *granule functions*

$$d_r = f_r - f_{r+1}, \tag{29}$$

the nonzero regions of which are the *granules* of only that scale. The sieves defined using these functions are summarized in Table 1.

In the 1D case, (23) becomes the set of intervals containing r elements

$$C_r(x) = \{[x,\ x + r - 1] \mid x \in \mathbf{Z}\} \qquad r \geq 1 \tag{30}$$

which is identical to morphological filtering using a flat structuring element.

So far, we have described standard morphological filter based sieves. For example, an $\mathcal{M}$-filter is a two-pass operation which removes positive then negative extrema and, by applying the filters in the opposite order, an $\mathcal{N}$-filter removes negative extrema before positive. A further bipolar extrema processing variant is the recursive median filter

$$\rho_s f(x) = \begin{cases} \operatorname{med}(\rho_s f(x - s + 1), \ldots, \rho_s f(x - 1), \\ \qquad f(x), \ldots, f(x + s - 1)) & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{31}$$

$$r = (s - 1)/2. \tag{32}$$

The recursive median filter differs from $\mathcal{M}$- and $\mathcal{N}$-filters as it processes extrema in the order they occur in the signal. In practice, a recursive median sieve, or *m*-sieve, is a single pass method that gives similar results to *M*- or *N*-sieves but inherits the greater noise robustness of the recursive median filter [41]. It is also fast enough, $O = f(n)$ [6], to analyze the images used in this paper in real-time on a Silicon Graphics O2.

TABLE 1
Overview of Sieve Filter Types

Filter	Symbol	Sieve	Extrema Processing
opening	Ψ	*o*-sieve	maxima
closing	γ	*c*-sieve	minima
$\mathcal{M}$-filter	$\mathcal{M}$	*M*-sieve	bipolar $\pm$
$\mathcal{N}$-filter	$\mathcal{N}$	*N*-sieve	bipolar $\mp$

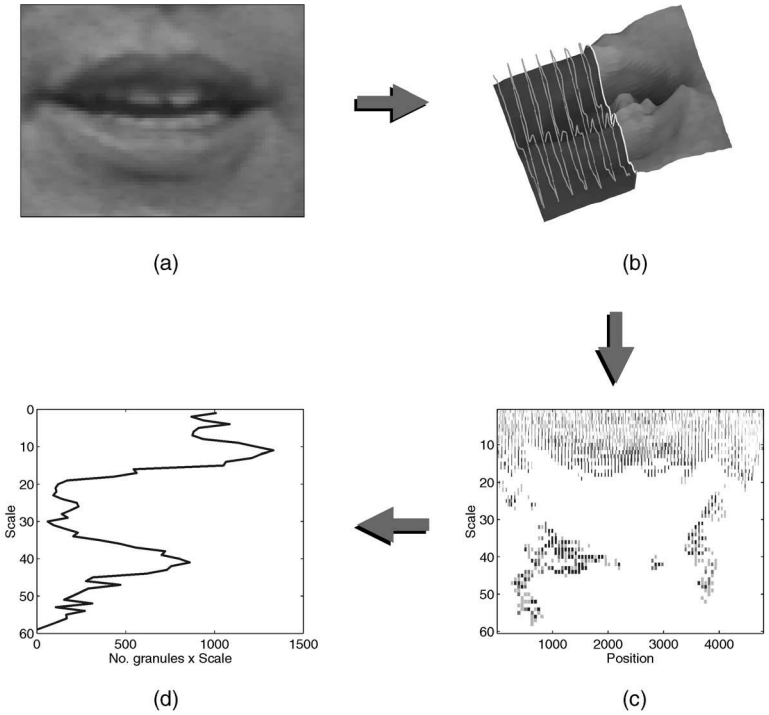

Fig. 10. Multiscale spatial analysis. The image (a) is vertically raster scanned. An example cut away (b) highlights a single slice in white and some previous scanlines in gray. The entire image is decomposed by scale (vertical length) into granularity (c) using an *m*-sieve. Positive granules are plotted in red, negative in blue. A scale histogram formed by summing or counting over position at each scale (d).

The mathematical properties of sieves have been well discussed [8]. There are two properties that are important here. First, the granule domain (29) is a mapping of the original signal and, so, all information present in the original is also present in the transformed granularity domain. In other words, the transform is *invertible* [3]. The second is that no new features (extrema) are created as scale is increased. In other words, a sieve preserves scale-space causality [7], [8] and large scale features are a faithful reflection of characteristics of the original image. The importance of this has been discussed at length [53] since the concept of scale-space was introduced [50], [86]. A comparison of sieves and several other scale space processors can be found in [12].

4.1 Feature Extraction

Mapping the image into a scale-space allows position, intensity, and scale to be dissociated. Earlier reports [42] explored features extracted with a 2D area-sieve, essentially a method of monitoring the area of the mouth aperture, but it appears that a more robust approach is to use a 1D length-sieve on the 2D image. A 1D analysis of a 2D image can be obtained by raster scanning, i.e., for a vertical scan start at the top left and scan down the first column, then repeat starting at the top of the second column, and so on. The resulting granularity describes the image in terms of granules that characterize amplitude, scale, and position relative to the raster scan. All of the sieve properties are maintained, the image is simply treated as a series of one-dimensional signals. An example image is shown in Fig. 10a and a vertical scan line highlighted in white in Fig. 10b with the preceding scan lines amplitude compressed to demonstrate the piece-wise 1D nature of this decomposition. The resulting granularity is shown in Fig. 10c, plotting scale on the vertical axis and vertical raster position relative to the top left of the image along the horizontal. Granules for this bipolar recursive median *m*-sieve decomposition are plotted in gray, which represents the absolute amplitude. The maximum scale is determined by the longest raster line. For this example, the maximum scale is the height of the image, 60 pixels.

The 1D decomposition separates the image features out according to length, in this case vertical length. Other 1D decompositions are possible by changing the direction of the raster scanning, for example, horizontal or at an arbitrary angle. For lipreading, where most mouth movement is up and down, it is preferable to scan the image vertically.

