UNIVERSIDAD NACIONAL DE COLOMBIA

Lip region feature extraction analysis by means of stochastic variability modeling

David Augusto Cárdenas Peña

Universidad Nacional de Colombia
Faculty of Engineering and Architecture
Departement of Electrics, Electronics and Computation Engineering
Manizales, Colombia
2011

Lip region feature extraction analysis by means of stochastic variability modeling

David Augusto Cárdenas Peña

Thesis submitted as a partial requirement to receive the grade of:
Magister en Ingeniería - Automatización Industrial

Director:
Ph.D. Germán Castellanos Domínguez

Academic Research Group:
Signal Processing and Recognition

Universidad Nacional de Colombia
Faculty of Engineering and Architecture
Departement of Electrics, Electronics and Computation Engineering
Manizales, Colombia
2011

Caracterización del contorno labial en video empleando análisis de variabilidad estocástica

David Augusto Cárdenas Peña

Thesis submitted as a partial requirement to receive the grade of:
Magister en Ingeniería - Automatización Industrial

Director:
Ph.D. Germán Castellanos Domínguez

Academic Research Group:
Signal Processing and Recognition

Universidad Nacional de Colombia
Faculty of Engineering and Architecture
Departement of Electrics, Electronics and Computation Engineering
Manizales, Colombia
2011

(Dedicatoria)

A los dos mejores ejemplos, Islena y Didier, por enseñarme a ser persona.
A mi hermosita por su apoyo incondicional y el aguante en los momentos difíciles.
A los Serna Morales por hacerme uno más de la familia.

La mención recibida se la quiero dedicar a mis compis del grupo, con bonus al Oso, Leo, el Negro y el Paya. Se necesita de un grupo como ustedes para hacer trabajos de gran calidad y evitar enloquecerse.

Más pasión, menos técnica y 1e-10 táctica.

Acknowledgement

The author wants to thank research project number 1127-405-20332 *Identificación de posturas labiales en pacientes con labio y/o paladar hendido corregido*, of the Universidad de Caldas, for providing scholarship funding. Additional thanks are given to the COOPEN project (Erasmus Mundus External Cooperation Windows Programme), for allowing an academic exchange at the Universidad Politécnica de Valencia (Spain).

Abstract

On this thesis work, an analysis of lip region characterization techniques used to model lip dynamics was performed. To carry out the analysis a video sequence database of Spanish alphabet was built and used to train a visual speech recognition system with several feature extraction methodologies. The aim of the experiment is to evaluate the ability of each feature set to model accurately lip movement. Appearance-based, shape-based and spatiotemporal-based feature extraction methodologies were tested. Reported results let choose the spatiotemporal features as the best descriptors for visual speech dynamics.

Keywords: Visual Speech Recognition, Visual Feature Extraction, Lip Movement Modeling, Image Processing, Stochastic Modeling, Pattern Recognition

Resumen

En este trabajo de tesis se analizaron diferentes técnicas de carateterización de la región labial, usadas para modelar la dinámica labial. Para llevar a cabo este análisis, se construyó una base de datos de secuencias de video de la pronunciación del alfabeto español. Esta base de datos se utilizó para entrenar un sistema de reconocimiento visual del habla usando diferentes metodologías de extracción de características. El objetivo del experimento es evaluar la habilidad de cada conjunto de características para modelar adecuadamente el movimiento labial. Se probaron metodologías basadas en apariencia, forma y una representación espacio-temporal. Los resultados reportados permiten seleccionar las características espacio-temporales como los mejores descriptores, dentro de los evaluados, de la dinámica visual del habla.

Palabras clave: Reconocimiento visual del habla, Extracción visual de características, Modelado del movimiento labial, Procesamiento de imágenes, Modelado estocástico, Reconocimiento de patrones.

