Extraction and Modeling of Visual Speech Features

Probabilistic Models for Speechreading and Face Recognition

Speechreading and Recognition of Talking Faces using Probabilistic Models

Speechreading using Flexible Models

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

[1]Institut Dalle Molle d'Intelligence Artificielle Perceptive
C.P. 592, Rue du Simplon 4, CH-1920 Martigny, Switzerland
Tel : +41 27 721 27, Fax : +41 27 721 77 12
email: luettin@idiap.ch

[2]Department of Electronic and Electrical Engineering
University of Sheffield
Mappin Street, Sheffield S1 3JD, UK

[3]Department of Medical Biophysics
University of Manchester
Oxford Road, Manchester M13 9PT, UK

Abstract. We describe a robust method for locating and tracking lips in gray-level image sequences. Our approach learns patterns of shape and appearance variability from a training set. This constrains the model 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 or visibility of teeth and tongue. Visual features are recovered from the tracking results and represent both, shape and intensity information of the mouth. Their utilization for a speechreading (lip-reading) system is described, where the features are modeled by Gaussian distributions and their temporal dependencies by Hidden Markov Models. While some of the visual information accounts for speech data, other information is intuitively related to the identity of the talker. We describe a novel approach for person recognition by building models of talking persons based on the extracted features. Experimental results are presented for locating lips, tracking lips, speechreading and person recognition. For a speaker independent digit recognition task, the speechreading system achieved an accuracy equivalent to that of untrained humans. The Person recognition system obtained an identification rate of 97.9 % on a database of 12 persons. The database used consists of a broad variety of speakers and was recorded in a natural environment, with no special lighting, lip markers or constraints used.

Some Questions:

- should I include the chapter ‘recognizing talking faces’?
- should I cut the simple equations how to calculate the normalization, mean and covariance matrix? (A reviewer claimed once that the model can not model asymmetric lips and that the corners of the lips are required to be at the same horizontal level, which is of course wrong)
- Are there too many references?
- Can I include the reference to the panel discussion [29]?
- can you think of any good titles?
- I included quite a lot of experiments and figures, so please give me comments on what I should leave out.

Comments:

- I will complete some of the missing Figures and update some of the old ones.

1. Introduction

Facial speech feature extraction and modeling has become an important issue in both, automatic speech processing and automatic face processing. Potential applications include audio-visual speech recognition [1], person recognition [2], lip synchronization [3], speech-driven talking heads [4] and speech based image coding [5].

Robust and accurate facial feature analysis is a difficult object recognition problem, because of the large variability between different subjects and the variability in the appearance of a specific subject due to changes in pose, lighting and mouth opening. It has received much attention in the face recognition community [6, 7, 8, 9, 10, 11] but is even harder for image sequences due to motion and changing appearance. 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.

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 [12]. Such environments are for example an office, car, factory or aircraft. 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 [13]. 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. Especially in the presence of noise, visual information is often complementary to acoustic information and can therefore improve speech perception [14, 15]. Reisberg et al. [16] has shown that normal-hearing subjects who see the talker's face perceive speech more accurately, even in noise-free environments. Hearing impaired and deaf persons make extensive use of visual cues and some few individuals perform this task to such a high degree which enables almost perfect speech perception [17].

Motivated by these psychological studies, several researchers [1, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28] 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 [29]. We apply our lip-tracker to this task. The parameters describing the talking mouth are recovered from the tracking result and describe the shape of the lip contours and the gray-level appearance of the mouth. We show how the extracted features can be modeled by mixtures of Gaussians and their temporal dependencies by Hidden Markov Models and be used for speechreading. How visual articulation can influence human perception of speech has been demonstrate by the McGurk effect [30] in which subjects mistakenly hear sounds which are biased by visual articulation.

Whereas the multi-modal nature of speech has received much attention from the speech recognition community, person identification research has mainly been treated uni-modal with face recognition [31, 32, 33] and speaker recognition [34] being two of these modalities. The appearance of a face can change considerably during speech and due to facial expressions. In particular the mouth is subject to considerable changes but at the same time it is also one of the most distinctive parts of a face. Face recognition research has largely ignored this variability and focused on static images with neutral facial expressions. Likewise, acoustic speaker recognition research has ignored the visual information inherent in speech production. Although several approaches for facial expression analysis have been proposed in the literature [35, 36, 37, 38, 39] their main objective was to analyze facial expressions rather than to model face variability for the purpose of face recognition. Person recognition combining face recognition and speaker recognition have recently been proposed in [40, 41], but face recognition was performed on the static face image with a neutral facial expression and did not make use of the visual motion of the face during speech production. We describe a novel approach for person recognition by building spatio-temporal models of talking persons based on features extracted by lip-tracking. The modeling of these features is similar to that of speechreading but here we aim to train models specific to a persons articulators and to the way that person speaks and therefore construct person dependent HMMs. Person identification is performed by tracking the lips of the unknown talking person, estimating the likelihood of each model of having generated the observed sequence of features and the model with the highest likelihood is chosen as the recognized person.

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 persent results for lip localisation and tracking. The next section describes the modeling of the extracted features for speechreading where we present results for a speaker independent digit recognition task. We also describe the modeling of talking faces and experiments for a identification task. The last section concludes the paper.

2. Background

The two main approaches for extracting visual speech information from image sequences can be grouped into the image based approach and the model based approach. In the image based approach the gray-level image containing the mouth is used directly as the feature vector. The advantage of this approach is that no data is thrown away. 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 the model based 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 model topology and the image feature analysis which maps the model to the image.

2.1 Image based approaches

Petajan [1] was probably the first researcher who developed a speechreading system. The system was based on geometric features of the mouth opening like height, with and area. A simple thresholding technique was used to find the mouth opening. A linear time-warping

algorithm was used to match test sequences to training templates. This system was later extended by Goldschen [24] to a continuous speechreading system using HMMs and context dependent sub-word models. The use of sub-word models is currently also the most promising method in acoustic speech recognition. Goldschen has determined distinct *viseme* classes, the visual equivalent to phonemes, and described modelling techniques commonly used in the acoustic domain which 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 which required manual assistance.

Finn et el. [18] described 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 [21].

Yuhas et al. [19] presented and approach where the whole gray-level image containing the mouth area was used as feature vector. Other approaches using the whole image as features have been reported in [22, 23, 26]. A method which only uses the horizontal and vertical gray-level vectors centered at the mouth has been described in [26].

Mase et al. [20] 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.

2.2 Model based approaches

Yuille et al. [42] have described the used of deformable templates for locating facial features. Their application to lip-tracking has been described in [43]. 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 the 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 [44].

Kass et al [45] 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 [46] 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. A gradient based image search is performed, similar to the one for deformable templates which does not take the actual appearance of typical gradients along lip contours into account. The energy function considers both, the distance of the contour to the sub-space and an internal energy which is minimal when the contour follows a straight line. The weights for each contribution were determined empirically.

An approach based on splines [47] has been described in [48]. 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. 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 [49]. 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. However, if the aim of such a system is to be used in such an environment, the robustness of the visual subsystem is also of great importance. We believe that a feature extraction method intended to be robust to all those variabilities encountered in real world application (illumination, speaker, 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.

One of the issue to address in a model based approach is to choose an appropriate description of the visible articulators. We are modelling 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 [20, 38]. 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 [50]. Furthermore, using optical flow computation to estimate such action might violate its underlying assumptions of rigid motion, i.e. when the mouth opens, and small movements, i.e. at a typical frame rate of 25-50 frames/sec mouth shape changes can occur very rapidly.