The next stage is to discard unnecessary information. This could be done in a number of ways. We use a *scale-histogram* that estimates the distribution of vertical granules obtained from the image. A scale histogram is shown in Fig. 10d, plotting scale on the vertical axis and number of granules at each scale along the horizontal, sh. This can be visualized as summing along the horizontal, position, and axis of Fig. 10c. A scale histogram formed in this way is a low-dimensional representation of the overall shape of the mouth and is most sensitive to vertical changes in the image.

If amplitude information is ignored by simply counting the number of granules at each scale, then the scale histogram is relatively insensitive to lighting conditions. An overall brightening or dimming of the image will not significantly change the granularity decomposition because it is the relative amplitudes of the pixels that define the

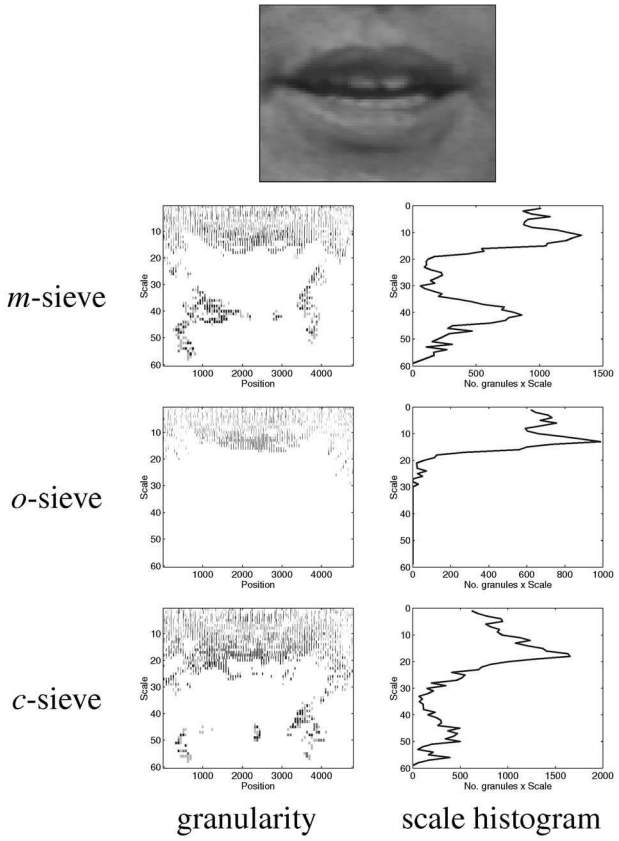

Fig. 11. Scale histogram sieve types. The full granularity of the image is plotted for recursive median, m, opening, o, and closing, c, sieves. Gray represents the absolute amplitude of the granules. The number of granules at each scale are plotted on the scale-histograms.

signal extrema. Until gray-level quantization or clipping effects are seen, a scale histogram is very stable to varying lighting conditions. However, it appears that significant information is described by the amplitude of granules, for example, when the granularity is inverted (transformed back to the image domain) the low amplitude granules have very little observable effect in the image. An alternative is summing the amplitudes at each scale (a). As amplitudes are generally bipolar, further alternatives are to sum the absolute ($|a|$) or squared amplitudes (a^2).

These measures are relatively insensitive to image translations, more so in the less sensitive horizontal direction. The most significant problem in practice is due to image scaling. Motion in the z-plane, toward or away from the camera, shifts granularity decompositions through scale. In practice, this could be solved by tracking head size as a separate process, for example, and using it for normalizing scale.

Any type of sieve can be used for a 1D rasterized image decomposition. A closing, *c*-sieve, is biased to process only negative extrema which are often associated with the dark mouth cavity. A recursive median, *m*-sieve, is more robust because it processes bipolar extrema and, hence, may be less sensitive to variations between talkers' appearance. Fig. 11 shows example granularity decompositions and scale count (sh) histograms for *m*-, *o*-, and *c*-sieves.

Extracting lipreading features in this way is abbreviated to Multiscale Spatial Analysis (MSA) to highlight the bottom-up scale based analysis used to derive lipreading

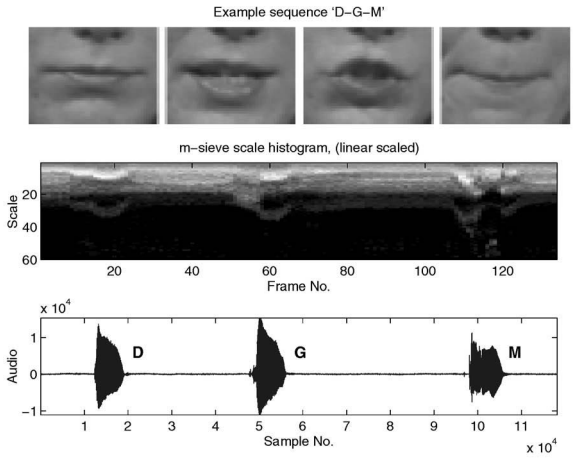

Fig. 12. Scale count histogram sequence. Top row shows example images from the sequence of isolated letters, "D-G-M." The middle row plots the scale count histogram (sh), the number of granules found at each of the 60 scales of the *m*-sieve vertical decomposition of the 80×60 images. The bottom row is the time-aligned audio waveform.

features. The scale-histogram obtained by counting the number of granules at each scale from an m-sieve is shown in Fig. 12 for the D-G-M image sequence.

4.2 Low-Level Statistical Model

The 1D scale-histograms discussed in the previous section extract 60-dimensional feature vectors from the 80×60 mouth images of the AVletters database. For recognition, a smaller feature space is desired that allows less complex statistical models which are easier to train. Principal component analysis (PCA) can again be used to identify orthogonal directions by their relative variance contribution to the multidimensional data. The values of the original data transformed along the top N directions can be used as decorrelated features in a reduced N-dimensional transformed feature space.

