IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

Speechreading using Probabilistic Models

Juergen Luettin [1 2] Neil A. Thacker [2 3]

IDIAP–RR 97-12

PUBLISHED IN
Computer Vision and Image Understanding, Vol. 65, No. 2, February, pp. 163-178, 1997

Dalle Molle Institute for Perceptive Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

[1] Institut Dalle Molle d'Intelligence Artificielle Perceptive, C.P. 592, Rue du Simplon 4, CH-1920 Martigny, Switzerland, email: luettin@idiap.ch.
[2] Department of Electronic and Electrical Engineering, University of Sheffield, Mappin Street, Sheffield S1 3JD, UK.
[3] Current address: Department of Medical Biophysics, University of Manchester, Oxford Road, Manchester M13 9PT, UK.

IDIAP Research Report 97-12

Speechreading using Probabilistic Models

Juergen Luettin Neil A. Thacker

PUBLISHED IN
Computer Vision and Image Understanding, Vol. 65, No. 2, February, pp. 163-178, 1997

Abstract. A robust method for locating and tracking lips in gray-level image sequences is described. The method learns patterns of shape variability from a training set which constrains the model during image search to only deform in ways similar to the training examples. Image search is guided by a learned gray-level model which is used to describe the large appearance variability of lips. Such variability might be due to different individuals, illumination, mouth opening, specularity or visibility of teeth and tongue. Visual speech features are recovered from the tracking results and represent both, shape and intensity information. We describe a speechreading (lip-reading) system, where the extracted features are modeled by Gaussian distributions and their temporal dependencies by Hidden Markov Models. Experimental results are presented for locating lips, tracking lips and speechreading. The database used consists of a broad variety of speakers and was recorded in a natural environment with no special lighting or lip markers used. For a speaker independent digit recognition task using visual information only, the system achieved an accuracy about equivalent to that of untrained humans.

1 Introduction

Robust and accurate facial feature analysis is a difficult object recognition problem, because of large appearance differences between subjects and large appearance variability of a specific subject due to changes in pose, lighting, specularity and mouth opening. It has received much attention in the face recognition community [30, 67, 10, 8, 50, 31] but is even harder for image sequences due to motion and large appearance variability. Most approaches for speechreading have constrained or circumvented the feature extraction problem by marking the subjects? lips with color or reflective markers, by recording the lip movements with a head mounted camera, by using one subject only, by hand segmenting the lip region or by using very controlled lighting conditions.

We describe a method for locating and tracking lips in gray-level image sequences which avoids such constraints. A deformable model is used which learns patterns of typical lip deformation from a training set. This constrains the model during image search to only deform in ways similar to the training examples. Image search is based on an appearance model which is learned from the training set and used to estimate the similarity between the image and the model. This similarity is assumed to be maximal when the model is placed exactly over the actual lip contours. We demonstrate the robustness of the approach on a database of image sequences of various persons.

There has been much progress in automatic speech recognition over the past decade and state of the art systems perform very well in controlled lab environments. However, once these systems are applied to real-world environments, where background noise or cross-talk is present, their performance degrades significantly [25]. Such application environments are for example an office, car, factory or aircraft and essentially all of these are subject to some interfering noise. Much research effort has therefore been directed to systems for noisy speech environments and the robustness of speech recognition systems has been identified as one of the biggest challenges in future research [12].

Most approaches for robust recognition make use of the acoustic speech signal only and ignore the multi-modal nature of human speech. It is however well known that humans make use of the visual modality and that their speech perception is enhanced by seeing the talker's face, particularly the mouth [18]. Hearing impaired and deaf persons make extensive use of visual cues and some few individuals perform lip-reading to such a high degree which enables almost perfect speech perception [60]. But also normal hearing persons make use of visual information to improve their speech perception, especially in noisy environments [58, 26]. In the presence of noise, the visual signal is often complementary to the acoustic signal, i.e. some phonemes which are difficult to understand acoustically are easier to distinguish visually, and vice versa. Thus, the visual signal often provides that information which is acoustically most sensitive to noise. Lip information can also be beneficial when no noise is present. Reisberg et al. [54] have shown that normal-hearing subjects who see the talker?s face perceive speech more accurately, even in noise-free environments. The influence of visual articulation on human perception of speech is demonstrated by the McGurk effect [46] in which subjects mistakenly hear sounds which are biased by visual articulation.

Motivated by these psychological studies, several researchers [21] [66] [44] [64] [57] [56] [4] [24] [5] [7] [63] [19] [48] [9] [42] [55] [33] have developed speechreading systems, mainly to demonstrate the potential use of visual information to improve the robustness of acoustic speech recognition systems in noise. While these systems have validated the benefit of visual speech information, there is still much discussion about determining which visual features are important for speechreading, how to represent them and how to extract them automatically in a robust manner [43].

We describe an approach for feature extraction based on the model we use for lip-tracking. The parameters describing the talking mouth are recovered from the tracking results and describe the shape of the lip contours and the gray-level appearance of the mouth. We describe the application of these features for speechreading, by modeling their distribution by mixtures of Gaussians and their temporal dependencies by Hidden Markov Models (HMMs).

The next section reviews some previous approaches for the extraction and modeling of visual speech features. This is followed by a description for our lip modeling approach. We then evaluate the method and present results for lip localization and tracking. The modeling of the extracted features for speechreading is described and results are presented for a speaker independent digit recognition task.

2 Background

The two main approaches for extracting visual speech information from image sequences can be grouped into image-based and model-based approaches. In the image-based approach, the gray-level image containing the mouth is either used directly or after some pre-processing as feature vector. The advantage of this method is that no data is disregarded. The disadvantage is that it is left to the classifier to learn the nontrivial task of finding the generalization for translation, scaling, rotation (2D and 3D), illumination and inter/intra speaker variability. Another disadvantage is the high dimension and high redundancy of the feature vector.

In the model-based approach, a model of the visible speech articulators, mainly the lip contours, is built and its configuration is described by a small set of parameters. The advantage of this approach is that important features can be represented in a low dimensional space and are normally invariant to translation, scaling, rotation and lighting. A disadvantage is that the particular model used may not consider all relevant speech information. The main difficulty in the model-based approach is the design of the model topology and the design of a robust algorithm which accurately maps the model to the image.

2.1 Image - based approaches

Yuhas et al. [66] presented an approach where the whole gray-level image containing the mouth area was used as feature vector. Similar feature extraction methods were described in [64, 4]. A method which only uses the horizontal and vertical gray-level vectors, centered at the mouth, has been proposed in [63]. Movellan [48] described a feature extraction method where the image was first normalised and symmetrized using the vertical mid-line of the image as the axis of symmetry. He also included pixel by pixel differences in the feature vector, which improved the recognition accuracy considerably.

Silsbee [56, 55] proposed a feature extraction method based on vector quantization, where 17 codevectors describing different mouth configurations were selected by hand. The approach was described as being extremely sensitive to differences in width and height of the lip opening and very sensitive to the presence/absence of teeth.

Mase et al. [44] described an approach based on optical flow, which was calculated on four windows near the mouth. Parameters were extracted from average flow vectors in each window. The underlying assumption of this approach was to estimate the major muscle activities involved in speech production.

A feature extraction method where principal component analysis was performed on the image containing the mouth area has been proposed by Bregler et al. [5] and Brook et al. [7]. Similar approaches were used in [19]. These methods reduce the feature space considerably and are generally less sensitive to noise.

2.2 Model - based approaches

Petajan [51] was probably the first researcher to develop a speechreading system. The system was based on geometric features of the mouth opening, like height, width and area. A simple thresholding technique was used to find the mouth opening and a distance measure, without time-warping, was applied to match test sequences to training templates. Petajan's system was later extended by Goldschen [24] to a continuous speechreading system using HMMs and context dependent sub-word models. Goldschen has determined distinct viseme classes which are visually distinguishable speech units, equivalent to phonemes. He trained context dependent sub-word models based on these visemes, similar to acoustic sub-word modelling, and achieved impressive recognition results, using visual information only. This study has demonstrated that acoustic modelling techniques are also applicable to the visual speech signal. The bottleneck in the system however was the feature extraction, which was not very robust and required manual assistance.

Finn et al. [21] presented a speechreading system where features were extracted by measurements on highly reflective dots, placed around the speaker?s mouth. A similar feature extraction method was used in [57]. These approaches mainly serve to determine the usefulness of visual features but are hardly applicable to practical applications.

Yuille et al. [67] have described the use of deformable templates for locating facial features. Their application to lip-tracking has been described in [29]. The outline of the lips is modeled by a set of hand coded polynomials, which are matched onto the outline of the lips, represented by the image gradient. Since the deformation of deformable templates is constrained by the initial choice of polynomials, they are often not able to resolve fine contour details. The image search is performed by fitting the template to image gradients, assuming strong edges at the lip contours. This assumption is however often violated as the gradient along the contour is dependent on the speaker, illumination, reflection, facial hair, visibility of teeth, and mouth opening. A similar approach but based on color information has been described in [11].

Kass et al. [32] have described active contour models, so-called snakes, for lip-tracking. These are able to resolve fine contour details but shape constraints are difficult to incorporate and one has to compromise between the degree of elasticity and the ability to resolve fine contour details. Bregler et al. [6] described a method based on snakes for tracking the outer lip contour, but where the contour is constrained to lie in a sub-space learned from a training set. The energy function for image search considers the distance of the contour to the sub-space, an internal energy, which is minimal when the contour follows a straight line, and the sum of gradients along the contour. The weights for each contribution were determined empirically. Similar to deformable templates, the approach assumes that image gradients are well suited to represent the lip contours. Bregler et al. found however, that the information provided by their tracking results were not distinctive enough to give reasonable recognition performance. They therefore used the components of the gray-level matrix, centered and scaled around the lips, as speech features.