We therefore chose to build and appearance based model. 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 [51, 52, 17]. Particularly for fricatives, the place of articulation can often be determined visually [53], i.e. for labiodental (upper teeth on lower lip), interdental (tongue behind front teeth) and alveolar (tongue touching gum ridge) place. We 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 [54, 55, 56, 57]. PDMs are flexible models which represent an object by a set of labeled points. The points describe the boundary or other significant locations 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 deformations 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 $\mathbf{x}_i$ with

$$\mathbf{x}_i = (X_{i0} Y_{i0} X_{i1} Y_{i1} \ldots X_{in-1} Y_{in-1})^T \qquad (\ )$$

$(X_{ij} \ Y_{ij})$ j $(j$ .. $)$.

$$\mathbf{x}_i = M\left(\frac{1}{s}, -\theta\right)[V_i - V_c] \qquad (2)$$

with translation

$$\mathbf{V}_c = (t_x t_y t_x t_y \ldots t_x t_y)^T \qquad (3)$$

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 $\bar{\mathbf{x}}$ using

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \qquad (4)$$

The covariance matrix $\mathbf{S}$ is found using

$$\mathbf{S} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \qquad (5)$$

The eigenvectors and eigenvalues of the covariance matrix are obtained using principal component analysis. The eigenvectors with the largest eigenvalues describe the most significant modes of variation, in particular the variance spanned by an eigenvector is equal to its corresponding eigenvalue.

A normalized shape $\mathbf{x}$ can now be approximated by

$$\hat{\mathbf{x}} = \bar{\mathbf{x}} + \mathbf{Pb} \qquad (6)$$

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 three 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 subsequent modes account for asymmetric properties and 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 [58]. 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 leads to less robust image search. They also require more modes of variation than the true number of degrees of freedom of the object. We have experienced no drawbacks by using a linear model of shape for our experiments.

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 week at the lower lip. 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 shows some examples from the database we used [27] 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 use *a priori* knowledge of the actual gray-level appearance at the contour as well as its global variation across different examples. 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.

Following [55] 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. But instead of calculating individual mean profiles and covariance matrices for each model point, we concatenate the profiles of each model point to construct a global profile vector $\mathbf{h}_j$ for each training image as described in [59]:

$$\mathbf{h}_i = (\mathbf{g}_{i0}\mathbf{g}_{i1}\cdots\mathbf{g}_{in-1})^T . \tag{7}$$

We then calculate a global mean profile $\bar{\mathbf{h}}$ with

$$\bar{\mathbf{h}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{h}_i \tag{8}$$

and covariance matrix $\mathbf{S}_g$ using

$$\mathbf{S}_g = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{h}_i - \bar{\mathbf{h}}\right)\left(\mathbf{h}_i - \bar{\mathbf{h}}\right)^T . \tag{9}$$

The eigenvectors and eigenvalues of the covariance matrix $\mathbf{S}_g$ are found using principal component analysis. The eigenvectors with the largest corresponding eigenvalues describe the main modes of gray level variation seen in the training set. Any profile $\mathbf{h}$ can be approximated using

$$\hat{\mathbf{h}} = \bar{\mathbf{h}} + \mathbf{P}_g\mathbf{b}_g \tag{10}$$

where $\mathbf{P}_g = (\mathbf{p}_{g1}, \mathbf{p}_{g2}, \ldots, \mathbf{p}_{gt})$ is t (n*n_p)×t ma x of f t column g nv c o co pond ng o la g g nvalu and $\mathbf{b}_g$ a v c o con a n ng g fo ac g nv c o . T app oac m la o *local gray-level models* d c b d by Lan al. [39], ou app oac d ff n a u *global* g ay-l v l mod l. I al o m la o *eigenlips* po d by B gl al. [25]. B gl al. plac d a ndo a ound mou a a on c PCA a p fo m d. S nc ndo do no d fo m l p , g nv c o of PCA ma nly accoun fo n n y va a on du o d ff n l p ap and mou op n ng. In compa on, ou PCA pac d fo m l p con ou and fo d c b n n y va a on nd p nd n of l p ap . T n n y nfo ma on ob a n focomplmnayoconounfomaonpovddbyapmodl.

In o d o v ual z g ay-l v l mod l n pola g ay l v l b n p of l v c o and moo m a M d an f l . T m an mod l and f 3 mod of va a on a d p c d n F gu 5. T mod o l p a a and k n a ound l p cov d by p of l . T f mod accoun fo global llum na on, cond mod ma nly d c b n n y of lo l p , d mod c ang con a b n k n and l p and ubqunmodaccounfofnvaaonucavblyofandongu.

T app oac a um a va anc of p of l a d ff n mod l po n a co la d ac o a y a xp c d o b du o llum na on ff c and d ff n k n and l p n n y. T p of l mod l cap u global va a on b n d ff n p ak and va a on fo nd v dual p ak . Pa cula ly n n y va a on n d mou ubj c olagnnyvaaondungpcpoducon.

4. Database

We used the image part of the Tulips 1 database of isolated digits [27] 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. The use of this database makes our experiments very challenging because it consists of speakers with different ethnic origin, some with make-up or facial hair and different illumination.

5. Locating and Tracking Lips

5.1 Cost Function

As 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 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

$$\mathbf{b}_g = \mathbf{P}_g{}^T (\mathbf{h} - \bar{\mathbf{h}}). \tag{11}$$

To measure how well a model fits the image we measures the mean squared error (MSE)

E_p between the image profile and the aligned model profile using

$$E_p = \left(\mathbf{h} - \bar{\mathbf{h}}\right)^T \left(\mathbf{h} - \bar{\mathbf{h}}\right) - \mathbf{b}^T \mathbf{b}\,. \tag{12}$$

Unlike other methods [43, 46, 49], 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 +/- 3 s.d, which accounts for 99% of each shape mode, and we assign equal prior probabilities for all model shapes within these limits.

5.2 Image Search

It is assume 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 [60, 61] 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 first few shape parameters 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. (11) followed by the calculation of the cost using Eq. (12). 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.

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 profiles 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.

We think it is important to evaluate the performance of computer vision algorithms quantitatively to be able to compare them. 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*.

All tests for locating and tracking lips were performed on Set 2. For localization tests the position of the model was initialized at a position off the center as seen in Figure 6. 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. For tracking tests the model was initialized at the center of the image. Examples of locating lips in the first image frame are shown in Figure 14. Tracking examples are shown in Figure 9. Results for both models can be found in Table 1. Only two results for localization were classified as a *Miss*, both results are depicted in Figure 8.

For comparison we also performed 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 [43, 46]. 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 1: Results for lip location using profile models with off-center initialization.

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

Table 2 shows the results for locating the lips by using gradient information instead of the profile model. Miss-location was mainly caused by the small gradient of the outer lower lips,

Table 2: Results for lip location using gradient search.

	Initialization	Good (%)	Medium (%)	Miss (%)
Model 1	off center	6.3	16.7	77
Model 2	off center	4.2	10.4	87.5
Model 1	center	8.3	18.8	72.9
Model 2	center	14.6	14.6	70.8

reflections on the lips and gradients originating from the teeth. Initializing the model in the center of the image improved the results only slightly.

Table 3 shows the results for tracking the lips for both models. For Model 1, results for tracking the lips are similar to the ones for location. However, for Model 2, the performance for tracking the lips is lower than for locating them. Most of the *Adequate* and *Miss* results were caused by the model aligning to the teeth instead of the inner lip contour. The lips were however never missed completely.

Table 3: Results for lip tracking.

	Good (%)	Adequate (%)	Miss (%)
Model 1	95.8	2.1	2.1
Model 2	91.7	6.2	2.1

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: mouth closed ; mouth open and teeth not visible ; mouth open teeth visible. It is therefore unlikely that this variability can accurately be described by a uni-modal distribution.