A problem using PCA on scale-histograms is that although the coefficients are all measures of scale they do not have similar variance. There are typically many more small scale objects in images than large ones and these often represent simple pixel-level noise. These will be identified as the major axes of variance and any correlation between small and large-scale coefficients is lost. The usual solution is to assume all variables are equally significant and normalize for variance by calculating PCA using the correlation matrix rather than the covariance matrix. However, if the variables are not equally important, then this is not recommended [21]. As the relative importance of each scale for lipreading is unknown, both methods were used to derive PCA features from scale-histograms. An example transformation calculated using the covariance matrix and taking the top 20 rotated directions is shown in Fig. 13 for a concatenated letter sequence.

5 Results

All recognition results were obtained using left to right, continuous density hidden Markov models (HMM's) with one or more diagonal covariance matrix Gaussian modes associated with each state. These were all implemented

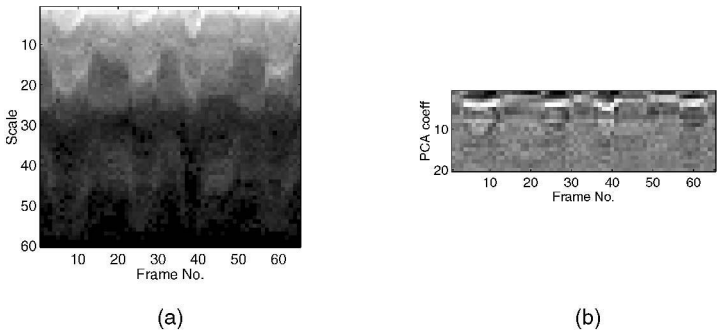

Fig. 13. Example PCA transformed scale-histogram sequence. The 60-dimensional scale count histogram (sh) from a *m*-sieve decomposition for concatenated isolated letters "A-V-S-P," (a), is transformed using the top 20 directions of PCA analysis using the covariance matrix calculated over the entire databse, (b).

using the hidden Markov model toolkit HTK version 2.1 [88]. In all cases, the recognition task is word-level and multitalker. Models are trained for all letters in the database using examples from all of the talkers. The training set was the first two utterances of each of the letters from all talkers (520 utterances) and the test set was the third utterance from all talkers (260 utterances). The HMM parameters for the number of states and number of Gaussian modes per state were systematically varied to find the model topology that gave the best recognition accuracy.

The effect of interpolating the data in time by a factor of two was also investigated. This was first used to resample the visual features to the same rate as the audio (from 40ms to 20ms) for an audio-visual integration experiment. However, this was found to have a beneficial effect even for the visual-only recognition task. This is partly due to the small size of the database. Interpolation creates more data, which is smoother, so the models can be better trained.

5.1 Visual-Only Recognition

For the ASM tracker, all results were obtained using a two-stage multiresolution fit initialized to the mean shape in the center of the coarse resolution 40×30 image for each frame. Three model fitting conditions were tested, talker dependent shape models and GLDM's (D_pD_g in Table 2), talker independent shape model with talker dependent GLDM's (I_pD_g in Table 2), and talker independent shape model and GLDM (I_pI_g). 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.

The AAM tracker was also initialized in the center of each course resolution image but was trained over all speakers and, so, is similar to the talker independent ASM. Recognition accuracy was tested using all 37 appearance parameters or just the top 20, 10, or 5. The results in Table 2 show that using only the top 20 gives almost identical performance to all 37. By modeling appearance as well as shape, the AAM tracker performs substantially better than the ASM tracker with best accuracies of 41.9 percent and 26.9 percent, respectively.

For MSA, there are a number of parameters that can be investigated; these are summarized in Table 3. The full set of 1,152 results represented by these parameters can be found in [60]. Several trends can be seen in the exhaustive test results that allow us to present only a subset of these results.

We have found it generally better to ignore the DC component (the maximum scale granule that is effectively the offset of the signal but may alter the resulting decomposition) and calculate the PCA using the covariance matrix. We also find that using the top 20 PCA components is better than using only the top 10. The best results are obtained using closing *c*-sieves on interpolated data from amplitude sum a scale-histograms. Note that

TABLE 2
Recognition Accuracy, % Correct, for Varying Number of HMM States and Gaussian Modes per State on ASM and AAM Tracker Data

States	5			7			9		
Modes	1	3	5	1	3	5	1	3	5
ASM I_pI_g	7.7	13.9	8.9	12.3	13.1	10.0	12.3	11.2	-
ASM I_pD_g	10.4	19.2	21.2	15.8	25.8	24.6	18.5	22.7	**26.9**
ASM D_pD_g	10.8	15.0	20.4	12.7	15.8	21.2	12.3	16.9	23.5
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

I_p independent speaker point distribution models, D_p dependent. I_g independent speaker gray-level distribution models, D_g dependent.

TABLE 3
Scale-Histogram Experimental Parameters

Attribute	Settings			
Sieve type	median, m	opening, o	closing, c	
Histogram type	sum, sh	amplitude, a	magnitude, $\|a\|$	squared, a^2
DC baseline	preserve	ignore		
Interpolate	no, 25Hz	yes, 50Hz		
PCA components	10	20		
PCA type	covariance	correlation		
HMM states	5	7	9	
Gaussian modes	1	3		

All 1,152 possibilities were tried.

there is no difference between amplitude sum a and magnitude sum $|a|$ for either o- or c-sieves because they extract, respectively, only positive or negative extrema. The MSA results for c- and m-sieves are summarized in Table 4 using the top 20 PCA components calculated using the covariance matrix from DC baseline ignoring data. The best result is 44.6 percent correct.

These results suggest that, when using sieves to extract visual speech features, it is the dark image regions that capture most information. In practice, if lighting conditions were such that the mouth cavity was brighter than the face, this would no longer capture the same information. This might occur if there is direct light into the mouth. Features derived using bipolar recursive median sieves have similar performance and might be expected to be more robust to different skin tone and lighting conditions.