An approach based on splines [1] and Kalman filters [23] has been described in [17]. Shape constraints were imposed on the deforming template by limiting the number of degrees of freedom. This was performed by fitting the spline to two extreme mouth shapes. Image search was performed by searching for high contrast edges. However, tracking was only stable when a lip-stick was worn to enhance the contrast around the lips. A similar approach based on Bezier curves and the use of color information was described in [62]. The energy function considers internal and external modal forces, dynamic constraints, and color information from the image. Their relative contribution to the energy function was determined empirically.

3 Lip Model

It is often argued that visual information in a speech recognition system is most beneficial in real world applications where noise is present. If the aim of a visual-acoustic speech recognition system is therefore to be used in such an environment, the robustness of the visual subsystem in such an environment is of crucial importance. We believe that a feature extraction method intended to be robust to all the sources of variability encountered in real world applications (illumination, individual appearance, 3D pose) should use as much knowledge about the scene as possible. One way of incorporating such knowledge is to build a model of the object. Most model-based systems developed so far have relied on heuristics about shape deformation and appearance. Our approach learns such knowledge by examining a representative training set.

One of the issue to address in a model based approach is to choose an appropriate description of the visible articulators. We are modeling a physical process, so we could describe this process in terms of physical movements and positions of the articulators that determine the vocal tract. Specifically, for visual analysis, we could attempt to estimate muscle action from the image such as in [44, 20]. However, the musculature of the face is complex, 3D information is not present, muscle motion is not directly observable and there are at least thirteen groups of muscles involved in the lip movements alone [27]. Furthermore, using optical flow computation to estimate such action might not be appropriate due to violation of its underlying assumptions.

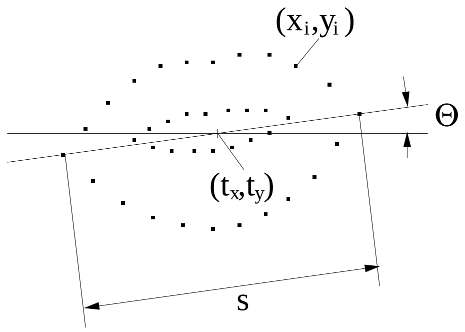

Figure 1: *Model 2* representing the outer and inner lip contour with 38 control points. To generate a certain lip shape, the translation (t_x, t_y), rotation (Θ), and scale (s) are required in addition to the coordinates of the model points.

We chose to use an appearance-based model. In order to construct such a model, we need to determine which features the model will incorporate and how they are represented. The model should represent those features important for speech recognition and disregard those which account for speaker variability and illumination.

It is generally agreed that most visual information is contained in the lip contours, especially the inner lip contour, but it has also been shown that the visibility of teeth and tongue provide important speech cues [47, 45, 60, 3]. Particularly for fricatives, the place of articulation can often be determined visually [59], i.e. for labiodental (upper teeth on lower lip), interdental (tongue behind front teeth) and alveolar (tongue touching gum ridge) place.

We therefore would like to have a model which describes both, the shape of the inner and outer lips and the intensity around the mouth area. We use models based on point distribution models (PDM), also called active shape models (ASM) when used in image search [16, 15, 13, 14]. PDMs are flexible models which represent an object by a set of labeled points. The points describe the boundary or other significant parts of an object. The average shape and the principal modes of variation are captured from a labeled training set.

The training examples need to be labeled in a consistent manner in order to be able to compare equivalent points from different shapes. We use the two outer corner points of the lips as reference points. Their distance is defined as scale, their orientation to the horizontal as the angle and their center as the origin. The other points are placed at equal horizontal distance along the lip contours, where the horizontal is the line connecting the two corners. We built two different models of the lips: *Model 1* describes the outer contour of the upper and lower lips and *Model 2* describes the outer and inner contour of the upper and lower lips. Figure 1 shows *Model 1* with translation t_x and t_y, scale s and angle θ.

3.1 Shape Modeling

By using PDMs we try to avoid the use of heuristic assumptions about legal shape deformation. Instead, knowledge about legal shape deformation is obtained by examining a representative training set. This leads to a description of local and global deformations with a small set of parameters and constrains the shape model to only deform to shapes similar to the ones seen in the training set.

In order to obtain the average shape and the main modes of variation of an object, a training set containing examples of the object is labeled by hand. Each example is labeled in the same way to be able to relate different examples to each other. The ith shape in the training set ($i = 1 \ldots N$) is described by a vector $\mathbf{v}_i$ with

$$\mathbf{v}_i = (x_{i0}, y_{i0}, x_{i1}, y_{i1}, \ldots, x_{iN-1}, y_{iN-1})^T \tag{1}$$

where (x_{ij}, y_{ij}) are the coordinates of the jth point ($j = 0 \ldots N-1$) of the ith shape. The training shapes are then normalized by scaling to unit width, zero translation and zero rotation using

$$\mathbf{x}_i = M\left(\frac{1}{s}, -\theta\right)[\mathbf{v}_i - \mathbf{t}_c] \tag{2}$$

with translation

$$\mathbf{t}_c = (t_x, t_y, t_x, t_y, \ldots, t_x, t_y)^T \tag{3}$$

and $M(s, \theta)[x]$ performing a rotation by θ and a scaling by s of x. The normalization process ensures that model points of different examples can be related without distortion from scale, translation or rotation.

Given a set of N normalized labeled shapes, we can calculate the mean shape $\overline{\mathbf{x}}$ and the covariance matrix $\mathbf{S}$. The eigenvectors and eigenvalues of the covariance matrix are obtained using principal component analysis (PCA). The eigenvectors with the largest eigenvalues describe the most significant modes of variation, in particular the variance described by an eigenvector is equal to its corresponding eigenvalue.

A normalized shape can now be approximated by

$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{P}\mathbf{b} \tag{4}$$

where $\mathbf{P} = (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_t)$ is the matrix of the first t ($t < n$) column eigenvectors corresponding to the largest eigenvalues and $\mathbf{b}$ a vector containing the weights for each eigenvector. The number of eigenvectors used to describe the main modes of variation is normally much smaller than the number of variables of the shape model.

Figure 2 and Figure 3 show the mean shape of the models and the first four modes of variation by ± 2 standard deviation (*s.d.*). We use 8 shape modes to describe *Model 1* and 10 shape modes to describe *Model 2*. In order to project a lip model into an image, the scale, rotation and translation parameters are needed in addition to the shape parameters.

For *Model 1*, the first mode of variation mainly changes the position of the lower lip contour. The second mode primarily describes the position of the upper lip contour and the third mode accounts for asymmetry due to horizontal rotation. Subsequent modes describe finer contour details. For *Model 2*, the first mode mainly changes the lower lip, the second mode changes the upper lip and the third mode describes the mouth opening. Subsequent modes account for asymmetry and finer contour details.

An earlier version of these models which only considered vertical deformation of the model points was described in [37]. The method was sufficient for the single contour model but could not fully describe the large variability of the inner contour of the double contour model.Our approach assumes that the principal modes are linearly independent, although there might be non-linear dependencies present. For objects with non-linear behavior, linear models reduce the specificity of the model and can generate implausible shapes, which lead to less robust image search. They also require more modes of variation than the true number of degrees of freedom of the object.

3.2 Intensity Modeling

In order to use PDMs for image search, we would like to have a cost function, which measures the fit between the model and the image. We therefore need to find a way of representing dominant image features of the lip contours. The most common approach for representing contours is to use edges or gradients. However, the appearance of lip contours is highly variable, even for the same person. The gradient values at the outer lip contour are often strong at the upper lip and weak at the lower lip.

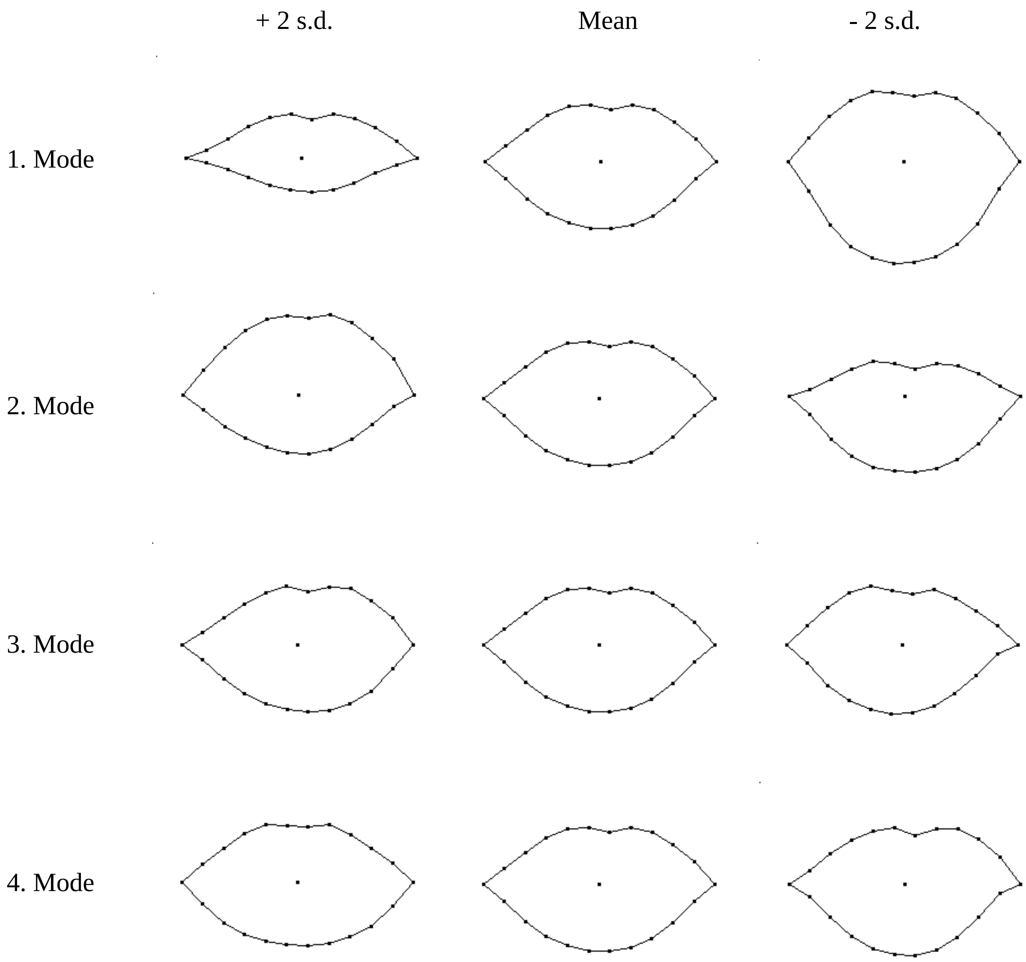

Figure 2: The middle column displays the mean shape for *Model 1* and the other columns show the four most significant modes of shape variation as the sum of the mean shape and different basis shapes.