6. Speechreading

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 the 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. The change in scale is likely to provide speech information and was also include it in the feature vector.

6.2 Speech Modeling

For modeling visual speech we use whole-word HMMs which is a standard approach in acoustic small vocabulary recognition systems [62].

A visual observation $\mathbf{O}$ of an utterance is represented by a sequence of feature vectors

$$\mathbf{O} = \mathbf{o}_1, \mathbf{o}_2, \ldots \mathbf{o}_T \tag{13}$$

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 Hidden Markov Model (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.

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

$$\lambda = (\mathbf{A}, \mathbf{B}, \pi). \tag{14}$$

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

$$b_i(\mathbf{o}) = \sum_{m=1}^{M} c_{im} N(\mathbf{o}, \mu_{im}, \Sigma_{im}) \tag{15}$$

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 [63]. The models are further re-estimated using the Baum-Welch procedure [64] (also known as the EM- method [65]), which maximises the likelihood of model λ for having generated the observed sequence **O**.

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)

$$\arg\max_i P(\lambda_i|\mathbf{O}) \tag{16}$$

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

$$P(\lambda_i|\mathbf{O}) = \frac{p(\mathbf{O}|\lambda_i)P(\lambda_i)}{p(\mathbf{O})}, \tag{17}$$

where $P(\lambda_i)$ represents the prior probability of class i and $P(\mathbf{O}|\lambda_i)$ the probability distribution of the feature sequence **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.

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 [66], 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 [67].

For experiments using *all* shape modes and *all* intensity modes in the features vectors as well 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 learns the features which are important for word discrimination, in practice the training set is rarely large and balanced enough to ensure this. We therefore performed recognition test by analysis the contribution of each shape and intensity mode individually. Results for Model 2 are depicted in Figure 11 and Figure 12. Results for Model 1 for shape features can be found in [68] and for intensity features in [69]. For our recognition task of 4 digits, the first and third shape mode and the scale and for the intensities, modes 12-14 make the largest contribution towards accuracy. The shape modes are shown in Figure 3 and the 12^{th}-14^{th} intensity modes are displayed in Figure 13. All three intensity modes mainly account for different intensity of the mouth opening.

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

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

Table 5: Word recognition rate using the six most discriminant features.

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

The word recognition rates using only the most 6 discriminant modes are shown in Table 5 (based on 2 shape features, scale and 3 intensity features, scale was only used as delta feature). Separate tests were performed by including delta parameters (Model 1 delta, Model 2 delta). Delta parameters also included delta scale. It is interesting to note, that Model 1 obtained higher rates for shape parameters than Model 2 when delta information was included. The recognition rates are considerably higher for Model 2 when using intensity information. This seems natural, since Model 2 provides gray-level information of the mouth opening. 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.

	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 about equivalent to the performance achieved by humans with no lip-reading knowledge who were asked to lip-read on the same database. Those subjects achieved and average of 89.93 % while the average performance of hearing impaired subjects with lip-reading knowledge was 95.49% [27].

7. Recognizing Talking Faces

Face recognition and facial expression recognition requires detailed analysis of the whole face but most facial motion during speech occurs around the mouth area. We are interested in facial changes due to speech and therefore use the same features we used for speechreading to build models of talking persons. We follow the approach where the face, or in this work part of the face, is represented in 2D, with no explicit 3D information. This approach is also motivated by psychophysical experiments, which suggest that humans may represent faces in a caricature like manner by considering shape information of individual parts and their spatial relationship [70]. We assume that the extracted lip parameters and their temporal dependencies during speech production are specific to a particular talking person and can therefore serve for a person identification system.

7.1 Modeling talking faces

The approach is similar to acoustic speaker recognition methods and based on the same modeling structure which we use for speechreading. A model of a talking person is shown in Figure 14. Speaker recognition tests can be classified into text dependent (TD) and text independent (TI) tasks. For text dependent tasks the test utterance is known while for text independent tasks it is not known. We describe our method for both TI and TD mode, where TI mode here is restricted by the size of the vocabulary. For the TD mode, we build one HMM per word class and speaker while for TI mode, only one HMM is built per speaker, representing all word classes. In TD mode, the spoken word is known and only HMMs of that word class are used for identification. In TI mode, the spoken word is not known, thus text independent HMMs representing all word classes are used for identification.

We trained HMMs which only allow 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 training set. Training is done the same way as for speechreading. Speaker identification 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) and the subject model with the highest likelihood is chosen to be the identified person.

7.2 Experiments

Set 1 was used for training and Set 2 for testing. We performed experiments for different feature vectors and different HMM architectures. Best results were obtained by HMMs with 4 states and 3 mixture components. The performance generally increased by using more mixture components, but the database only permitted to train about 3 or 4 components. The feature vector either contained shape parameters or intensity parameters or both. All experiments were

performed with Model 2. 10 parameters were used for describing the shape and 20 parameters for describing the profile.

Speaker recognition performance is often evaluated as a function of training duration and test duration and typically performance increases if either of those periods is increased. Typical periods used in speaker recognition are 10-30 seconds for training and 3-10 seconds for testing. In our experiments, average training duration was 0.3 seconds for TD tests and 1.3 seconds for TI tests. Average test duration was 0.3 seconds for both tasks. Both periods were therefore considerably shorter than typical periods used in the acoustic domain. Particularly for TD mode, the small size of the database, providing only one example of each word for a given speaker, caused difficulties in estimating the HMM parameters. We tried to alleviate these problems by applying different training and parameter tying methods [71].

7.3 Text Dependent Test (TD)

For TD recognition we built a separate HMM for each subject and each word class, resulting in a total of 48 models. Due to the small size of the database we used a sequential training procedure, where re-estimation is based on the models trained in the previous step:

- Estimating variances for one *global model* using all training data.
- Re-estimating means, mixture weights and transition probabilities for *subject independent word models*.
- Re-estimating the mean and mixture weights for *subject dependent word models*.

All HMMs therefore share the same variances and the transition probabilities of any word class are tied for all subjects. Only the means and mixture weights are estimated individually for each class and each subject.

Identification is based on speaker dependent models of the spoken word. The likelihood for each speaker is estimated and the speaker with the highest likelihood is chosen as the identified person. Results for TD tests are summarised in Only the variances are therefore tied across all models. Table 8 shows the results for text independent person identification. Best performance was obtained by using both, shape and intensity parameters. Although the performance for TI mode is generally worse than for TD mode in the acoustic domain, our system performed better for TI mode. But this is likely to be due to the very small training set for TD mode and the constrained training procedure needed to train the models. For both tasks, performance was higher for intensity parameters than for shape parameters. This could however be due to the first few intensity modes which probably account for different illumination and therefore bias recognition results.

Table 8. Best performance was achieved by using both, shape and intensity parameters.

7.4 Text Independent Test (TI)

For text independent person recognition we built one HMM for each subject, representing all utterances. The motivation behind this approach is to construct one model which represents different word classes by different mixture components. Parameter estimation was not as critical as in text dependent mode and was performed as follows:

- Estimating variances, means and mixture weights for one *global model*.

- Re-estimating mean, mixture weights and transition probabilities for a *text independent speaker model*.

Only the variances are therefore tied across all models. Table 8 shows the results for text independent person identification. Best performance was obtained by using both, shape and intensity parameters. Although the performance for TI mode is generally worse than for TD mode in the acoustic domain, our system performed better for TI mode. But this is likely to be due to the very small training set for TD mode and the constrained training procedure needed to train the models. For both tasks, performance was higher for intensity parameters than for shape parameters. This could however be due to the first few intensity modes which probably account for different illumination and therefore bias recognition results.