5.2 Audio-Visual Recognition

Lipreading is not a goal in itself rather it can be used to improve the reliability of audio-visual speechreading. To get some intuition on how lipreading might improve speechreading, the audio signal is degraded by adding white noise. This increases the error rate for audio speech recognition and the aim of the experiment is to use the video stream to reduce the error rate again. A first strategy for integrating the audio and video streams is to linearly combine the probabilities output by each recognizer, so that we recognize word w^*, where

$$w^* = \underset{i=1,2,\ldots,v}{\operatorname{argmax}} \{\alpha \Pr(w_i|A) + (1-\alpha)\Pr(w_i|v)\}, \tag{33}$$

where $\Pr(w_i|A)$ and $\Pr(w_i|v)$ are the respective probabilities of the ith word from the audio and video recognizers and α is a weighting factor.

Several approaches may be used for deriving the value of α; we have used a confidence measure based on the uncertainty of the audio recognizer about a word at a given signal to noise ratio (SNR). Let $H_A(X|Y)$ be the average uncertainty of the audio recognizer about the input word X given knowledge of the output word Y and let $H_A(X|Y)_{max}$ be the maximum uncertainty ($log_2 v$). Then, α is defined as,

$$\alpha = 1 - \frac{H_A(X|Y)}{H_A(X|Y)_{max}}. \tag{34}$$

Previous results [30] 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.

Fig. 14a plots recognition accuracy over a range of SNR's. Spectrum subtraction [11] 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.

Fig. 14b plots the same using the best AAM visual information and Fig. 14c likewise for the best MSA results. A comparison between ASM, AAM, and MSA is shown in Fig. 14d. The results obtained using AAM and MSA are remarkably similar. Other, more elaborate, schemes for combining the audio and video streams have been investigated [82] but our concern here is to investigate the performance of the visual features and, so, we have used a very simple technique.

TABLE 4
Shows How Varying the HMM Parameters: Number of States, S, and Gaussian Mixtures (1 or 3) Affect Recognition Accuracy, %, for Interpolated, I, and Noninterpolated, N, Data for Both *C*-Sieve and *M*-Sieves for all Scale-Histogram Types, T

T	S	c-sieve				m-sieve			
		I		N		I		N	
		1	3	1	3	1	3	1	3
sh	5	26.2	36.2	24.6	34.6	21.9	38.1	18.5	36.2
	7	24.2	36.6	28.5	34.6	27.3	40.8	25.8	41.2
	9	30.8	37.7	30.8	39.2	27.3	38.9	28.5	40.8
a	5	24.6	36.5	22.3	40.0	18.9	31.9	21.9	33.1
	7	27.3	36.5	26.9	36.2	24.2	30.8	20.4	33.1
	9	32.7	**44.6**	30.0	41.5	26.5	33.4	25.8	35.8
$\|a\|$	5	24.6	36.5	22.3	40.0	19.6	36.2	20.8	35.8
	7	27.3	36.5	26.9	36.2	28.1	36.9	25.8	38.9
	9	32.7	**44.6**	30.0	41.5	30.0	40.8	28.1	39.6
a^2	5	17.7	34.2	13.1	31.9	18.1	31.9	19.6	29.6
	7	23.1	34.6	21.2	34.6	20.0	32.6	21.5	30.4
	9	21.5	37.3	27.7	28.4	23.4	31.5	21.2	30.8

6 Discussion

It seems fair to say that the development of features for lipreading is currently at about the same stage as the development of features for automatic speech recognition (ASR) was in the early 1980's. At that time, features for ASR were generally either derived from a model-based approach, linear prediction [2], or a data-driven approach using the short-term spectrum of the speech signal [31]. It was not until the mid-1980's that mel-frequency cepstral coefficients (MFCC's) [32], which can be derived from both