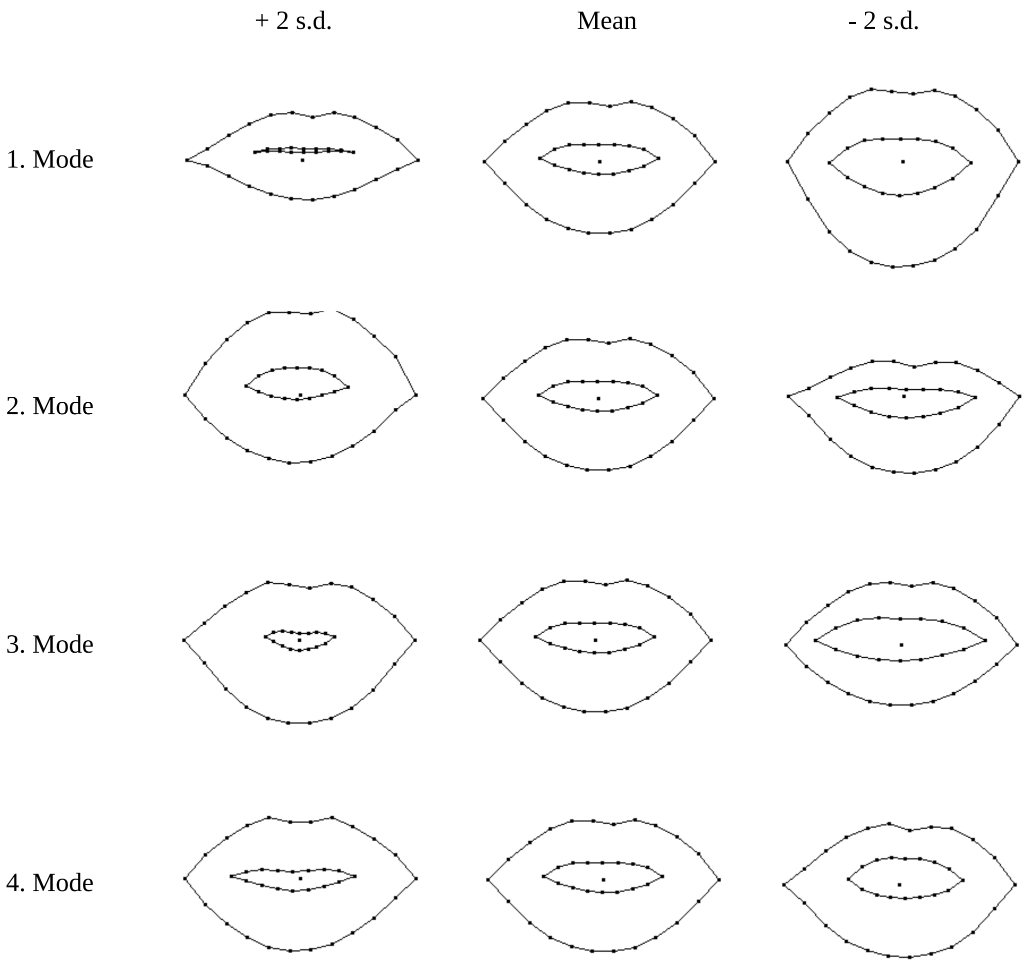

Figure 3: The middle column displays the mean shape for *Model 2* and the other columns show the four most significant modes of shape variation as the sum of the mean shape and different basis shapes.

The gradients at the inner contour are highly dependent on mouth opening and the visibility of teeth and tongue. Gradients are also dependent on the speaker (make up, facial hair, ethnic origin) and illumination. Furthermore, edges of the lip contours are often confused with other gradients, which can originate from specularity, visibility of teeth and tongue, shadows, or facial hair.

Figure 4: Example images of the Tulips1 database (first row) and their gradient magnitude images (second row).

Figure 4 shows some examples from the database we used [48] and their gradient magnitude images after Gaussian smoothing. The examples show clearly the difficulties gradient-based search methods are faced with and that gradient information is not an appropriate way to represent dominant features of lip contours.

In analogy to the statistical description of the lip deformation we want to avoid the use of heuristics for image search and rather learn the gray-level appearance at the contour from a training set. Assuming that gray-level changes are not only important at each contour point but also in regions around each point, we capture the statistics of the actual gray-level appearance around each model point and estimate their main modes of variation within a training set.

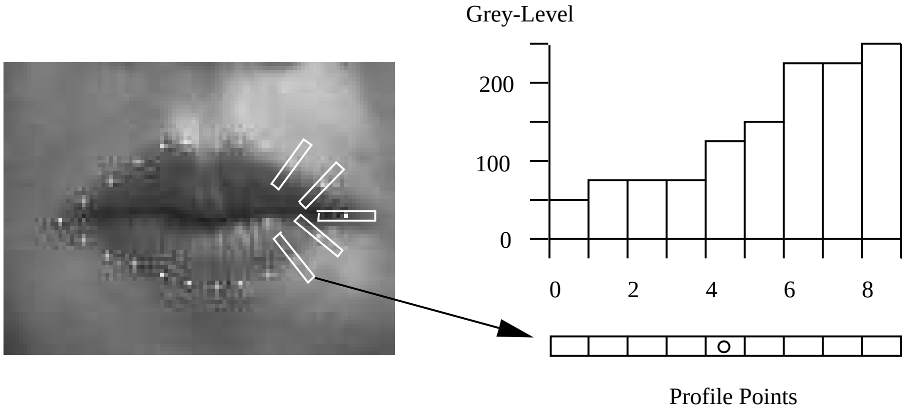

Figure 5: Grey-level profile extraction. The grey-level vectors are sampled perpendicular to the lip contour and centred at the model points.

Following [15], we choose to sample one dimensional profiles $\mathbf{g}_{ij}$ of length n_p perpendicular to the contour and centered at point j for each training image i as shown in Fig. 5. But instead of calculating individual mean profiles and covariance matrices for each model point, we concatenate the profiles of all model points to construct a global profile vector $\mathbf{h}_j$ for each training image as described in [28]:

$$\mathbf{h}_i = (\mathbf{g}_{i0}, \mathbf{g}_{i1}, \ldots, \mathbf{g}_{iN-1})^T \tag{5}$$

We then calculate a global mean profile $\overline{\mathbf{h}}$ and covariance matrix $\mathbf{S}_g$ and compute the eigenvectors and eigenvalues of the covariance matrix $\mathbf{S}_g$. The eigenvectors with the largest corresponding eigenvalues describe the main modes of gray-level variation seen in the training set. Any profile can be approximated using

$$\mathbf{h} = \overline{\mathbf{h}} + \mathbf{P}_g \mathbf{b}_g \tag{6}$$

where $\mathbf{P}_g = (\mathbf{p}_{g1}, \mathbf{p}_{g2}, \ldots, \mathbf{p}_{gt}$ is the $(n \star n_p) \times t$ matrix of the first t column eigenvectors corresponding to the largest eigenvalues and $\mathbf{b}_g$ a vector containing the weights for each eigenvector.

These weights describe the intensity around the lip contour and can therefore be used as speech features. This approach is similar to the local gray-level models described by Lanitis et al. [34, 36, 35], who performed an individual PCA on each model point. We use global gray-level models, assuming that variances at different model points are correlated. It is also related to the eigenlips reported by Bregler et al. [5]. Bregler et al. placed a window around the mouth area on which PCA was performed. Since the window does not deform with the lips, the eigenvectors of the PCA mainly account for intensity variation due to different lip shape and mouth opening. In comparison, our PCA space deforms with the lip contours and therefore describes intensity variation independent of lip shape. The intensity information we obtain is therefor complementary to the contour information provided by the shape model.

In order to visualize gray-level models, we interpolate the gray-levels between the profile vectors and smooth them with a Median filter. The mean intensity and the first three modes of intensity variation are shown in Fig. 6. All gray-level models are displayed using the mean shape. Although this might generate unrealistic appearances, since the gray-levels are correlated with the mouth opening, we considered it to be acceptable for visualization purposes.

The gray-level models show the area covered by the profile vectors: the mouth opening, the lip area, and the skin around the lips. The first mode accounts for global illumination, the second mode mainly describes the intensity of the lower lip and the third mode explains the contrast between the skin and the lips. Subsequent modes describe finer variations, such as lighting direction, specularity, and visibility of teeth and tongue.

This approach assumes that the variances of profiles at different model points are correlated with each other as they are expected to be due to illumination effects and different skin and lip intensity. The profile model captures the global variation between different speakers and the variation for individual speakers. Particularly the intensity variation inside the mouth is subject to the largest intensity variation during speech production. In analogy to shape modeling, we assume that the intensity modes are linearly independent.

4 Database

We used the image part of the Tulips 1 database of isolated digits [48] for all experiments. It consists of 96 image sequences of 12 speakers (9 male, 3 female) each saying the first four digits in English twice. We will later refer to the set of words spoken the first time as Set 1 and the set of words spoken the second time as Set 2. The subjects were asked to talk into a video camera and to position themselves so that their lips be roughly centered in a feed-back display. The gray-scale images were digitized at 30 frames/sec., 100 x 75 pixels, 8 bits per pixel. The database contains a total of 934 images and consists of speakers with different ethnic origin, some with make-up or facial hair and different illumination.