Table 8: Accuracy for text dependent (TD) and text independent (TI) person identification tests using shape and intensity parameters.

	Shape	Intensity	Shape + Intensity
TD	72.9 %	89.6 %	91.7 %
TI	83.3 %	95.8 %	97.9 %

To show the advantage of this modelling technique over static approaches we performed identification tests were only the first frame of each image sequence is used for training and for recognition. The Euclidean distance is performed on the feature vector containing all shape and intensity modes and the template with the smallest distance is selected as the recognised person. We can extend this model by including variance parameters and using the Mahalanobis distance. This model can be further extended by considering all images of an image sequence which are modelled by a mixture of Gaussian distribution, a so-called Gaussian Mixture Model (GMM) [72] which is equivalent to a HMM with only one state. The Results for these experiments can be found in Table 9.

Table 9: Person identification for Euclidean distance, Mahalanobis distance and Gaussian Mixture Model.

	Euclidean distance	Mahalanobis distance	GMM
TD	54.2 %	70.8 %	87.0 %
TI	66.7 %	77.1 %	93.8 %

Only the performance of the GMMs is very close to the performance using HMMs. GMM have the advantage that they can also be used for continuous text indépendant speaker recognition without the need to segment the sentences.

8. Conclusions

We have described a system for robust location 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 learned from training data. These two techniques avoid the use of heuristic assumptions about shape deformation and the representation of dominant image features and therefore enables robust image search. Our database did not include images with rotations in 3D, but is has been shown in [39] that a similar model of the face is able to recover vertical and horizontal rotation of at least +/- 20 degrees.

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.

A speaker independent digit recognition system has been presented, purely based on visual features, which obtained performance levels similar to humans. 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.

A new approach for person recognition was described based on spatio-temporal person models built with our extracted mouth features. While we have shown promising results for a speaker identification task of 12 persons, our primary goal was to build a link between the face recognition and speaker recognition communities and to set the path for a multi-modal person verification approach.

9. Acknowledgments

This work was partially funded by the University of Sheffield, the German Academic Exchange Service (DAAD) and the European ACTS-M2VTS project. The authors would like to thank Javier Movellan for the use of his database and the anonymous reviewers for their comments. The image processing algorithms were developed under the TINA vision research environment (http://www.shef.ac.uk/~eee/esg/research/tina.html).

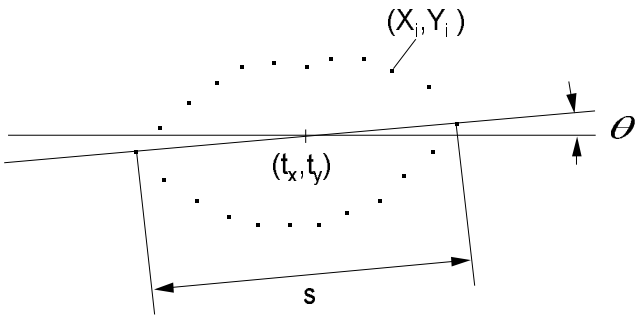

Figure 1: Model 1, representing the outer lip contour with 22 points.

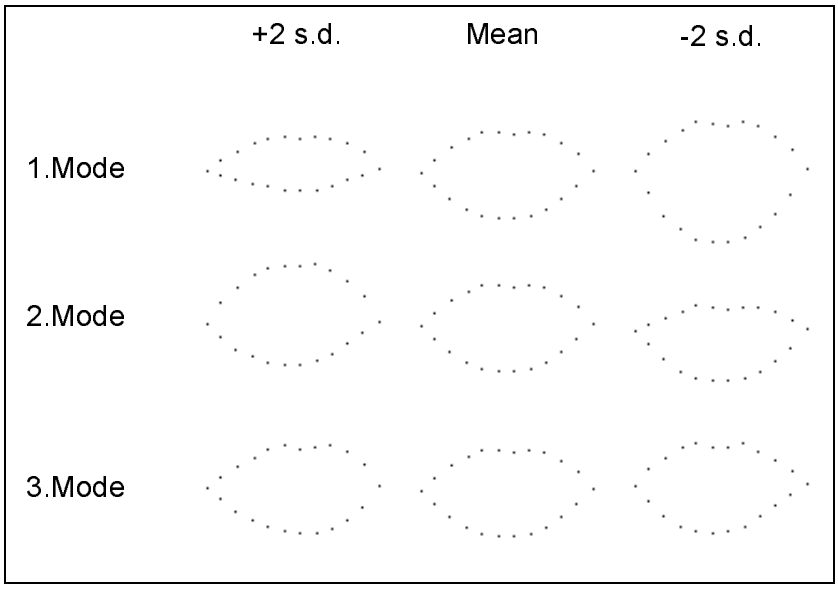

Figure 2: Mean shape and main modes of variation of Model 1.

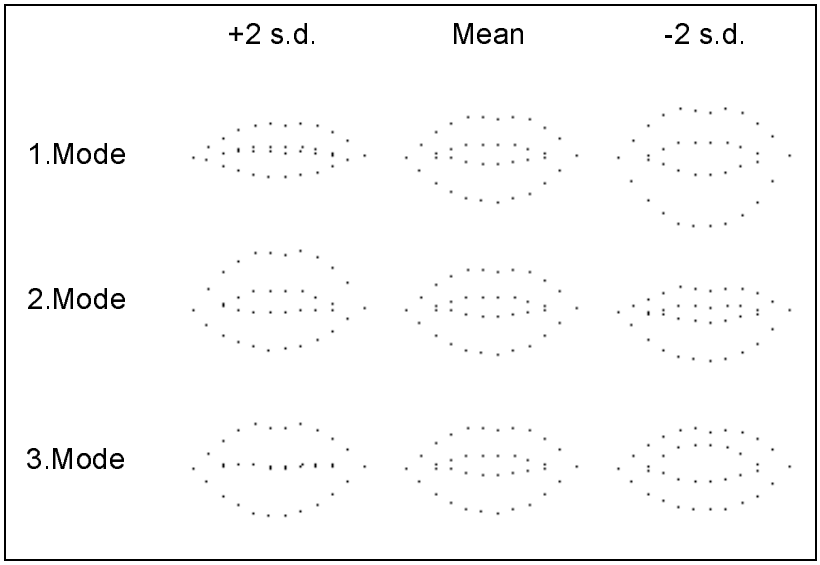

Figure 3: Mean shape and main modes of variation of Model 2.

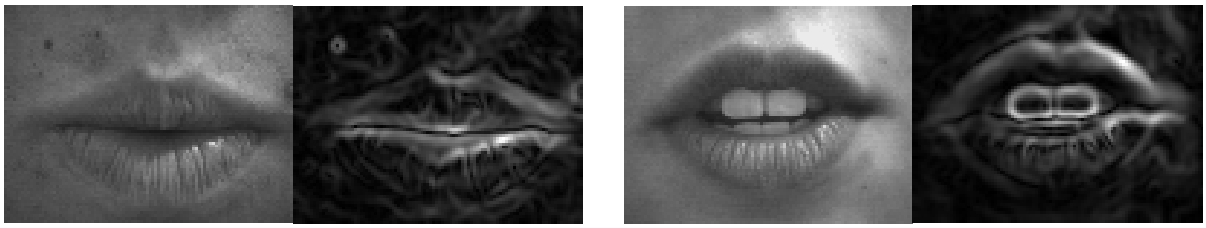

Figure 4: Example images and their gradient images.