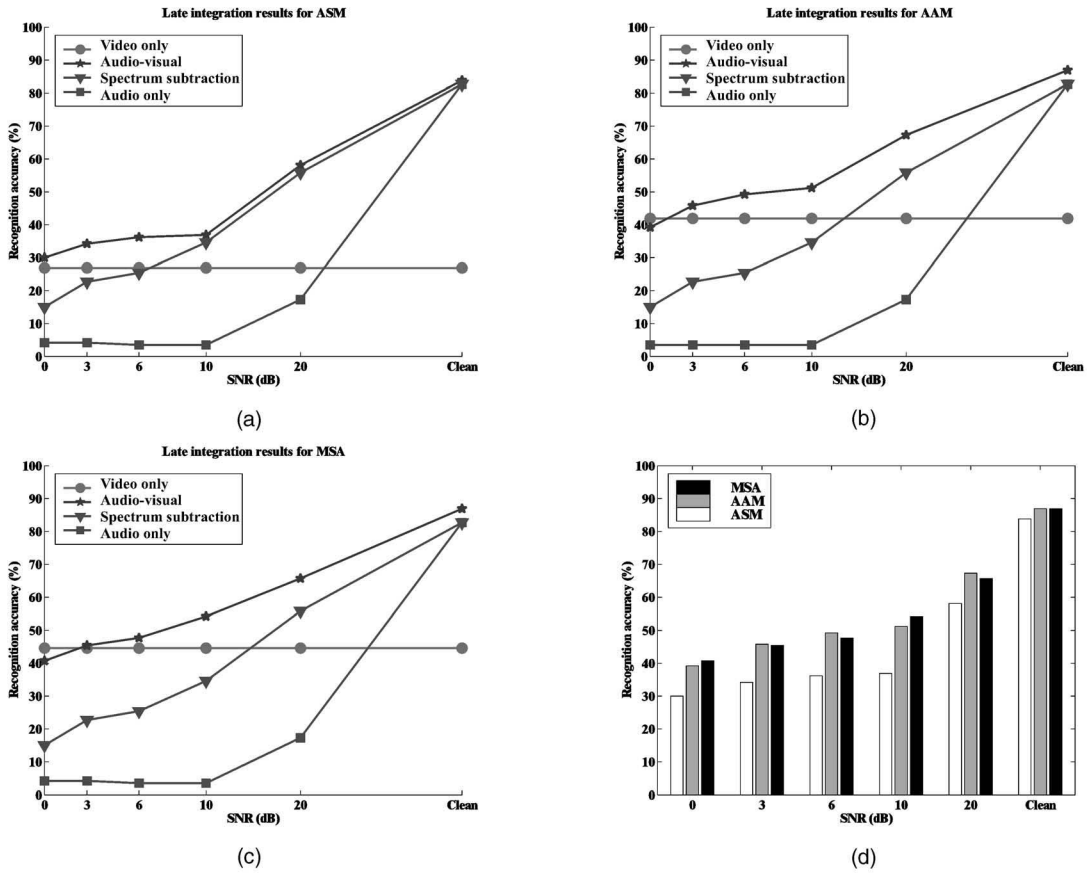

Fig. 14. Late integration results for all methods. (a) ASM, (b) AAM, and (c) MSA. (d) Direct comparison of the audio-visual speechreading methods.

approaches, emerged as clearly superior and were adopted almost universally by the speech recognition community. Since then, there have been some refinements and extensions [37] to the use of MFCC's in ASR, but they have remained the feature of choice.

In this work, we have compared some approaches for extracting features for lipreading. The first important finding is that, by including gray-level appearance information, AAM's achieve better recognition rates than ASM's that use shape information alone. AAM's are also found to have advantages for handling textures [25] and interpreting facial images [35]. It is, perhaps, interesting to note that, in a completely different context, line drawings (cf., ASM's) are widely used to illustrate hand signing for the deaf and photographs with full gray-level information (cf., AAM's) are more usually used to teach facial expressions.

The second finding is that comparable recognition rates are achieved with the MSA approach. The sieves used here are emerging as a useful way to extract patterns from signals and images [47], [64] (note: the MSA used here is typically faster than a discrete Fourier transform). Sieves are based on connected-level-sets and, so, differ from those mathematical morphology based systems first proposed for analysing images (*granulometries*) for textures [59] and shapes [72]. Granulometries reflect the match between the underlying image and multiple scale structuring elements and we are not aware of evidence that these discriminate shapes more effectively than other methods.

There is scope to improve both methods. For example, the use of a predictive temporal tracking framework can be expected to reduce shape model tracking errors [34], [49] and MSA should be improved with lip tracking to better identify the mouth area and correct for image scale. We have not quantified how robust any of these methods are to variation in lighting, pose, angle, etc. and this should now be investigated. However, experience demonstrating a small vocabulary, real-time speechreading system, using MSA on a Silicon Graphics O2, under varying lighting conditions suggests that lighting is less of a problem than pose.

Perhaps the most significant observation made during these experiments is illustrated in Table 5. It appears that, although features obtained by MSA and AAM's yield similar recognition rates, they fail in different ways. This suggests that it might be possible to construct a system that exploits the benefits of both, which is a similar data-fusion problem to that of audio-visual integration. The simplest approach is to form a combined feature space by concatenating the two feature vectors. Another option is to have two classifiers, one for each feature set, and in the event of a disagreement between the two using some form of confidence measure to resolve the issue. However, the current database contains too few repetitions to train and test such a system and, so, this investigation awaits further work.

TABLE 5
Table Comparing the Numbers of Correctly and Incorrectly Recognized Utterances for the Best MSA and AAM Recognizers

		MSA		AAM total
		correct	incorrect	
AAM	correct	62	47	109 (41.9%)
	incorrect	54	97	151
MSA total		116 (44.6%)	144	260

Note that many utterances are correctly recognized by one, but not the other method.

The motivation for lipreading lies in the contribution that it, and systems for recognizing other gestures, can make to the process of reliably communicating naturally with a computer. For this, we have combined audio with visual speech recognition. Section 5.2 shows that speech recognition accuracy is significantly improved when a noisy audio signal is augmented with visual information, even if the audio signal has already been enhanced using noise cancellation. The results stand compared with those from the classical work on the importance of visual clues for human recognition [68]. Further experiments using alternative noise sources, particularly the "babble" of irrelevant talkers in the background, might show still better improvements as the visual information cues the audio. The potential gains of multimodal speech recognition extend further than simply improving recognition accuracy. Summerfield [82] notes that the other benefits of vision to speech include the ability to identify the talker, find their location, and determine to whom they are speaking to. Such information is clearly useful for a truly easy to use, intelligent man-machine interface.

REFERENCES

[1] A. Adjoudani and C. Benoît, "On the Integration of Auditory and Visual Parameters in an HMM-Based ASR," *Speedreading by Humans and Machines: Models, Systems, and Applications,* vol. 150, pp. 461–471, 1996.

[2] B. Atal and L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *J. Acoustical Soc. of America,* vol. 50, pp. 637-655, 1971.

[3] J.A. Bangham, P. Chardaire, C.J. Pye, and P.D. Ling, "Mulitscale Nonlinear Decomposition: The Sieve Decomposition Theorem," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 5, pp. 529-539, May 1996.

[4] J.A. Bangham, R. Harvey, P. Ling, and R.V. Aldridge, "Morphological Scale-Space Preserving Transforms in Many Dimensions," *J. Electronic Imaging,* vol. 5, no. 3, pp. 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 Conf. Computer Vision,* vol. 1, pp. 189-198, 1996.

[6] J.A. Bangham, S.J. Impey, and F.W.D. Woodhams, "A Fast 1D Sieve Transform for Multiscale Signal Decomposition," *Proc. European Signal Processing Conf.,* pp. 1621-1624, 1994.

[7] J.A. Bangham, P. Ling, and R. Harvey, "Scale-Space from Nonlinear Filters," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 5, pp. 520-528, May 1996.

[8] J.A. Bangham, P. Ling, and R. Young, "Mulitscale Recursive Medians, Scale-Space, and Transforms with Applications to Image Processing," *IEEE Trans. Image Processing,* vol. 5, no. 6, pp. 1043-1048, 1996.