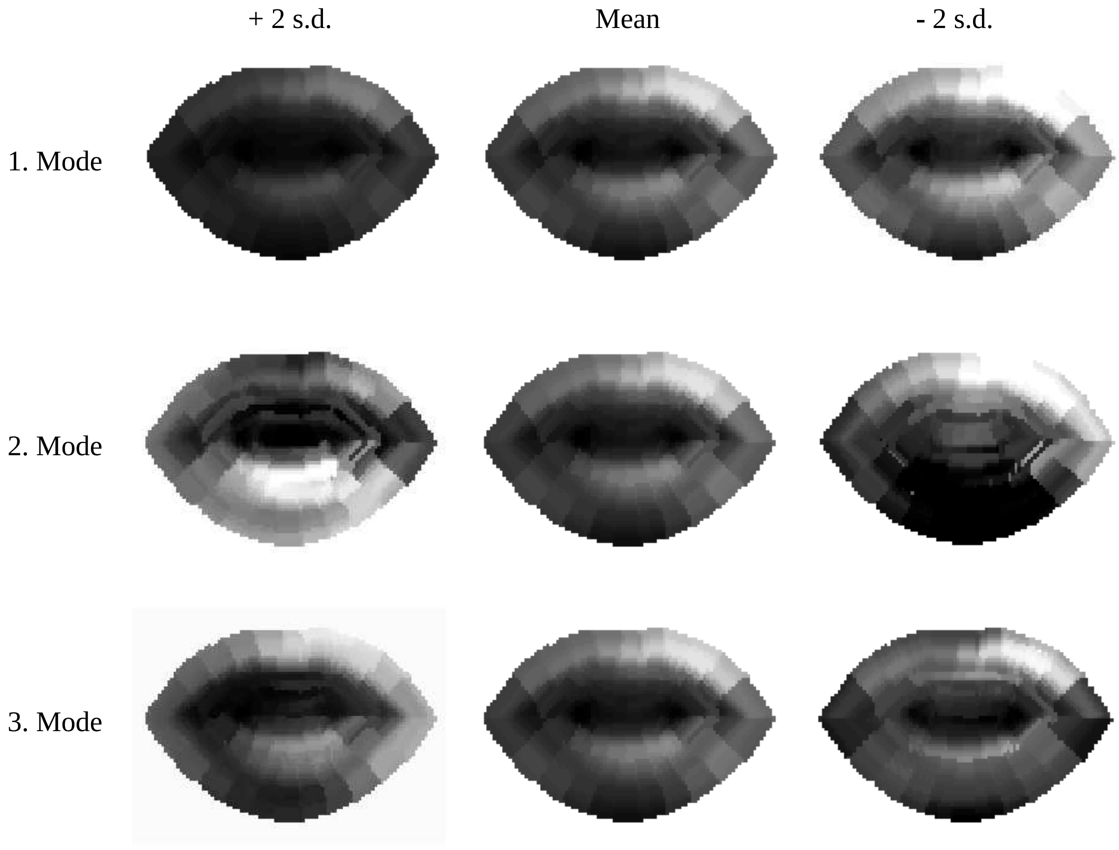

Figure 6: The principal modes of intensity variation of *Model 2*. The middle column displays the mean intensity and the other columns represent the three most significant modes of intensity variation for ± 2 s.d.

5 Locating and Tracking Lips

5.1 Cost Function

As a cost function we use a measure which describes the fit between the profile model and the image. To account for gray level variation captured in the training set, the model profile is first aligned to the image profile $\mathbf{h}$ as closely as possible, using the mean mode and the first main modes of variation. As the eigenvectors in $\mathbf{P}_g$ are orthogonal, the parameter vector $\mathbf{b}_g$ describing the weights for the modes can be found using

$$\mathbf{b}_g = \mathbf{P_g}^T(\mathbf{h} - \overline{\mathbf{h}}). \tag{7}$$

To measure how well a model fits the image, we measure the mean squared error (MSE)E_p between the image profile and the aligned model profile using

$$E_p = (\mathbf{h} - \overline{\mathbf{h}})^T(\mathbf{h} - \overline{\mathbf{h}}) - \mathbf{b_g}^T\mathbf{b_g}. \tag{8}$$

Unlike other methods [29, 6, 62], we do not use an energy component for shape deformation. Using a shape constraint enforces conformity of the solution with a grouped behavioral statistic, while attempting to locate a specific instance from the distribution. Instead we restrict each shape mode to stay within $\pm 3 s.d.$ during image search, which accounts for 99% of variation. We assign equal prior probabilities to all model shapes within these limits.

5.2 Image Search

It is assumed that a coarse estimate of the region of interest (ROI) containing the lips and of the scale of the lips is known. The Downhill Simplex Method [49, 52] is used for image search to match the model to the image. The model is initialized with the mean shape and placed in a random location in the ROI which is the starting point for the search algorithm. The algorithm uses the translation parameters t_x and t_y, the scale s, the rotation θ and the vector of shape parameters $\mathbf{b}$ as variables for the multidimensional optimization process. The search process begins by evaluating initial perturbations of each parameter, which were chosen to be $2 s.d.$. An evaluation is performed in two steps, at first the optimal profile weights are calculated using Eq. 7 followed by the calculation of the cost using Eq. 8. It gradually moves and deforms the model to shapes which give a lower cost and changes the perturbations dynamically until a certain number of iterations is reached or until the difference in cost falls below a certain threshold. The algorithm has proven to be robust to various initial parameters and copes well with local minima.

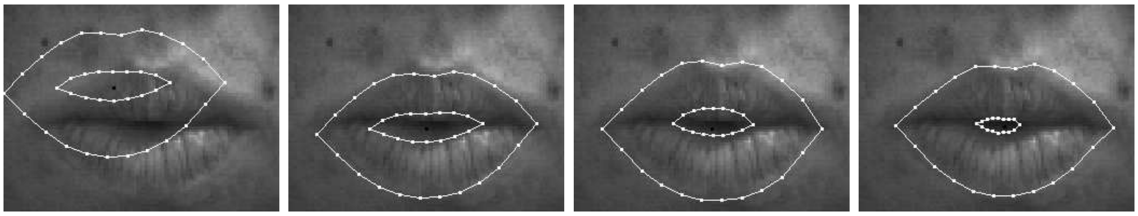

Figure 7: Off-centre initialisation of *Model 2* and image search results after 5, 10, and 20 iterations.

Lip-tracking is initiated by locating the lips in the first image as described above. For consecutive frames, the previous frame is used as the initial estimate of the lip position and the search is performed using the simplex algorithm. Although constraints could be introduced to limit the search to stay within certain limits during tracking, for simplicity, we used the same constraints as for locating the lips.

5.3 Experiments

The shape and profile models were built using 190 training examples for *Model 1* and 250 examples for *Model 2* drawn from all 12 speakers and covering a representative set of mouth shapes. All examples were taken from Set 1 and contained hand labeled points. *Model 1* was represented by 22 points, *Model 2* by 38 points and the dimension of a gray-level profile was 19. We used 8 shape modes for *Model 1*, which covers 87.1% of shape variation and 10 shape modes for *Model 2*, which covers 84% of shape variation, estimated from the corresponding covariance matrices. 12 intensity modes were used in image search for *Model 1* and 20 modes for *Model 2*.

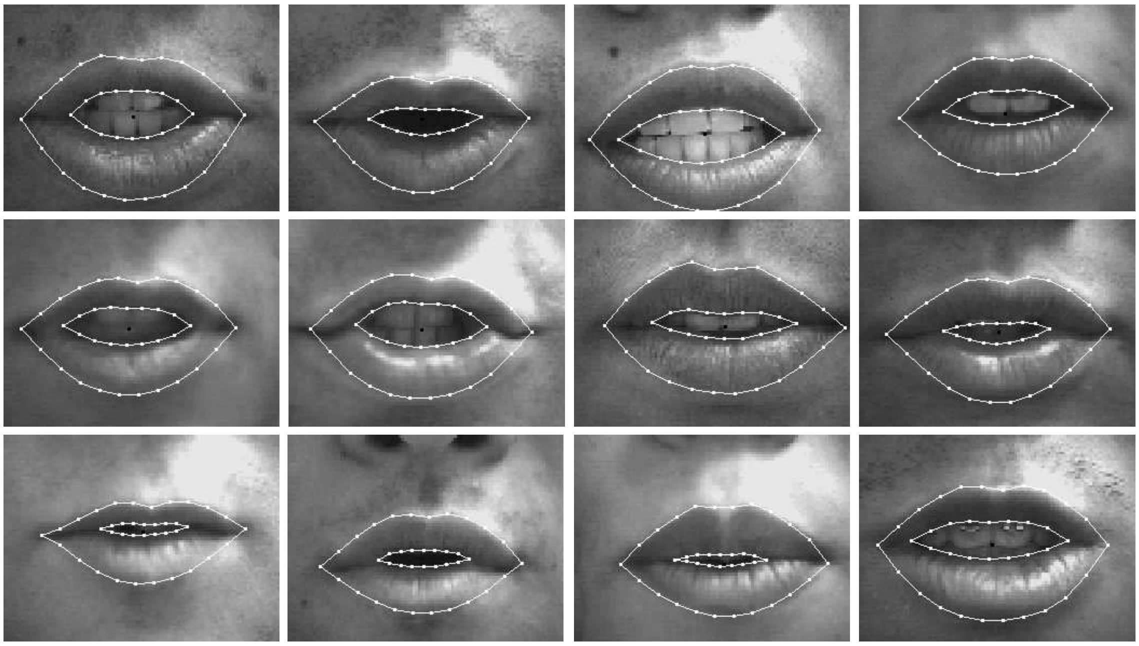

Figure 8: Examples of lip localisation results for all subjects, using *Model 2* and intensity model based search, which were classified as *Good*. Image search copes well with different subjects and different illumination and finds the inner lip contour despite the large intensity differences due to the visibility of the teeth and mouth opening.