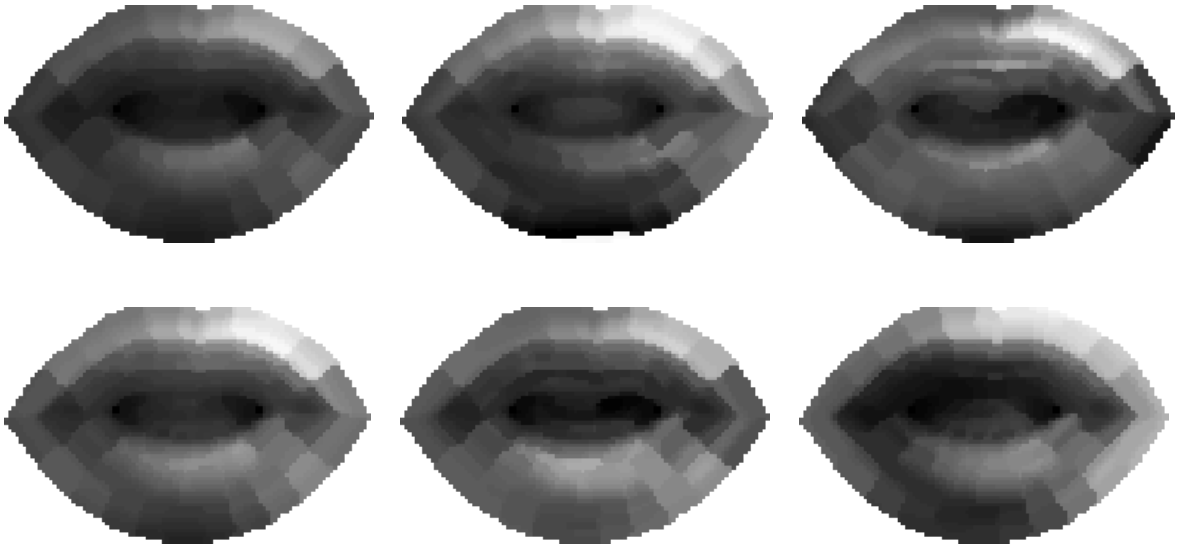

Figure 5: Main modes of grey-level variation for the mean shape of Model 2.

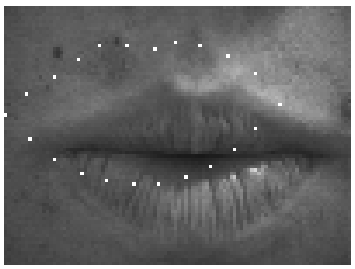

Figure 6: Off center initialization of the model.

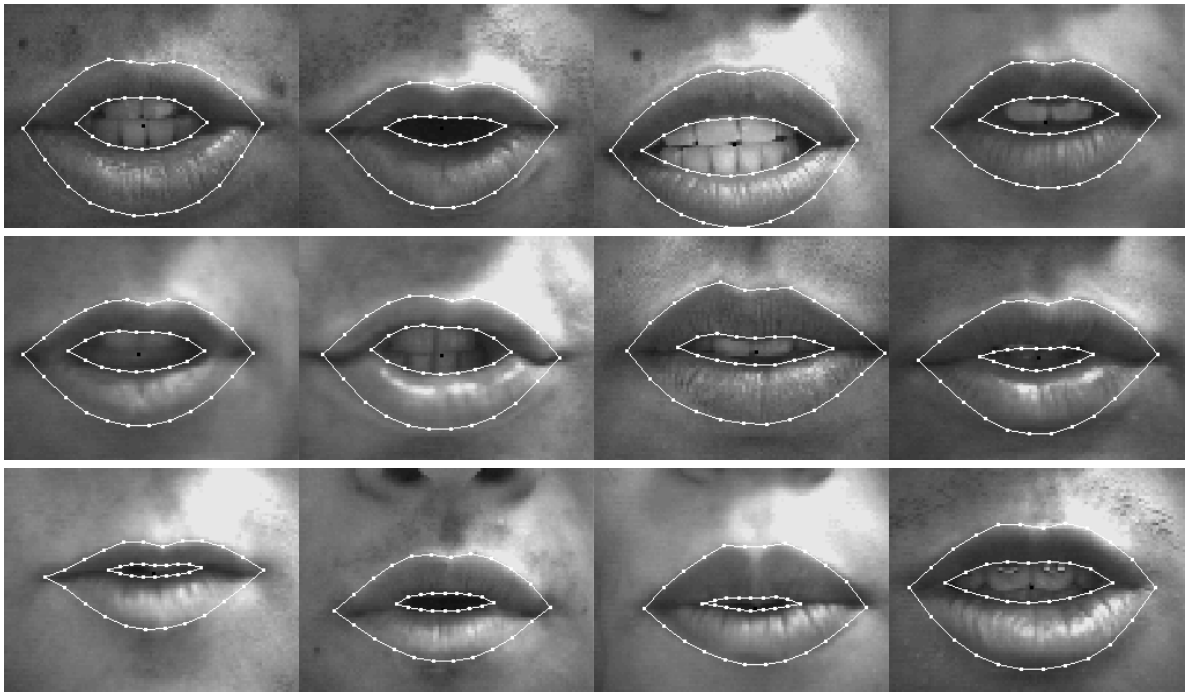

Figure 7: Examples of *Good* lip location results with Model 1 (top row) and Model 2 (middle row). Examples of *Miss* and *Adequate* results for Model 1 and Model 2 (bottom row from left to right).

Figure 8: The only two images where lip location was judged as *Miss*.

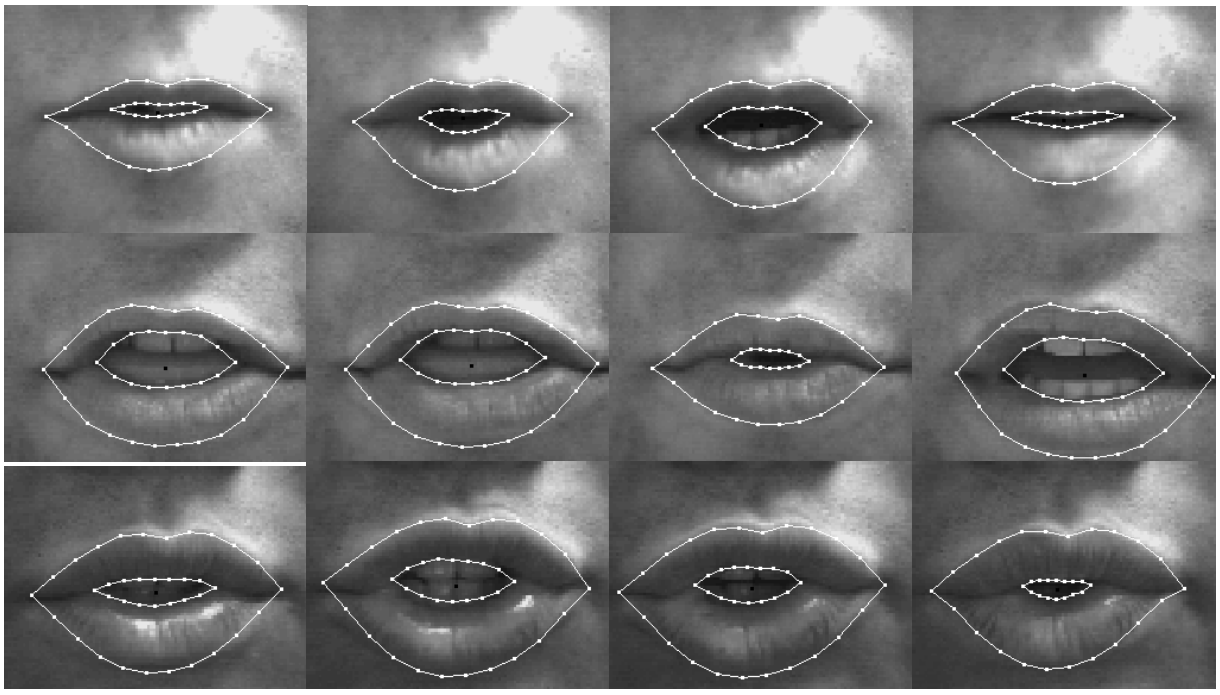

Figure 9: Tracking examples for Model 2.