Contents

List of Figures

List of Tables

1 Preliminaries

1.1 Introduction

Automatic speech recognition (ASR) is currently an important research field, since wide range of applications can be deployed by means of it. Mainly, human-computer interfaces are being developed with ASR systems, that is because speech is the natural human communication channel. Although current speech based tasks, such as dictation and automatic translation, have been improved in recent years, their performance is strongly determined by variables as the speaker dependence and the environment noise. Some state-of-the-art ASR systems present high performance on "clean" environments, but under hard environment conditions their performance is reduced drastically[1]. To overcome those issues some approaches propose the use of another channel of information to complement the audio signal and make ASR systems robust enough to be deployable in field applications. Clearly, visual speech is the first candidate to be a noise-robust source of information. The benefit of the use of visual information on speech recognition tasks has been demonstrated [2]. Furthermore, bimodal integration of audio and visual stimuli in perceiving speech has been demonstrated by the McGurk effect [3]: when, for example, the spoken sound */ga/* is superimposed on the video of a person uttering */ba/*, most people perceive the speaker as uttering the sound */da/*. In addition, visual speech is of particular importance to the hearing impaired: mouth movement is known to play an important role in both sign language and simultaneous communication between the deaf.

There are three key reasons why vision benefits human speech perception [4]: it helps the speaker localization (audio source), it contains speech segmental information that supplements the audio, and it provides complimentary information about the place of articulation. The latter is due to the partial visibility of articulators, such as the tongue, teeth, and lips, which can help disambiguate some phonemes sets. For example, the unvoiced consonants */p/* (a bilabial) and */k/* (a velar), the voiced consonant pair */b/* and */d/* (a bilabial and alveolar, respectively), and the nasal */m/* (a bilabial) from the nasal alveolar */n/*, all three pairs are highly confusable based on acoustics alone. These facts have motivated significant interest in automatic recognition of visual speech, formally known as automatic lipreading or speechreading. Works in this field aims at improving ASR by exploiting the visual modality of the speaker's mouth region in addition to the traditional audio modality, leading to Visual Speech Recognition (VSR) and Audiovisual Speech Recognition (AVSR) systems.

Compared to audio-only speech recognition, AVSR introduces new tasks on video signals, which are shown in Figure 1-1. First, a face detection, as well as a mouth or lips detection are required to locate the informative area. Then, visual features must be extracted from the Region of Interest in video recording. Finally, a combination strategy of audio and visual features is needed aiming to improve the performance of the two single modality recognizers. The last two issues, namely, visual feature extraction and audiovisual fusion, are important research fields in the scientific community.

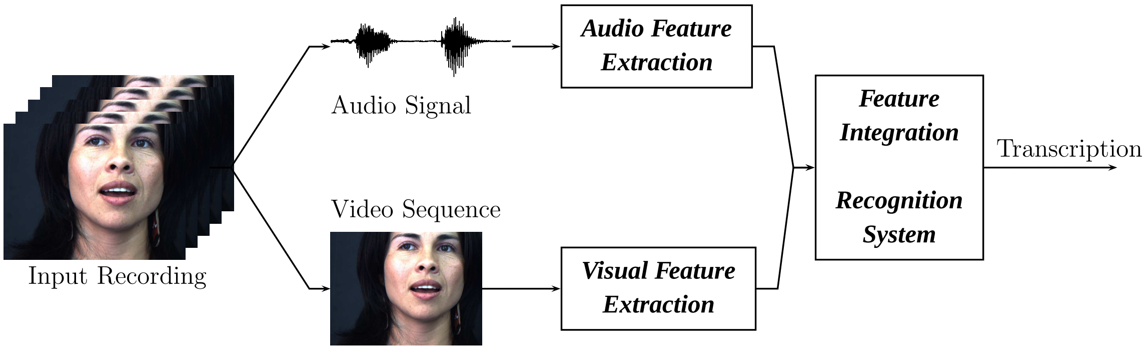


Figure 1-1: General AVSR system

The first automatic speechreading system was reported by Petajan [5] on 1984. Since then, a lot of works have been introduced. Most of them deal with recognition tasks of speaker (authentication) [6], isolated words [5–8], isolated digits [9, 10], letters [9, 11] and closed-set sentences [12, 13], mostly in English. The wide range of applications and strategies to perform AVSR tasks make those works hard to compare.

The main goal of this thesis work is to compare and select mouth region feature extraction techniques by means of their performance on a VSR task. First, a mouth region segmentation and a lip contour extraction algorithms are presented as the initial stage of the VSR system. Then, the set of visual feature extraction techniques used are presented. Finally, each technique is used to develop a recognition system for isolated Spanish alphabet digits by using a classical HMM-based classifier.

1.2 State of the art

The interest in facial image processing is leaded by the wide range of applications designed to decode, read and recognize these images. Some applications are:

- Automatic face recognition systems.
- Gesture and face pose detection systems.
- Automatic speech recognition systems.