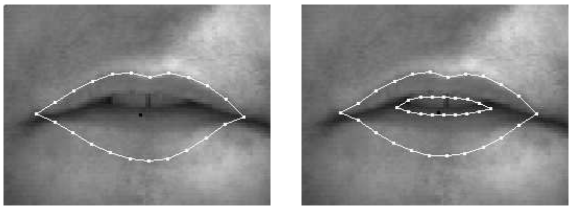

Figure 9: Examples of lip localisation results using intensity model based search, which were classified as *Miss*.

It is important to evaluate the performance of computer vision algorithms quantitatively. Ideally, this would require all images to be labeled with the correct outline of the lips, which would need to be done by hand. Instead we choose to judge the performance by visual inspection into three categories. A search result was classified as Good if the lip contour was found within about one quarter of the lip thickness deviation. It was classified as Adequate if the outline of the contour was found between one quarter and half the lip thickness deviation and it was classified as a Miss otherwise. For tracking results, all of the images had to be classified with Good in order to classify the whole sequence as Good. The same was true for Adequate.

Table 1: Results for lip localization using profile models with off-center initialization.

Model	Good (%)	Adequate (%)	Miss (%)
Model 1	97.9	0	2.1
Model 2	97.9	0	2.1

All tests for locating and tracking lips were performed using Set 2. For localization tests, the position of the model was initialized at a position off the center. This makes the search process more realistic, assuming that only a rough estimate of the lip position is known. The shape parameters were initialized with the mean shape parameters. Figure 7 shows the initial placement of the model in the image and results of the image search after a few iterations. Examples of lip localization for all subjects are shown in Fig. 8. The two localization results classified as Miss are depicted in Fig. 9. Localization results for both models can be found in Table 1.

For comparison, we also performed localization tests by using the gradient magnitude of the image instead of the profile model. The images were first Gaussian filtered before calculating the gradient magnitude. The resulting images were smoothed with an exponential function to blur the gradients over a large image area and to make the search more robust. The cost was defined as the negative sum of the gradient image at all model points. This method is similar to the ones described in [29, 6]. Because the results for locating the lips were so poor, we performed another test where the model was initialized at the center of the image.

Table 2: Results for lip localization using gradient search.

Model	Initialization	Good (%)	Adequate (%)	Miss (%)
Model 1	off center	6.3	6.3	87.5
Model 2	off center	4.2	12.5	83.3
Model 1	center	18.8	22.9	58.3
Model 2	center	20.8	8.3	68.8

Table 2 shows the results for locating the lips by using gradient information. Miss-localization was mainly caused by the small gradient of the outer lower lips, reflections on the lips and gradients originating from the teeth. Initializing the model in the center of the image improved the results only slightly. These results show clearly the superiority of gray-level models over gradients in image search.

Table 3: Results for lip tracking.

Model	Good (%)	Adequate (%)	Miss (%)
Model 1	95.8	2.1	2.1
Model 2	91.7	6.3	2.1

For tracking tests, the model was initialized at the center of the image. Tracking examples of Good performance are shown in Fig. 10 and for Adequate performance in Fig. 11. Table 3 summarizes tracking results for both models. For *Model 1*, results for lip-tracking are similar to the ones for localization. For *Model 2*, the performance for lip-tracking is lower than for localization. This was mainly due to the model aligning to the teeth instead of the inner lip contour.

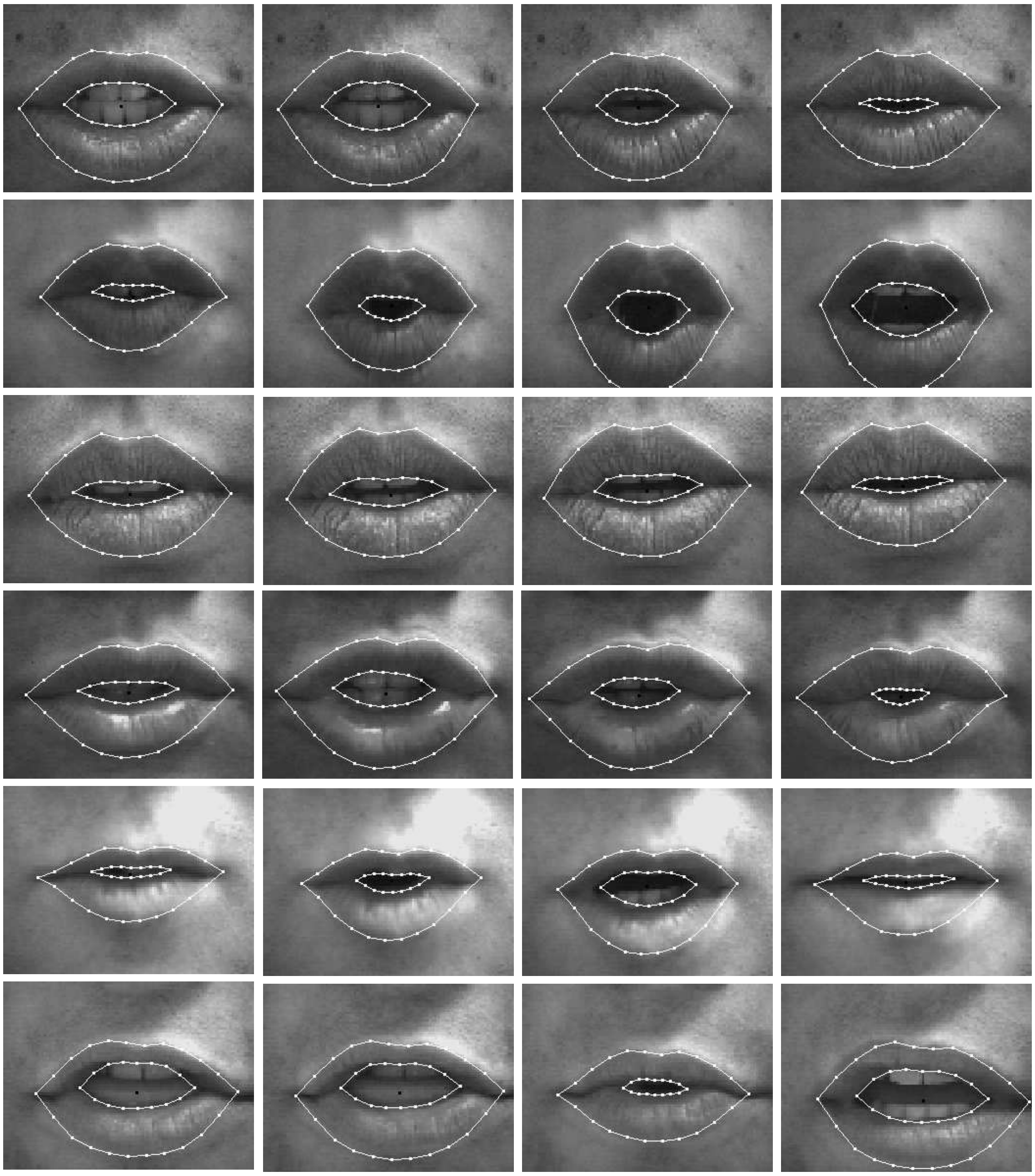

Figure 10: Examples of lip tracking results using *Model 2* and intensity model based search, which were classified as *Good*. The tracking results are very accurate across subjects despite large appearance differences and varying visibility of teeth and tongue.

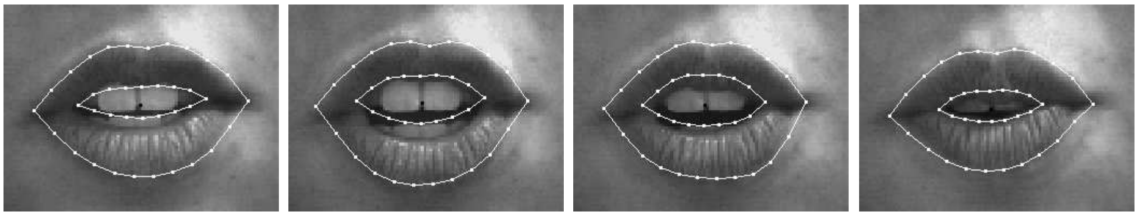

Figure 11: Example of a lip tracking sequence using *Model 2* and intensity model based search, which was classified as *Miss*. The inner model contour first latched onto the teeth and later missed the upper lip contour.

The assumption of a linear model seems to be appropriate for shape modeling but might be inadequate for profile modeling. Most of the errors were due to the inner lip contour latching on to the teeth. The profiles covering the mouth opening basically change between three different intensity levels: (1) mouth closed, (2) mouth open and teeth not visible, (3) mouth open and teeth visible. It is therefore unlikely that this variability can accurately be described by a uni modal distribution.

6 Speechreading

This section describes the modeling of the extracted features for a speech recognition system. The task is to recognize the first four digits in English using visual information only. One of the main difficulties in speechreading is to cope with the large variability across speakers, due to individual appearance and individual lip movements (see for example Fig. 10.). We therefore performed speaker independent tests, using different speakers for training and testing, to see how well the system generalizes to new speakers.

6.1 Visual Speech Features

We use two kind of parameters as visual speech features, shape and intensity parameters. Both sets of parameters are extracted at each image frame and are recovered from the tracking results. The shape vector is represented by $\mathbf{b}$ and the intensity vector by $\mathbf{b}_g$. These parameters describe variability due to different speakers and variability due to speech. The aim of our speech recognition system is to learn which features account for different utterances and which for different speakers.

We do not include translation, rotation and scale parameters in the feature vector since they are unlikely to provide speech information and since they could distort our results. All features are therefore invariant to translation, rotation and scale. The shape features are also invariant to illumination.