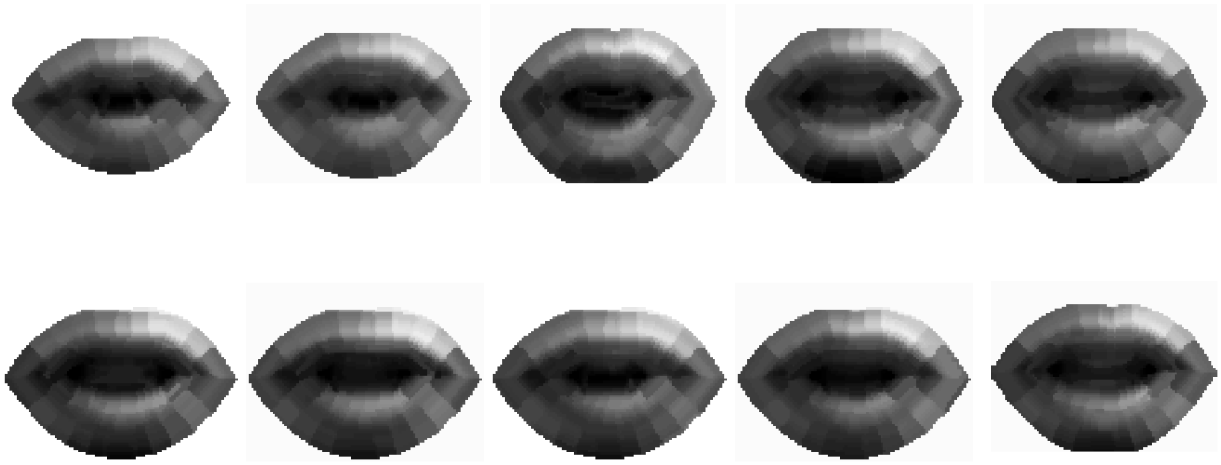

Figure 10: Learned HMM states for word 'one' (top) and 'two' bottom.

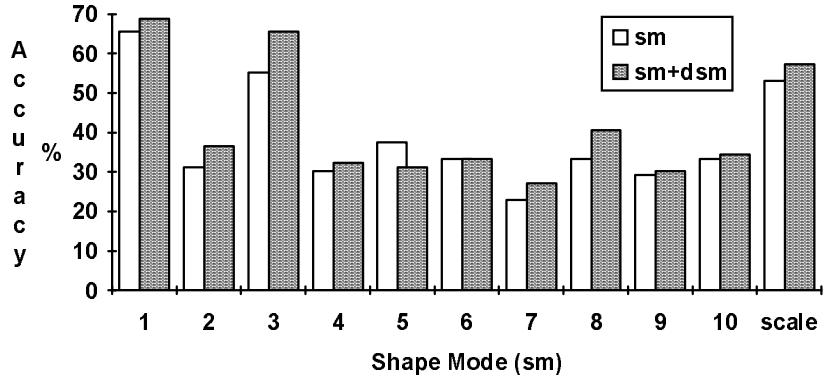

Figure 11: Recognition accuracy for each individual **shape** mode (sm) and delta shape mode (dsm) using Model 2.

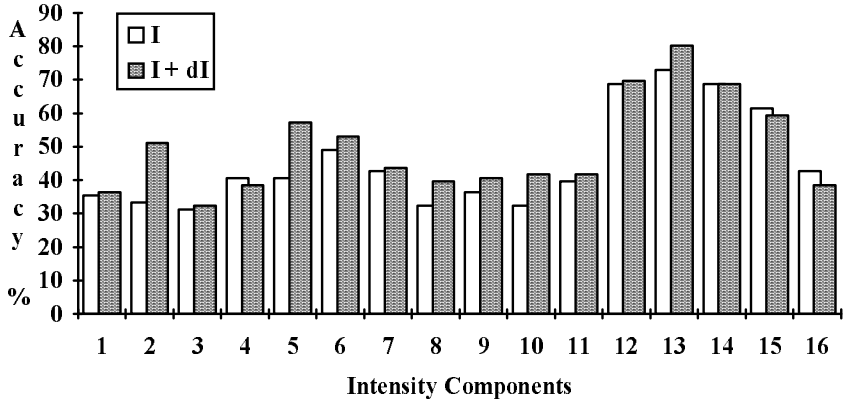

Figure 12: Recognition accuracy for each individual **intensity** mode (I) and delta intensity mode (dI) for the first 16 modes of Model 2.

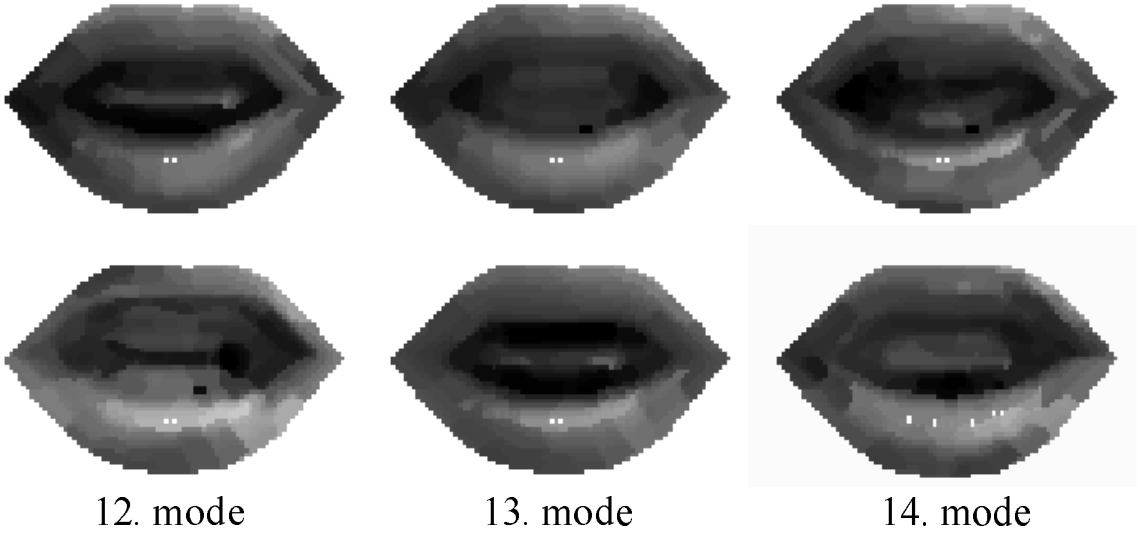

Figure 13: Intensity modes 12-14 with the highest individual recognition accuracy.