Automatic face recognition is, probably, the first field which takes advantage of the information on photographic images and video sequences to recognize facial structures. Specialized classification algorithms, such as EigenFaces [14], allow an adequate recognition on a small set of subjects on isolated images. In Takimoto et al. [15] and Bao et al. [16] several approaches to face contour detection are shown. They aim to locate the face and facial structures on photographic images of groups of people. These algorithms have a simple design and are straightforward, although they loose accuracy on the facial structure location and their segmentation.

On the other hand, there are applications which aim to detect face gestures. The main objective on this approaches is to achieve a high quality segmentation and structure parametrization, no matter the computational cost required to perform it. This applications are common on virtual character generation, in which the most realistic representation of the face movement is required. [17].

Besides, visual information transmission and Visual Speech Recognition (VSR) applications are common [18]. In some cases, detected gestures are used to drive robotic systems, and a low recognition time is reached, which enables a real time operation of the application.

Several visual features have been proposed on literature. In general, they can be grouped in two sets, *low level* and *high level* features. On the latter, lip contour is extracted from image sequences and represented by a parametric [19] or statistical [20] model, and the parameters of the model are used as features; alternatively, geometric features are used as a complement [21, 22]. On the former approach, full lip region image is considered as informative and some appropriate transformations can be used as visual features [9, 19, 21–25]. From feature extraction methodologies used to build databases for training and develop of automatic recognition systems, most of the signals exhibited a stochastic behavior, which has important discriminative information. The usual analysis of the feature vectors does not show this behavior, on the contrary, it hides the information required to identify functional states. Temporal information of lip movement has been used to develop more robust speech recognition systems. In this field, the dynamic has been modeled by the inclusion of the first and second time derivatives of the features. This approach improves the overall performance of the system, even in noisy environments [26].

1.3 Objectives

1.3.1 General objective

To analyze and develop image feature extraction methodologies to model the lip region on video sequences, with a representation quality suitable to perform a lip dynamic analysis by means of stochastic variability modeling.

1.3.2 Specific objectives

- To build a set of visual features, by means of artificial vision techniques, which allow a suitable mouth modeling.

- To build a visual speech recognition system based on the stochastic modeling of the lip dynamics.

- To test and compare the performance of the methodology against the most common used feature extraction methodologies.

2 Materials and Methods

2.1 Lip Segmentation

Lip segmentation is one of the most important issues on Visual and Audio/Visual Speech Recognition (AVSR), since it is the starting point of the recognition task. An accurate lip detection improves the quality of visual information extracted from the speech video sequences. Nonetheless, achieving a correct lip extraction has proved to be difficult due to the weak color contrast and the significant overlap in color features between the lip and the face regions. Moreover, features as lip shape variation in different people, skin color in different human races, presence of facial hair and uncontrolled illumination conditions have negative impact on the performance of lip segmentation algorithms.

Many techniques have been proposed for this task. Active contour algorithm was applied by Delmas et al. [27] to extract the lip contour. The main drawback on this approach is the parameter tunning and the algorithm initialization. Active shape methods [28] have also been used, but they often converge to inaccurate results, when the lip edges are unclear or lip color is close to face color. Also, these methods need a large training set to extract lip shapes. Fuzzy c-mean clustering [29–31] techniques and parametric modeling [32, 33] have been used to segment pixels from various facial structures. Although, those models are highly dependent of the color contrast among facial structures.

In this chapter, a lip segmentation technique is introduced; the result of this stage will be called the Region-of-Interest (RoI). Then, the *jumping snake* technique proposed by Eveno et al. [34], which is employed to extract the lip contour, is explained. The RoI and the extracted contour will be used in the visual feature extraction stage.