Much visual speech information is contained in the dynamics of lip movements rather than the actual shape. Furthermore, dynamic information might be more robust to linguistic variability, i.e. intensity values of the lips and skin will remain fairly constant during speech, while intensity values of the mouth opening will vary during speech. On the other hand, intensity values of the lips and skin will vary between speakers, but temporal intensity changes might be similar for different speakers and robust to illumination. We therefore performed some recognition tests by including first order differential parameters of the feature vector (delta features). The change in scale is likely to provide speech information and was also included in the delta feature vector.

6.2 Speech Modeling

For modeling visual speech, we use whole-word Hidden Markov Models (HMMs) which is a standard approach in acoustic small vocabulary recognition systems [53]. A visual observation $\mathbf{O}$ of an utterance is represented by a sequence of feature vectors

$$\mathbf{O} = \mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T, \tag{9}$$

where $\mathbf{o}_t$ is the feature vector extracted at time t. We assume that the feature vectors follow continuous probability distributions which we model by mixtures of Gaussians. We further assume that temporal changes during speech are piece-wise stationary and follow a first-order Markov process. Thus, each HMM state represents several consecutive feature vectors. These assumptions are not strictly true, but are also often made in the acoustic domain. They can be improved by increasing the number of states, which decreases the number of times an observation frame remains at a particular state. But this increases the complexity of the models and the number of parameters to estimate.

A HMM representing a particular utterance class is defined by the parameter set

$$\lambda = (\mathbf{A}, \mathbf{B}, \pi) \tag{10}$$

$\mathbf{A} = \{a_{ij}\}$ is the matrix of state transition probabilities from state i to state j, $\mathbf{B}$ is the vector of observation probabilities $b_j(\mathbf{o})$ for state j, and π is the vector with probabilities π_i of entering the model at state i. The observation probabilities are modelled as mixtures of Gaussian distributions:

$$b_i(\mathbf{o}) = \sum_{m=1}^{M} c_{im} N(\mathbf{o}, \mu_{im}, \Sigma_{im}) \tag{11}$$

where c_{im} is the mixture weight for state i and mixture m, and $N(\mathbf{o}, \mu, \Sigma)$ a multivariate Gaussian with mean μ and covariance matrix Σ.

We trained one HMM for each word class on the corresponding training set. The HMMs only allowed self-loops and sequential transitions between the current and the next state. The initial state probabilities are set to zero for all states but the first. The remaining parameters are estimated from the extracted model parameters of the training set. Each HMM is initialised by linear segmentation of the training vectors onto the HMM states, followed by iterative segmental k-means clustering and Viterbi alignment [61]. The models are further re-estimated using the Baum-Welch procedure [2], which maximises the likelihood of model λ for having generated the observed sequence $\mathbf{O}$.

Figure 12 shows the HMM states learned from 11 speakers, for the words 'one' and 'three'. The feature vector consisted of 10 shape and 20 profile parameters. The differences of lip shape and intensity at the mouth opening, as well as their temporal differences can be clearly seen on these examples.

The mouth opening in the first two states is smaller for model 'one' than for model 'three' but at state three it is smaller for model 'three'. The intensity inside the mouth is higher at the end for both models, indicating the visibility of teeth, while for model 'three' it is also high at the first state, indicating the visibility of the tongue.

Recognition is performed using the Viterbi algorithm which calculates the most likely state sequence for each HMM of having generated the observed sequence. Classification is performed by estimating the maximum *a posteriori* probability (MAP)

$$\arg\max_i P(\lambda_i|\mathbf{O}) \tag{12}$$

where λ_i represents the model of word class i and $\mathbf{O}$ the observation sequence. The *a posteriori* probability can be obtained using Bayes rule ,

$$P(\lambda_i|\mathbf{O}) = \frac{P(\mathbf{O}|\lambda_i)P(\lambda_i)}{p(\mathbf{O})}, \tag{13}$$

where $P(\lambda_i)$ represents the *prior* probability of class i and $P(\mathbf{O}|\lambda_i)$ the probability distribution of the feature sequence $\mathbf{O}$ for model λ_i. The terms $P(\lambda_i)$, which are assumed to be equal for all classes, and $p(\mathbf{O})$ are constant for all classes and can therefore be ignored in the MAP calculation. $P(\mathbf{O}|\lambda_i)$ is simply the product of the transition probabilities a and the output probabilities $b(y)$ of the most likely state sequence.

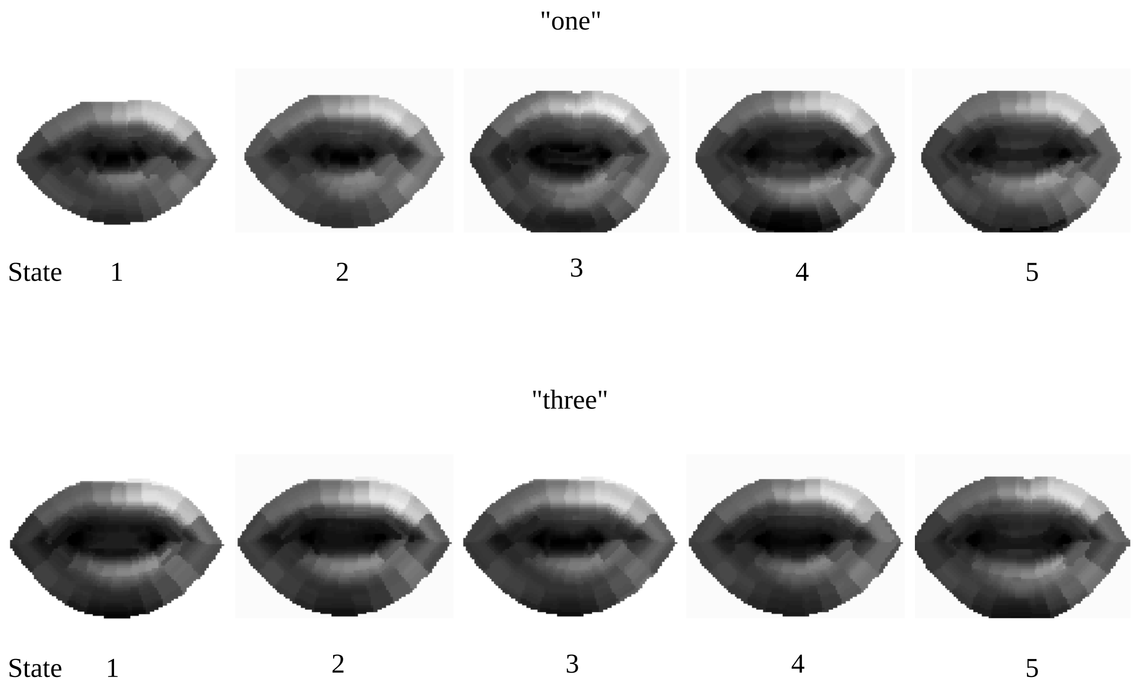

Figure 12: Learned sequence of HMM states for the words *one* and *three* using *Model 2*.

6.3 Experiments

We performed speaker independent recognition tests, using different speakers for training and testing to see how well the system generalizes for new speakers. Experiments were performed using the jack-knife or leaving-one-out procedure [22], using 11 subjects for training and the 12th subject for testing. The whole procedure was repeated 12 times, each time leaving a different subject out for testing. Recognition results were then averaged over all speakers. All tests were performed for *Model 1* and *Model 2* using the HMM toolkit HTK V1.5 [65].

We trained HMMs with 6 states and one diagonal variance vector. Experiments using full covariance matrices or more than one diagonal variance vector resulted in lower performance. This indicates that the training set was not large enough to estimate more than one diagonal mixture reliably.

Table 4: Word recognition rate using all shape and intensity features.

Model	Shape	Intensity	Shape + Intensity
Model 1	81.3 %	78.1 %	82.3 %
Model 2	77.1 %	83.3 %	88.5 %

For experiments using all shape modes and all intensity modes in the features vectors as well as all delta features, we obtain recognition rates of 82.3% for *Model 1* and 88.5% for *Model 2*. More details are given in Table 4. However, features which describe principal shape and intensity variability might not be directly related to speech information. For example, shape variability might account for lip shapes of different persons and intensity variability might be due to illumination and different skin and lip intensity of different subjects. Although the training algorithm ideally learns the features which are important for word discrimination, in practice the training set is rarely large and balanced enough to ensure this.

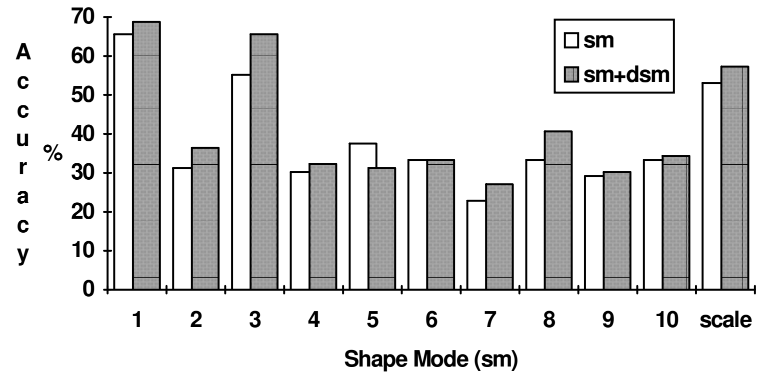

Figure 13: Recognition accuracy for each individual shape mode (sm) and optional delta shape mode (dsm) using *Model 2*.