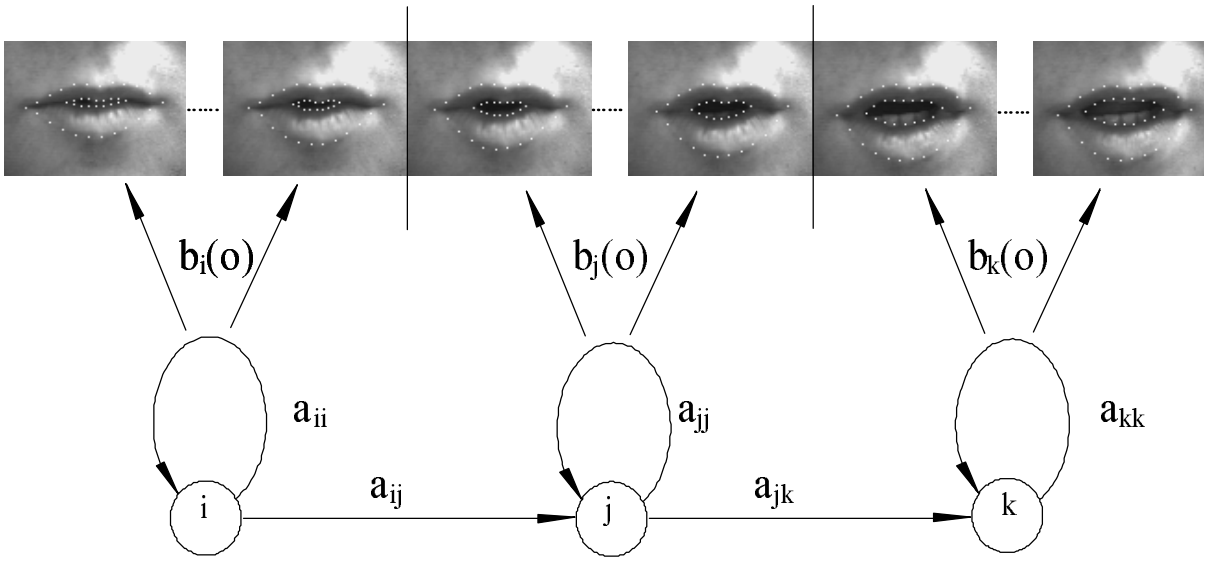

Figure 14: Spatio-temporal model of a talking face for a 3 state HMM with observation probabilities $\boldsymbol{b}(o)$ and transition probabilities $\boldsymbol{a}$.

1. E. Petajan, "Automatic Lipreading to Enhance Speech Recognition", Ph.D. thesis, University of Illinois at Urbana-Campaign, 1984.
2. J. Luettin, N. A. Thacker, S. W. Beet, "Learning to recognise talking faces", Proc International Conference on Pattern Recognition, Vienna, vol. 4, pp. 55-59.
3. T. Chen, H. P. Graf and K. Wang, "Lip synchronization using speech-assisted video processing", IEEE Signal Processing Letters, vol. 2, no. 4, pp. 57-59, 1995.
4. F. I. Parke, "A model of human faces that allows speech synchronized animation" Journal of Computers and Graphics, vol. 1, no. 1, pp. 1-4, 1975.

5. S. Morishima, K. Aizawa and H. Harashima, "An intelligent facial image coding driven by speech and phoneme", Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 1795-1798, 1989.
6. L. Huang and C.-W. Chen, "Human Facial Feature Extraction for Face Interpretation and Recognition", Pattern Recognition, vol. 25, no. 12, pp. 1435-1444, 1992.
7. A. L. Yuille, P. Hallinan, D. S. Cohen, "Feature extraction from faces using deformable templates", Int. J. Computer Vision, Vol. 8, pp. 99-112, August 1992.
8. G. Chow and X. Li, "Towards a system for automatic facial feature detection", Pattern Recognition, vol. 26, no. 12, pp. 1739-1755, 1993.
9. R. Brunelli and T. Poggio, "Face Recognition: Features versus Templates", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 15, no. 10, pp. 1042-1052, 1993.
10. A. Pentland, B. Moghaddam, T. Starner, "View-based and Modular Eigenspaces for Face Recognition", Proc. IEEE Conf. Computer Vision and Pattern Recognition, 1994.
11. N. Intrator, D. Reisfeld, Y. Yeshurun, "Extraction of Facial Features for Recognition using Neural Networks", Proc. IWAFG95, International Workshop on Automatic Face- and Gesture-Recognition, 1995.
12. Y. Gong, "Speech recognition in noisy environments : A survey", Speech Communication, vol. 16, pp. 261-291, 1995.
13. R. Cole, L. Hirschmann, L. Atlas et al., "The challenge of spoken language systems : research directions for the nineties", IEEE Trans. Speech and Audio Processing, vol. 3, no. 1, pp. 1-20, 1995.
14. W. H. Sumby and I. Pollack, "Visual Contribution to Speech Intelligibility in Noise", Journal of the Acoustical. Society of America., Vol.26, pp. 212-215, 1954.
15. K. W. Grant and L. D. Braida, "Evaluating the articulation index for auditory-visual input", Journal of the Acoustical Society of America, vol. 89, no. 6, pp. 2952-2960, 1991.
16. D. Reisberg, J. McLean and A. Goldfield, "Easy to hear but hard to understand: a lipreading advantage with intact auditory stimuli", B. Dodd and R. Campbell (eds.) Hearing by eye: the psychology of lip-reading, pp. 97-113. London: Lawrence Erlbaum Associates, 1987.
17. Q. Summerfield, "Lipreading and audio-visual speech perception", Phil. Trans. R. Soc. Lond. B 335, pp. 71-78, 1992.
18. E. K. Finn and A. A. Montgomery, "Automatic optically based recognition of speech", Pattern Recognition Letters, vol. 8, no. 3, pp. 159-164, 1988.
19. B. P. Yuhas, M. H. Goldstein, T.J. Sejnowski, "Integration of Acoustic and Visual Speech Signals using Neural Networks", IEEE Communications Magazine, pp. 75-81, 1989.
20. K. Mase and A. Pentland, "Automatic Lipreading by Optical-Flow Analysis", Systems and Computers in Japan, vol. 22, no. 6, pp. 67-75, 1991.
21. D. G. Stork, G. Wolff, E. Levine, "Neural Network Lipreading System for Improved Speech Recognition", Proc. International Joint Conference on Neural Networks, Vol. II, pp. 154-159, 1992.
22. P. L. Silsbee, "Computer Lipreading for Improved Accuracy in Automatic Speech Recognition", Ph.D. Dissertation, University of Texas at Austin, 1993.

23. C. Bregler, H. Hild, S. Manke and A. Waibel, "Improving connected letter recognition by lipreading", Proc. Int. Conf. Acoustics, Speech, and Signal Processing, vol. 1, pp. 557-560, Minneapolis, 1993.

24. A. J. Goldschen, "Continuous Automatic Speech Recognition by Lipreading", Ph.D. Dissertation, George Washington University, Washington DC, 1993.

25. C. Bregler and Y. Konig, "'Eigenlips' for Robust Speech Recognition", Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 669-672, 1994.

26. 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. D. Cowan, G. Tesauro and J. Alspector (editors), Advances in Neural Information Processing Systems, vol. 6, pp. 1027-1034. Morgan Kaufmann, 1994.

27. J. R. Movellan, "Visual Speech Recognition with Stochastic Networks", G. Tesauro, D. Touretzky, T. Leen (eds.) Advances in Neural Information Processing Systems. Volume 7, MIT Press Cambridge, 1995.

28. J. Luettin, N. A. Thacker, S. W. Beet, "Visual Speech Recognition using Active Shape Models and Hidden Markov Models", Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Atlanta, vol. 2, pp. 817-820, 1996.

29. J. Luettin, M. Vogt and C. Bregler, "Panel Discussion : Machine Recognition and Applications" in D. G. Stork and M. E. Hennecke (editors), Speechreading by Humans and Machine, NATO ASI Series, Series F : Computer and Systems Sciences, pp. 549-555, Berlin, 1996, Springer Verlag.

30. H. McGurk and J. W. MacDonald, "Hearing lips and seeing voices", Nature, vol. 264, pp. 746-748, 1976.

31. D. Valentin, H. Abdi, A. O'Toole and G. W. Cottrell, "Connectionist Models of Face Processing : A Survey", Pattern Recognition, vol. 27, no. 9, pp. 1209-1230, 1994.