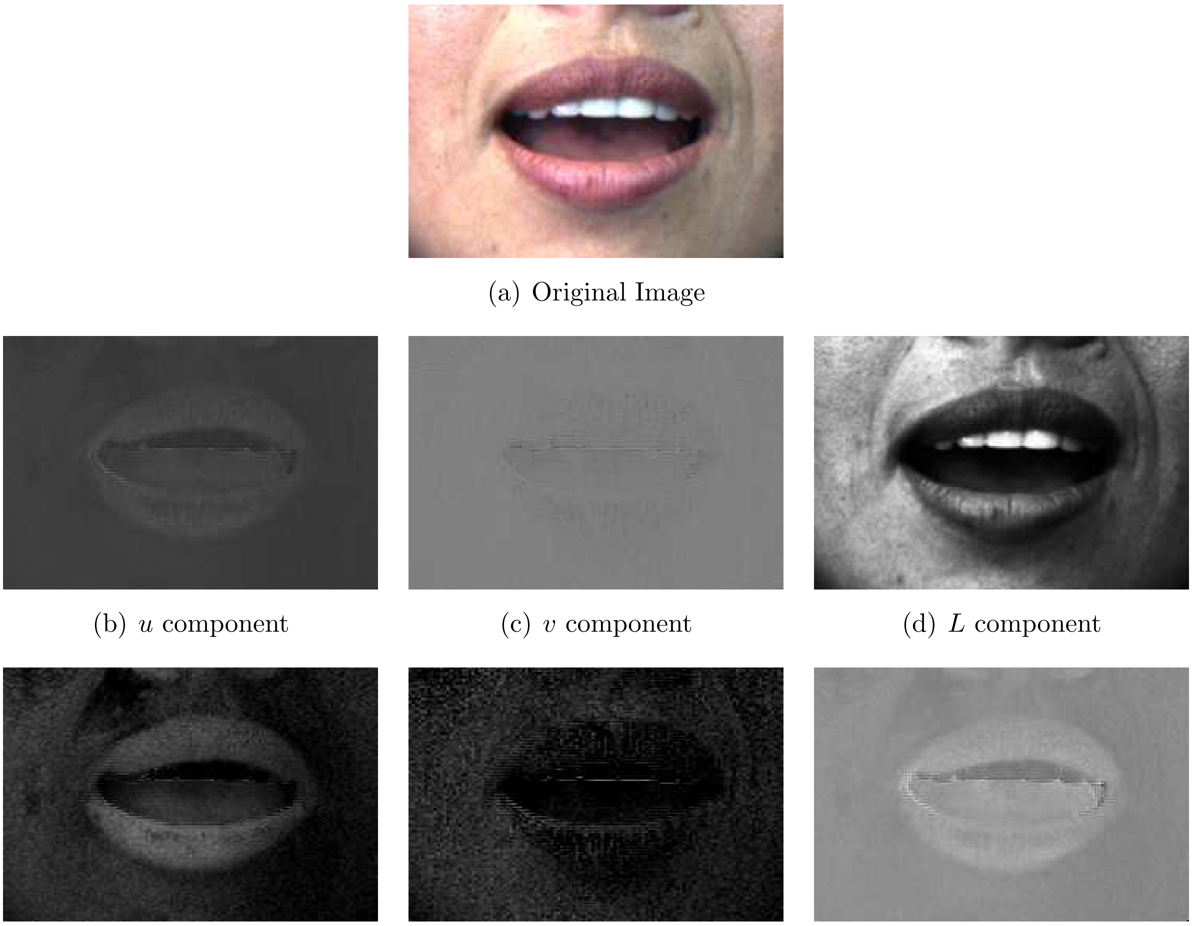
2.1.1 Pixel-based Mouth Structures Segmentation Techniques

Pixel-based techniques to mouth structure segmentation can be seen as a classifier, which uses features extracted from each pixel in an image. Usually, pixel features consist of color values from several color spaces. Since only the information of a single pixel is considered in pixel-based approaches, the mapping of the pixel features must be robust to achieve satisfactory results for the whole image segmentation.

Some approaches use standard color spaces, such as $L^*a^*b^*$, $L^*u^*v^*$ and Pseudo-hue [35] (Equation 2-1), as features, which have shown high contrast between skin and lip component intensity [29–31, 33].

$$Ph(x,y) = \frac{Red(x,y)}{Red(x,y) + Green(x,y)} \tag{2-1}$$

Other approaches attempt to find a combination of the standard color maps, aiming to reduce the number of features, increase the mouth-lip discrimation and increase the accuracy rate [32, 36]. Common color transformations used for lip segmentation are shown in Figure **2-1**.



(a) Original Image

(b) u component

(c) v component

(d) L component

(e) a component

(f) b component

(g) Pseudo hue

Figure 2-1: Common color transformations for lip segmentation

The most common pixel classifiers are based on clustering or parametric modeling of the feature space. K means [36] and Fuzzy C means [29–31] are examples of the former. In the latter category, the most common model used to estimate the pixel feature distribution is the Gaussian Mixture Model (GMM)[32, 33], which usually uses the K means algorithm as an initialization stage. However, it is common to find an overlapping of mouth structures on the feature space, which reduces the clustering techniques performance.