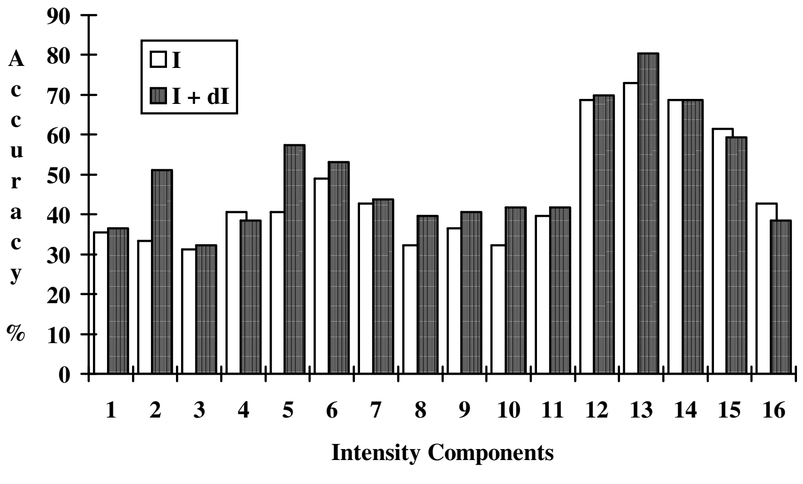

Figure 14: Recognition accuracy for each individual intensity mode (I) and optional delta intensity mode (dI) using *Model 2*.

We therefore performed another set of tests by using each shape and intensity mode individually. Results for *Model 2* are depicted in Fig. 13 and Fig. 14. Results for *Model 1* for shape features can be found in [41] and for intensity features in [40]. For our recognition task of 4 digits, the first and third shape mode, the scale and the intensities modes 12-14 contribute most to recognition performance. The shape modes are shown in Fig. 3 and the 12th-14th intensity modes are displayed in Fig. 15. The three intensity modes are hard to interpret but mainly seem to account for different intensity at the mouth opening.

Table 5: Word recognition rate using the five most discriminant features and optional delta parameters.

Features	Model 1	Model 2
Shape	72.9 %	75.0 %
Intensity	65.6 %	89.6 %
Shape + Intensity	84.4 %	87.5 %
Shape + Δ	83.3 %	81.3 %
Intensity + Δ	82.3 %	89.6 %
Shape + Intensity + Δ	86.5 %	90.6 %

The word recognition rates using only the 5 most discriminant modes (2 shape features and 3 intensity features) are shown in Table 5. Separate tests were performed by including delta parameters and delta scale. All recognition tests were higher by using only the most discriminant features. It is interesting to note that *Model 1* obtained higher rates for shape parameters than *Model 2*, when delta information was included. But this might be due to the small training set. For the intensity features, recognition rates are considerably higher for *Model 2* than for *Model 1*. This seems natural, since *Model 2* provides gray-level information of the mouth opening. Including delta parameters increased the accuracy in almost all cases, indicating the importance of dynamic information and their robustness to speaker variability and illumination. The confusion matrix is shown in Table 6 and the word accuracy for each individual test subject in Table 7.

Table 6: Confusion matrix for the system with 90.6 % recognition rate (rows represent the actual digits, columns the recognised digits).

	one	two	three	four
one	23	0	1	0
two	0	24	0	0
three	2	0	21	1
four	2	0	3	19

Table 7: Word accuracy for each individual test person using the system with 90.6% recognition rate.

Subject	1	2	3	4	5	6	7	8	9	10	11	12
Accuracy (%)	100	87.5	87.5	75	100	100	75	100	100	75	100	87.5

Overall best results of 90.6% were obtained by *Model 2*, using shape and intensity features with additional delta features. This is approximately equivalent to the performance achieved by humans with no lip-reading knowledge who were asked to lip-read on the same database. Those subjects

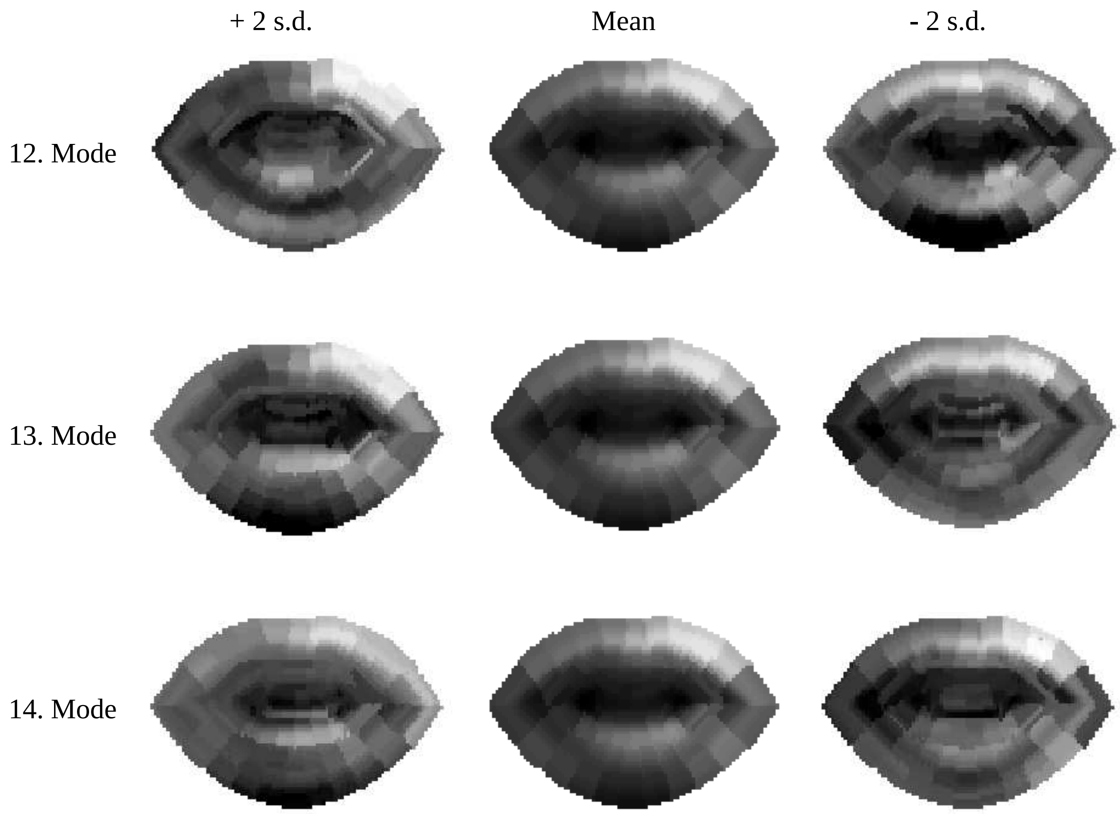

Figure 15: Intensity modes 12-14 for the mean shape of *Model 2*, which contributed most to recognition performance.

achieved and average of 89.93% while the average performance of hearing impaired subjects with lip-reading knowledge was 95.49% [48].

7 Conclusions

We have described a system for robust localization and tracking of lip-movements in image sequences of various speakers and various lighting conditions, without the use of aids like reflective markers or lipstick. The model can only deform in ways consistent with the training set. Similarly, image search is based on the statistical appearance of the mouth, lea rned from training data. These two techniques avoid the use of heuristic assumptions about shape deformation and the representation of dominant image features and therefore enable robust image search. We have shown that the performance for locating and tracking lips is much higher using a gray-level model than by using gradient search.

Our database did not include images with large rotations in 3D, but is has been shown in [36] that a similar model of the face is able to recover vertical and horizontal rotation of at least $\pm$ 20 degrees. It should therefore also be possible for our lip model to recover such rotation and to provide features which are robust to such variability.

Speech features are recovered from tracking results and describe shape and intensity appearance of the mouth. The features are invariant to scale, translation and rotation. The shape features are also invariant to illumination. The sample space of the intensity model deforms with the shape model and therefore represents intensity features which are independent of the lip shape and which can account for information about teeth, tongue and protrusion.

We have described a speechreading system, purely based on visual features, which obtained performance levels similar to humans on a digit recognition task. Visual features vary considerably across speaker due to different appearance and different mouth movements, which makes speaker independent speechreading very difficult. This is one of the first studies to perform speaker independent speechreading and we have shown that very few features are sufficient to achieve high recognition performance for this task. Result for our recognition task suggest that intensity information is more important than shape information.

The emphasis of our work was to describe a robust feature extraction method rather than to describe a large vocabulary speechreading system. We believe that the bottleneck of most existing systems developed so far is their feature extraction method. Several approaches have already shown that once visual features are available, their modeling for continuous speech recognition tasks is straight forward.

The extracted features provide detailed information about shape, intensity, and their temporal dependencies during speech production. They therefore do not only contain speech information but also specific information about a persons articulators and about the way a person speaks. We have described elsewhere how to exploit this information by building models of talking faces and using them for person identification [38, 39].

8 Acknowledgments

This work was partially funded by the University of Sheffield, the German Academic Exchange Service (DAAD), and the Swiss Office for Education and Science (BBW) within the framework of the European ACTS-M2VTS project. The authors would like to thank Steve Renals and the anonymous reviewers for critical reading of the manuscript and Javier Movellan for the use of his database. The image processing algorithms were developed under the TINA vision research environment, which is available at http://www.shef.ac.uk/~eee/esg/research/tina.html.

References

[1] R. Bartels, J. Beatty, and B. Barsky. *An Introduction to Splines for Use in Computer Graphics and Geometry Modelling*. Morgan Kaufmann, 1987.

[2] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occuring in the statistical analysis of probabilistic functions of markov chains. *Annals of Mathematical Statistics*, 41:164–171, 1970.

[3] C. Benoît, T. Guiard-Martigny, B. Le Goff, and A. Adjoudani. Which components of the face do humans and machines best speechread? In David G. Stork and Marcus E. Hennecke, editors, *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, pages 315–328. Springer Verlag, Berlin, 1996.

[4] C. Bregler, H. Hild, S. Manke, and A. Waibel. Improving connected letter recognition by lipreading. In *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, pages 557–560, Minneapolis, 1993. IEEE.

[5] C. Bregler and Y. Konig. Eigenlips for robust speech recognition. In *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, pages 669–672, Adelaide, 1994.

[6] C. Bregler and S. M. Omohundro. Surface learning with applications to lipreading. In J.D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6. Morgan Kaufmann, 1994.