[9] S. Basu, N. Oliver, and A. Pentland, "3D Modeling and Tracking of Human Lip Motions," *Proc. Int'l Conf. Computer Vision,* 1998.

[10] *Proc. ESCA Workshop Audio-Visual Speech Processing,* C. Benoît and R. Campbell, eds., Rhodes, Sept. 1997.

[11] S. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoustics, Speech, and Signal Processing,* vol. 27, pp. 113-120, 1979.

[12] A. Bosson and R. Harvey, "Using Occlusion Models to Evaluate Scale Space Processors," *Proc. IEEE Int'l Conf. Image Processing,* 1998.

[13] C. Bregler, H. Hild, S. Manke, and A. Waibel, "Improving Connected Letter Recognition by Lipreading," *Proc. Int'l Conf. Acoustics, Speech and Signal Processing,* vol. 1, pp. 557-560, 1993.

[14] C. Bregler and Y. Konig, "'Eigenlips' for Robust Speech Recognition," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing,* pp. 669-672, 1994.

[15] C. Bregler and S.M. Omohundro, "Learning Visual Models for Lipreading," *Computational Imaging and Vision,* chapter 13, vol. 9, pp. 301-320, 1997.

[16] C. Bregler, S.M. Omohundro, and J. Shi, "Towards a Robust Speechreading Dialog System," *NATO ASI Series F: Computer and Systems Sciences,* pp. 409-423, Sept. 1996.

[17] N.M. Brooke and S.D. Scott, "PCA Image Coding Schemes and Visual Speech Intelligibility," *Proc. Inst. of Acoustics,* vol. 16, no. 5, pp. 123-129, 1994.

[18] N.M. Brooke, M.J. Tomlinson, and R.K. Moore, "Automatic Speech Recognition that Includes Visual Speech Cues," *Proc. Inst. of Acoustics,* vol. 16, no. 5, pp. 15-22, 1994.

[19] J. Bulwer, *Philocopus, or the Deaf and Dumbe Mans Friend.* Humphrey and Moseley, 1648.

[20] *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech,* R. Campbell, B. Dodd, and D. Burnham, eds., Psychology Press, 1998.

[21] C. Chatfield and A. J. Collins, *Introduction to Multivariate Analysis.* Chapman and Hall, 1991.

[22] T. Chen and R.R. Rao, "Audio-Visual Integration in Multimodal Communication," *Proc. IEEE,* vol. 86, no. 5, pp. 837-852, May 1998.

[23] C.C. Chibelushi, S. Gandon, J.S.D. Mason, F. Deravi, and R.D. Johnston, "Design Issues for a Digital Audio-Visual Integrated Database," *IEE Colloquium on Integrated Audio-Visual Processing,* number 1996/213, pp. 7/1-7/7, Nov. 1996.

[24] T. Coianiz, L. Torresani, and B. Caprile, "2D Deformable Models for Visual Speech Analysis," *IEEE Trans. Speech and Audio Processing,* pp. 391-398, Sept. 1996.

[25] T. Cootes, G.J. Edwards, and C. Taylor, "Comparing Active Shape Models with Active Appearance Models," *Proc. British Machine Vision Conf.,* vol. 1, pp. 173-183, 1999.

[26] T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active Appearance Models," *Proc. European Conf. Computer Vision,* pp. 484-498, June 1998.

[27] 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,* vol. 12, no. 6, pp. 355-366, 1994.

[28] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham, "Active Shape Models—Their Training and Application," *Computer Vision and Image Understanding,* vol. 61, no. 1, pp. 38-59, Jan. 1995.

[29] T.F. Cootes, C.J. Taylor, and A. Lanitis, "Active Shape Models: Evaluation of a Multiresolution Method for Improving Image Search," *Proc. British Machine Vision Conf.,* E. Hancock, ed., pp. 327-336, 1994.

[30] S. Cox, I. Matthews, and A. Bangham, "Combining Noise Compensation with Visual Information in Speech Recognition," *Proc. ESCA Workshop Audio-Visual Speech Processing,* pp. 53-56, 1997.

[31] B. Dautrich, L. Rabiner, and T. Martin, "On the Effects of Varying Filter Bank Parameters on Isolated Word Recognition," *IEEE Trans. Acoustics, Speech and Signal Processing,* vol. 31, pp. 793-807, Aug. 1983.

[32] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoustics, Speech, and Signal Processing,* vol. 28, pp. 357-366, 1980.

[33] P. Duchnowski, M. Hunke, D. Büsching, U. Meier, and A. Waibel, "Toward Movement-Invariant Automatic Lip-Reading and Speech Recognition," *Proc. Int'l Conf. Spoken Language Processing,* pp. 109-112, 1995.

[34] G.J. Edwards, T.F. Cootes, and C.J. Taylor, "Face Recognition Using Active Appearance Models," *Proc. European Conf. Computer Vision,* pp. 582-595, June 1998.

[35] G.J. Edwards, C. Taylor, and T.F. Cootes, "Interpreting Face Images Using Active Appearance Models," *Proc. Third Int'l Conf. Automatic Face and Gesture Recognition,* pp. 300-305, 1998.

[36] N.P. Erber, "Auditory-Visual Perception of Speech," *J. Speech and Hearing Disorders,* vol. 40, pp. 481-492, 1975.

[37] S. Furui, "Speaker Independent Isolated Word Recognition Using Dynamic Features of the Speech Spectrum," *IEEE Trans. Acoustics, Speech, and Signal Processing,* 1984.

[38] A.J. Goldschen, "Continuous Automatic Speech Recognition by Lipreading," PhD thesis, George Washington Univ., 1993.

[39] A.J. Goldschen, O.S. Garcia, and E.D. Petajan, "Continuous Automatic Speech Recognition by Lipreading," *Computational Imaging and Vision,* chapter 14, pp. 321-343, 1997.