32. A. Samal and P. Iyengar, "Automatic Recognition and Analysis of Human Faces and Facial Expressions : A survey", Pattern Recognition, vol. 25, no. 1, pp. 65-77, 1992.

33. R. Chellappa, C. L. Wilson and S Sirohey, "Human and Machine Recognition of Faces: A Survey", Proc IEEE, vol. 83, no. 5, pp. 705-740, 1995.

34. S. Furui, "An overview of Speaker recognition technology", Proc. ESCA Workshop on automatic speaker recognition, identification and verification, Martigny, pp. 1-9, 1994.

35. K. Mase, "Recognition of facial expression from optical flow", IEICE Transactions, vol. E 74, no. 10, pp. 3474-3483, 1991.

36. D. Terzopoulos and K. Waters, "Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 15, no. 6, 1993.

37. Y. Yacoob and L. S. Davis, "Computing Spatio-Temporal Representations of Human Faces", Pro. IEEE Computer Vision and Pattern Recognition, 70-75, 1994.

38. I. A. Essa and A. P. Pentland, "Facial Expression Recognition using a Dynamic Model and Motion Energy", Proc. Int. Conf. on Computer Vision, pp. 20-23, 1995.

39. A. Lanitis, C. J. Taylor and T. F. Cootes, "A Unified Approach To Coding and Interpreting Face Images", Proc. Int. Conf. Computer Vision, 1995.

40. R. Brunelli, D. Falavigna, L. Stringa and T. Poggio, “Automatic Person Recognition by Acoustic and Geometric Features”, Machine Vision Applications, vol. 8, no. 5, pp. 317-325, 1995.
41. R. Brunelli and D. Falavigna, “Person Identification Using Multiple Cues” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 17, no. 10, pp. 955-966, 1995.
42. A. L. Yuille, P. Hallinan, D. S. Cohen, “Feature extraction from faces using deformable templates”, Int. J. Computer Vision, Vol. 8, pp. 99-112, August 1992.
43. M. E. Hennecke, K. V. Prasad, D. G. Stork, “Using Deformable Templates to Infer Visual Speech Dynamics”, 28th Annual Asilomar Conference on Signals, Systems and Computers, 1994.
44. T. Coianiz, L. Torresani and B. Caprile, “2D deformable models for visual speech analysis“, in D. G. Stork and M. E. Hennecke (editors), Speechreading by Humans and Machine, NATO ASI Series, Series F : Computer and Systems Sciences, pp. 391-398, Berlin, 1996, Springer Verlag.
45. M. Kass, A. Witkin and D. Terzopoulos, “Snakes: active contour models”, Int. J. Computer Vision, pp. 321-331, 1988.
46. C. Bregler and S. Omohundro, “Surface Learning with Applications to Lip-Reading”, J. D. Cowan, G. Tesauro and J. Alspector (eds.) Advances in Neural Information Processing Systems 6. San Francisco, CA: Morgan Kaufmann Publishers, 1994.
47. R. Bartels, J. Beatty and B. Barsky, ”An Introduction to Splines for Use in Computer Graphics and Geometry Modelling”, Morgan Kaufmann, 1987.
48. R. Kaucic, B. Dalton and A. Blake, “Real-time lip tracking for audio-visual speech recognition applications“ Proc. 4th European Conference on Computer Vision, vol. 2, pp. 376-387, 1996.
49. M. Vogt, “Fast Matching of a Dynamic Lip Model to Color Video Sequences under Regular Illumination Conditions“, in D. G. Stork and M. E. Hennecke (editors), Speechreading by Humans and Machine, NATO ASI Series, Series F : Computer and Systems Sciences, pp. 399-408, Berlin, 1996, Springer Verlag.
50. W. J. Hardcastle, “Physiology of Speech Production“, Academic Press, 1976, New York.
51. A. A. Montgomery, P. L. Jackson, “Physical characteristics of the lips underlying vowel lipreading performance”, J. Acoust. Soc. Am., Vol. 73, pp. 2134-2144, 1983.
52. M. McGrath, A. Q. Summerfield and N. M. Brook, ”Roles of lips and teeth in lipreading vowels”, Proc. Inst. Acoustics, pp. 401-408, 1984.
53. Q. Summerfield, “Audio-visual Speech Perception, Lipreading and Artificial Stimulation“, in M. E. Lutman and M. P. Haggard (editors), Hearing Science and Hearing Disorders, pp. 131-182, 1983, London, Academic Press.
54. T. F. Cootes and C. J. Taylor, “Active Shape Models - Smart Snakes”, Proc. Br. Machine Vision Conf., Springer Verlag, Berlin, pp. 266-275, 1992.
55. T. F. Cootes and C. J. Taylor, A. Lanitis, D. H. Cooper and J. Graham, ”Building and using flexible models incorporating grey-level information”, Proc. International Conference on Computer Vision, pp. 242-246, IEEE Comp. Soc. Press, 1993.
56. 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, 1995.

57. J. Luettin, N. A. Thacker, S. W. Beet, "Locating and tracking facial speech features", Proc. International Conference on Pattern Recognition, Vienna, vol. 1, pp. 652-656, 1996.
58. J. Luettin, N. A. Thacker, S. W. Beet, "Active Shape Models for Visual Speech Feature Extraction" in D. G. Stork and M. E. Hennecke (editors), Speechreading by Humans and Machine, NATO ASI Series, Series F : Computer and Systems Sciences, pp. 383-390, Berlin, 1996, Springer Verlag.
59. J. Haslam, C. J. Taylor and T. F. Cootes, "A Probabilistic Fitness Measure for Deformable Template Models", Proc. British Machine Vision Conference, pp. 33-42, 1994.
60. J. A. Nelder, R. Mead, "A simplex method for function minimization", Comput. J. Vol. 7(4), pp. 308-313, 1965.
61. W. H. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, "Numerical Recipes in C", Cambridge University Press, Cambridge, 1992.
62. L. R. Rabiner and B. H. Juang, " Fundamentals of Speech Recognition", Prentice Hall, New Jersey, 1993.
63. A. J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm", IEEE Trans. Information Theory, vol. 13, no. 2, pp. 260-269, 1967.
64. L. E. Baum, T. Petrie, G. Soules and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains", Annals Math. Stat., vol. 41, pp. 164-171, 1970.
65. A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", J. Royal Stat. Soc. Series B, vol. 39, pp. 1-38, 1977.
66. K. Fukunaga, "Introduction to statistical pattern recognition" Academic Press Inc., 1990.
67. S. J. Young, P. C. Woodland, P. C. Byrne, "HTK Version 1.5: User, Reference and Programmer Manual", Publ. Entropic Research Laboratories, Washington DC, 1993.
68. J. Luettin, N. A. Thacker, S. W. Beet, "Statistical lip modelling for visual speech recognition", Proc. 8th European Signal Processing Conference, Trieste, vol. 1, pp. 137-140, 1996.
69. J. Luettin, N. A. Thacker, S. W. Beet, "Speechreading using shape and intensity information", Proc. International Conference on Spoken Language Processing, Philadelphia, vol. 1, pp. 58-61, 1996.
70. G. Rhodes, S. E. Brennan and S. Carey, "Identification and rating of caricatures: implications for mental representations of faces", Cognitive Psychology, vol. 19, pp. 473-497, 1987.
71. L. R. Bahl, F. Jelinek and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition", IEEE Trans. Pattern Anal. Machine Intell., Vol. 5, pp. 179-190, 1983.
72. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", IEEE Trans. Speech and Audio Processing, vol. 3, pp. 72-83, 1995.