[7] N. M. Brooke, M. J. Tomlinson, and R. K. Moore. Automatic speech recognition that includes visual speech cues. In *Proceesings of the Institute of Acoustics*, volume 16, pages 15–22, 1994.

[8] R. Brunelli and T. Poggio. Face recognition: Features versus templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(10):1042–1052, Oct 1993.

[9] G. I. Chiou. *Active Contour Models for Distinct Feature Tracking and Lipreading*. PhD thesis, University of Washington, 1995.

[10] G. Chow and X. Li. Towards a system for automatic facial feature detection. *Pattern Recognition*, 26(12):1739–1755, 1993.

[11] T. Coianiz, L. Torresani, and B. Capril. 2D deformable models for visual speech analysis. In David G. Stork and Marcus E. Hennecke, editors, *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, pages 391–398. Springer Verlag, Berlin, 1996.

[12] R. Cole, L. Hirschmann, L Atlas, and et al. The challenge of spoken language processing: research directions for the nineties. *IEEE Trans. on Speech and Audio Processing*, 3(1):1–20, 1995.

[13] T. F. Cootes, A. Hill, C. J. Taylor, and J. Haslam. Use of active shape models for locating structures in medical images. *Image and Vision Computing*, 12:355–365, Jul-Aug 1994.

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

[15] T. F. Cootes, C. J. Taylor, A. Lanitis, D. H. Cooper, and J. Graham. Building and using flexible models incorporating grey-level information. In *Proceedings of the International Conference on Computer Vision*, pages 242–246, 1993.

[16] T. F. Cootes and C. J. Taylor. Active shape models - smart snakes. In *Proceedings of the British Machine Vision Conference*, pages 266–275. Springer Verlag, 1992.

[17] B. Dalton, R. Kaucic, and A. Blake. Automatic speechreading using dynamic contours. In David G. Stork and Marcus E. Hennecke, editors, *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, pages 373–382. Springer Verlag, Berlin, 1996.

[18] B. Dodd and R. Campbell, editors. *Hearing by Eye: The Psychology of Lip-reading.* Lawrence Erlbaum Associates Ltd., London, 1987.

[19] P. Duchnowski, U. Meier, and Alexander Waibel. See me, hear me: Integrating automatic speech recognition and lip–reading. In *International Conference on Spoken Language Processing*, 1994.

[20] I. A. Essa and A. P. Pentland. Facial expression recognition using a dynamic model and motion energy. In *Proc. 5th Int. Conf. on Computer Vision*, pages 360–367. IEEE Computer Society Press, July 1995.

[21] E. Kathleen Finn and Alan A. Montgomery. Automatic optically based recognition of speech. *Pattern Recognition Letters*, 8(3):159–164, 1988.

[22] K. Fukunaga. *Introduction to statistical pattern recognition.* Academic Press Ltd., London, 2 edition, 1990.

[23] A. Gelb, editor. *Applied Optimal Estimation.* MIT Press, Cambridge, MA, 1974.

[24] A. J. Goldschen. *Continuous Automatic Speech Recognition by Lipreading.* PhD thesis, George Washington University, Washington, D. C., 1993.

[25] Y. Gong. Speech recognition in noisy environments: A survey. *Speech Communication*, 16:261–291, 1995.

[26] K. W. Grant and L. D. Braida. Evaluating the articulation index for auditory-visual input. *Journal of the Acoustical Society of America*, 89(6):2952–2960, 1991.

[27] W. J. Hardcastle. *Physiology of Speech Production.* Academic Press, New York, NY, 1976.

[28] J. Haslam, C. J. Taylor, and T. F. Cootes. A probabilistic fitness measure for deformable template models. In *Proceedings of the British Machine Vision Conference*, pages 33–42, 94.

[29] M. E. Hennecke, K. V. Prasad, and D. G. Stork. Using deformable templates to infer visual speech dynamics. In *28th Annual Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, November 1994.

[30] C. L. Huang and C. W. Chen. Human facial feature extraction for face interpretation and recognition. *Pattern Recognition*, 25(12):1435–1444, Dec 1992.

[31] N. Intrator, D. Reisfeld, and Y. Yeshrum. Face recognition using a hybrid supervised/unsupervised neural network. *Pattern Recognition Letters*, 1996.

[32] M. Kass, A. Witkin, and Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, pages 321–331, 1988.

[33] R. Kaucic, B. Dalton, and A. Blake. Real-time lip tracking for audio-visual speech recognition applications. In *Proceedings of the European Conference on Computer Vision*, volume 2, pages 376–386. Cambridge, 1996.

[34] A. Lanitis, C. J. Taylor, and T. F. Cootes. An automatic face identification system using flexible apperance models. In *Proceedings of the British Machine Vision Conference*, pages 65–74, 1994.

[35] A. Lanitis, C. J. Taylor, and T. F. Cootes. Automatic face identification system using flexible appearance models. *Image and Vision Computing*, 13:393–401, Jun 1995.

[36] A. Lanitis, C. J. Taylor, and T. F. Cootes. A unified approach to coding and interpreting face images. In *Proc. 5th Int. Conf. on Computer Vision*, pages 368–373. IEEE Computer Society Press, July 1995.

[37] J. Luettin, N. A. Thacker, and S. W. Beet. Active shape models for visual speech feature extraction. In D. G. Storck and M. E. Hennecke (editors), editors, *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, pages 383–390. Springer Verlag, Berlin, 1996.

[38] J. Luettin, N. A. Thacker, and S. W. Beet. Learning to recognise talking faces. In *Proceedings of the International Conference on Pattern Recognition (ICPR'96)*, volume IV, pages 55–59. IAPR, 1996.

[39] J. Luettin, N. A. Thacker, and S. W. Beet. Speaker identification by lipreading. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP'96)*, volume 1, pages 62–65, 1996.

[40] J. Luettin, N. A. Thacker, and S. W. Beet. Speechreading using shape and intensity information. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP'96)*, volume 1, pages 58–61, 1996.

[41] J. Luettin, N. A. Thacker, and S. W. Beet. Statistical lip modelling for visual speech recognition. In *Proceedings of the 8th European Signal Processing Conference (Eusipco'96)*, volume I, pages 137–140, 1996.

[42] J. Luettin, N. A. Thacker, and S. W. Beet. Visual speech recognition using active shape models and hidden Markov models. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96)*, volume 2, pages 817–820, 1996.

[43] J. Luettin, M. Vogt, and C. Bregler. Machine recognition and applications. In D. G. Storck and M. E. Hennecke (editors), editors, *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, pages 549–555. Springer Verlag, Berlin, 1996.

[44] K. Mase and A. Pentland. Automatic lipreading by optical flow analysis. *Systems and Computers in Japan*, 22(6), 1991.

[45] M. McGrath, A. Q. Summerfield, and N. M. Brook. Roles of lips and teeth in lipreading vowels. In *Proceedings of the Institute of Acoustics*, volume 6, pages 401–408, 1984.

[46] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[47] A. A. Montgomery and P. L. Jackson. Physical characteristics of the lips underlying vowel lipreading performance. *Journal of the Acoustical Society of America*, 73(6):2134–2144, 1983.

[48] 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. MIT press, Cambridge, 1995.

[49] J. A. Nelder and R. Mead. A simplex method for function optimization. *Computing Journal*, 7(4):308–313, 65.

[50] A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 84–91. IEEE, Los Alamitos, CA, USA, 1994.

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

[52] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C.* Cambridge University Press, second edition, 1992.

[53] L. R. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition.* Prentice-Hall, Englewood Cliffs, NJ, 1993.

[54] D. Reisberg, J. McLean, and A. Goldfield. Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In Dodd and Campbell [18], pages 97–113.

[55] P. L. Silsbee and A. C. Bovik. Computer lipreading for improved accuracy in automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5):337–351, 1996.

[56] P. L. Silsbee. *Computer Lipreading for Improved Accuracy in Automatic Speech Recognition.* PhD thesis, University of Texas, 1993.

[57] D. G. Stork, G. J. Wolff, and E. P. Levine. Neural network lipreading system for improved speech recognition. In *Proceedings International Joint Conference on Neural Networks*, volume 2, pages 289–295, 1992.

[58] W.H. Sumby and I. Pollak. Visual contributions to speech intelligibility in noise. *J. Acoustical Society of America*, 26:212–215, 1954.

[59] A. Q. Summerfield. Audio-visual speech perception, lipreading and artificial simulation. In Lutman M. E. and M. P. Haggard, editors, *Hearing Science and Hearing Disorders*, pages 131–182. Academic Press, New York, 1983.

[60] A. Q. Summerfield. Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London, Series B*, 335:71–78, 1992.

[61] A. J. Viterbi. Error bounds for convolutional codes and an asymtotically optimum decoding algorithm. *IEEE Trans. on Information Theory*, 13(2):260–269, 1967.

[62] M. Vogt. Fast matching of a dynamic lip model to color video sequences under regular illumination conditions. In David G. Stork and Marcus E. Hennecke, editors, *Speechreading by Humans and Machines*, volume 150 of *NATO ASI Series, Series F: Computer and Systems Sciences*, pages 399–408. Springer Verlag, Berlin, 1996.

[63] G. J. Wolff, K. V. Prasad, D. G. Stork, and M. E. Hennecke. Lipreading by neural networks: Visual preprocessing, learning and sensory integration. In J. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6. MIT Press, 1994.

[64] J. Wu, S. Tamura, H. Mitsumoto, H. Kawai, K. Kurosu, and K. Okazaki. Neural network vowel recognition jointly using voice features and mouth shape image. *Pattern Recognition*, 24(10):921–927, 1991.

[65] S. J. Young, P. C. Woodland, and P. C. Byrne. *HTK Version 1.5: User, Reference and Programmer Manual.* Entropic Research Laboratories, Washington, DC, 1993.

[66] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski. Integration of acoustic and visual speech signals using neural nets. *IEEE Commun. Mag.*, pages 65–71, November 1989.

[67] A. L. Yuille, P. Hallinan, and D. S. Cohen. Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, 8:99–112, 1992.