[40] K.P. Green, "The Use of Auditory and Visual Information During Phonetic Processing: Implications for Theories of Speech Perception," *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech,* pp. 3-25, 1998.

[41] R. Harvey, A. Bosson, and J.A. Bangham, "Robustness of Some Scale-Spaces," *Proc. British Machine Vision Conf.,* vol. 1, pp. 11-20, 1997.

[42] R. Harvey, I. Matthews, J.A. Bangham, and S. Cox, "Lip Reading from Scale-Space Measurements," *Proc. Conf. Computer Vision and Pattern Recognition* pp. 582-587, June 1997.

[43] J. Haslam, C.J. Taylor, and T.F. Cootes, "A Probabilistic Fitness Measure for Deformable Template Models," *Proc. British Machine Vision Conf.,* pp. 33-42, 1994.

[44] M.E. Hennecke, "Audio-Visual Speech Recognition: Preprocessing, Learning and Sensory Integration," PhD thesis, Stanford Univ., Sept. 1997.

[45] M.E. Hennecke, D.G. Stork, and K.V. Prasad, "Visionary Speech: Looking Ahead to Practical Speechreading Systems," *NATO ASI Series F: Computer and Systems Science,* pp. 331-349, 1996.

[46] A. Hill and C.J. Taylor, "Automatic Landmark Generation for Point Distribution Models," *Proc. British Machine Vision Conf.,* pp. 429-438, 1994.

[47] A. Holmes and C. Taylor, "Developing a Measure of Similarity between Pixel Signatures," *Proc. British Machine Vision Conf.,* vol. 2, pp. 614-623, 1999.

[48] R. Kaucic and A. Blake, "Accurate, Real-Time, Unadorned Lip Tracking," *Proc Sixth Int'l Conf. Computer Vision,* 1998.

[49] R. Kaucic, B. Dalton, and A. Blake, "Real-Time Lip Tracking for Audio-Visual Speech Recognition Applications," *Proc. European Conf. Computer Vision,* B. Buxton and R. Cipolla, eds., pp. 376-387, Apr. 1996.

[50] J.J. Koenderink, "The Structure of Images," *Biological Cybernetics,* vol. 50, pp. 363-370, 1984.

[51] 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 J.,* vol. 62, no. 4, pp. 1035-1074, Apr. 1983.

[52] N. Li, S. Dettmer, and M. Shah, "Visually Recognizing Speech Using Eigensequences," *Computational Imaging and Vision,* chapter 15, pp. 345-371, 1997.

[53] T. Lindeberg, *Scale-Space Theory in Computer Vision.* Kluwer Academic, 1994.

[54] J. Luettin, "Visual Speech and Speaker Recognition," PhD thesis, Univ. of Sheffield, May 1997.

[55] J. Luettin and N.A. Thacker, "Speechreading Using Probabilistic Models," *Computer Vision and Image Understanding,* vol. 65, no. 2, pp. 163-178, Feb. 1997.

[56] J. Luettin, N.A. Thacker, and S.W. Beet, "Speechreading Using Shape and Intensity Information," *Proc. Fourth Int'l Conf. Spoken Language Processing (ICSLP '96),* vol. 1, pp. 58-61, 1996.

[57] J. MacDonald and H. McGurk, "Visual Influences on Speech Perception Processes," *Perception and Psychophysics,* vol. 24, pp. 253-257, 1978.

[58] K. Mase and A. Pentland, "Automatic Lipreading by Optical-Flow Analysis," *Systems and Computers in Japan,* vol. 22, no. 6, pp. 67-75, 1991.

[59] G. Matheron, *Random Sets and Integral Geometry.* Wiley, 1975.

[60] I. Matthews, "Features for Audio-Visual Speech Recognition," PhD thesis, School of Information Systems, Univ. East Anglia, Oct. 1998.

[61] 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," *Proc. European Conf. Computer Vision,* pp. 514-528, June 1998.

[62] H. McGurk and J. McDonald, "Hearing Lips and Seeing Voices," *Nature,* vol. 264, pp. 746-748, Dec. 1976.

[63] U. Meier, R. Stiefelhagen, and J. Yang, "Preprocessing of Visual Speech Under Real World Conditions," *Proc. ESCA Workshop Audio-Visual Speech Processing,* pp. 113-116, Sept. 1997.

[64] K. Morovec, R.W. Harvey, and J.A. Bangham, "Scale-Space Trees and Applications as Filters, for Stereo Vision and Image Retrieval," *Proc. British Machine Vision Conf.,* vol. 1, pp. 113-122, 1999.

[65] J.R. Movellan and G. Chadderdon, "Channel Seperability in the Audio Visual Integration of Speech: A Bayesian Approach," *NATO ASI Series F: Computer and Systems Science,* pp. 473-487, 1996.

[66] K.K. Neely, "Effect of Visual Factors on the Intelligibility of Speech," *J. Acoustical Soc. of America,* vol. 28, no. 6, pp. 1275-1277, Nov. 1956

[67] J.A. Nelder and R. Mead, "A Simplex Method for Function Minimization," *Computing J.,* vol. 7, no. 4, pp. 308-313, 1965.

[68] J.J. O'Neill, "Contributions of the Visual Components of Oral Symbols to Speech Comprehension," *J. Speech and Hearing Disorders,* vol. 19, pp. 429-439, 1954.

[69] E. Petajan and H.P. Graf, "Robust Face Feature Analysis for Automatic Speechreading and Character Animation," *NATO ASI Series F: Computer and Systems Science,* pp. 425-436, 1996.

[70] E.D. Petajan, "Automatic Lipreading to Enhance Speech Recognition," PhD thesis, Univ. of Illinois, Urbana-Champaign, 1984.

[71] E.D. Petajan, B.J. Bischoff, D.A. Bodoff, and N.M. Brooke, "An Improved Automatic Lipreading System to Enhance Speech Recognition," Technical Report TM 11251-871012-11, AT&T Bell Labs, Oct. 1987.

[72] I. Pitas and A.N. Venetsanopoulos, "Morphological Shape Decomposition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 12, no. 1, pp 38-45, Jan. 1990.

[73] G. Potamianos, F. Cosatto, H.P. Graf, and D.B. Roe, "Speaker Independent Audio-Visual Database for Bimodal ASR," *Proc. ESCA Workshop Audion-Visual Speech Processing,* pp. 65-68, Sept. 1997.

[74] C.A. Poynton, *A Technical Introduction to Digital Video.* John Wiley & Sons, 1996.

[75] M.U.R. Sánchez, J. Matas, and J. Kittler, "Statistical Chromaticity-Based Lip Tracking with B-Splines," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing,* Apr. 1997.

[76] J.-L. Schwartz, J. Robert-Ribes, and P. Escudier, "Ten Years after Summerfield: A Taxonomy of Models for Audio-Visual Fusion in Speech Perception," *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech,* pp. 85-108, 1998.

[77] "Motion-Based Recognition," *Computational Imaging and Vision,* M. Shah and R. Jain, eds., vol. 9, Kluwer Academic, 1997.

[78] P.L. Silsbee, "Motion in Deformable Templates," *Proc. IEEE Int'l Conf. Image Processing,* vol. 1, pp. 323-327, 1994.

[79] P.L. Silsbee, "Computer Lipreading for Improved Accuracy in Automatic Speech Recognition," *IEEE Trans. Speech and Audio Processing,* vol. 4, no. 5, pp. 337-351, Sept. 1996.

[80] "Speechreading by Humans and Machines: Models, Systems, and Applications," *NATO ASI Series F: Computer and Systems Sciences,* D.G. Stork and M.E. Hennecke, eds., vol. 150, 1996.

[81] W.H. Sumby and I. Pollack, "Visual Contribution to Speech Intelligibility in Noise," *J. Acoustical Soc. of Am.,* vol. 26, no. 2, pp. 212-215, Mar. 1954.

[82] Q. Summerfield, "Some Preliminaries to a Comprehensive Account of Audio-Visual Speech Perception," *Hearing by Eye: The Psychology of Lip-Reading,* B. Dodd and R. Campbell, eds., pp. 3-51, 1987.

[83] M.J. Tomlinson, M.J. Russell, and N.M. Brooke, "Integrating Audio and Visual Information to Provide Highly Robust Speech Recognition," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing,* vol. 2, pp. 821-824, May 1996.

[84] M. Vogt, "Interpreted Multi-State Lip Models for Audio-Visual Speech Recognition," *Proc. ESCA Workshop Audio-Visual Speech Processing,* pp. 125-128, Sept. 1997.

[85] B.E. Walden, R.A. Prosek, A.A. Montgomery, C.K. Scherr, and C.J. Jones, "Effects of Training on the Visual Recognition of Consonants," *J. Speech and Hearing Research,* vol. 20, pp. 130-145, 1977.

[86] A.P. Witkin, "Scale-Space Filtering," *Proc. Eighth Int'l Joint Conf. Artificial Intelligence,* vol. 2, pp. 1019-1022, 1983.
[87] J. Yang, R. Stiefelhagen, U. Meier, and A. Waibel, "Real-Time Face and Facial Feature Tracking and Applications," *Proc. Workshop Auditory-Visual Speech Processing,* D. Burnham, J. Robert-Ribes, and E. Vatikiotis-Bateson, eds., pp. 79-84, Dec. 1998.
[88] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland, *The HTK Book.* Cambridge Univ., 1996.
[89] B.P. Yuhas, M.H. Goldstein Jr., and T.J. Sejnowski, "Integration of Acoustic and Visual Speech Signals Using Neural Networks," *IEEE Comm. Magazine,* vol. 27, pp. 65-71, 1989.
[90] A.L. Yuille, P.W. Hallinan, and D.S. Cohen, "Feature Extraction from Faces Using Deformable Templates," *Int'l J. Computer Vision,* vol. 8, no. 2, pp. 99-111, 1992.

Iain Matthews received the BEng (1994) and PhD (1998) degrees in electronic engineering from the University of East Anglia, Norwich. Since then, he has been at Carnegie Mellon University, Pittsburgh, and is currently a post-doctoral fellow in the Robotics Institute. His research interests include statistical modeling of faces and tracking. He is a member of the IEEE and the IEEE Computer Society.

Timothy F. Cootes received the BSc degree from Exeter University, England, in 1986, and the PhD degree in engineering from City Polytechnic, Sheffield, UK, in 1991. Since then, he has worked in computer vision at the University of Manchester, UK, and has received two fellowships from Engineering and Physical Sciences Research Council, UK. He is currently a research fellow in the Division of Imaging Science at the University of Manchester, UK, with research interests that include statistical models of shape and appearance and their applications in interpreting face images, industrial, and medical computer vision problems. He is a member of the IEEE and the IEEE Computer Society.

J. Andrew Bangham received the PhD degree in biophysics from University College London. Having worked as an electrophysiologist, he moved to the computer sciences at the University of East Anglia (UEA), UK, where he started a computer vision laboratory. He is currently a professor in the School of Information Systems at UEA. His primary research interest lies in nonlinear low-level computer vision systems but other recent projects have included automatic signing systems for the deaf, models of plant development, and the painterly rendering of digital photographs. He is a member of the IEEE and the IEEE Computer Society.

Stephen Cox received the BSc degree in physics and music from Reading University and the PhD degree in speech recognition from The University of East Anglia (UEA), UK. He headed a group at BT developing robust speech recognition algorithms for use over the telephone network and is currently a senior lecturer in the School of Information Systems with interests in speech processing and pattern recognition. He is a member of the IEEE and is chairman of the UK Institute of Acoustics Speech Group.

Richard Harvey received the BSc degree in electronics and the PhD degree in statistical estimation theory. After a period working in industry on sonar, he is now a senior lecturer in the School of Information Systems at the University of East Anglia (UEA), UK. He has interests in computer vision, pattern recognition, and the understanding of images. He is a member of the IEEE.