2.1.2 Texture-based Lip Segmentation Technique

The proposed technique performs a color mapping of a window centered on each pixel, then a k-nearest neighbor k–nn classifier is used in the new space. The $W \times W$ sized window is always considered odd, so it has a central pixel. The mapping can be achieved through Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA). Depending on the classification accuracy, one of those transformations is chosen. By using the color information of the pixel neighborhood, the representation distance of near pixels is reduced. This feature extraction methodology aims to reduce the number of spurious regions.
Given a RGB color image $\boldsymbol{I} \in \mathbb{R}^{N \times M \times 3}$, the feature vector $\boldsymbol{x}$ for each pixel (x, y) on the image is composed of each element of the window:

$$\boldsymbol{x}(x, y) = \boldsymbol{I}(x + w_x, y + w_y, k) \qquad w_x, w_y \in \left[-\frac{W-1}{2}, \frac{W-1}{2}\right]; k \in [1, 3]$$

such that $\boldsymbol{x} \in \mathbb{R}^d$, being $d = W \times W \times 3$. The feature vector building procedure is shown in Figure 2-2.

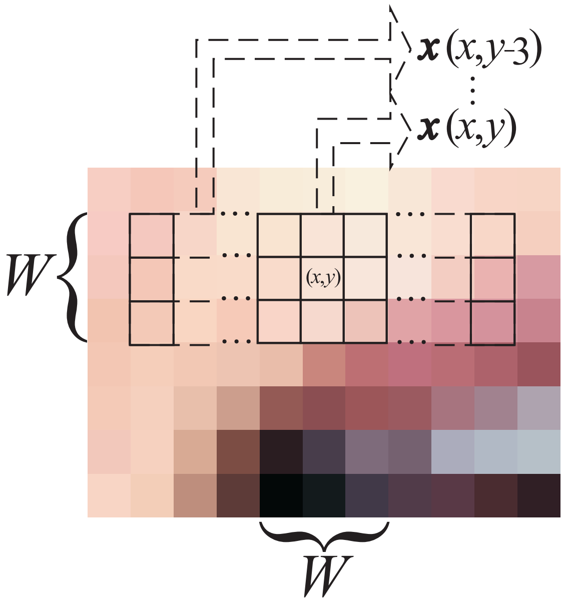


Figure 2-2: Texture feature building

On a training set with R elements and C different classes, each sample is a couple $(\boldsymbol{x_r}, c_r)$, where $c_r \in [1, C]$ is the label associated to the feature vector $\boldsymbol{x}_r$. The set of R training samples is denoted as $\boldsymbol{X} = [\boldsymbol{x}_1| \ldots \boldsymbol{x}_r| \ldots |\boldsymbol{x}_R]$.
Since PCA and LDA are both a linear transformation, the transformation matrix is needed for each of them. For PCA, the transformation is given by the matrix $\boldsymbol{W}_{PCA}$ of eigenvectors of the matrix $\boldsymbol{X}$. For LDA the transformation is given by the matrix $\boldsymbol{W}_{LDA}$, which is

composed of the eigenvectors of the matrix $S_w^{-1}S_b$, where S_w is the *within-class* scatter matrix and S_b is the *between-class* scatter matrix, shown in Equations 2-2 and 2-3.

$$S_w = \sum_{c=1}^{C}\sum_{r=1}^{N_c}(\boldsymbol{x}_r^c - \boldsymbol{\mu}_c)(\boldsymbol{x}_r^c - \boldsymbol{\mu}_c)^\mathsf{T} \tag{2-2}$$

$$S_b = \sum_{c=1}^{C}(\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^\mathsf{T} \tag{2-3}$$

Finally, the new feature vector is defined by $\boldsymbol{y}(x, y) = \boldsymbol{W}^\mathsf{T}\boldsymbol{x}(x, y)$

2.1.3 Lip Contour Extraction

Once the RoI is detected, the segmentation result can be used as the initialization procedure for the lip contour extraction stage. This task can be achieved by fitting a snake to lip boundaries. A snake is an elastic curve represented by a set of control points, and it is used to detect important visual features, such as lines, edges, or contours. The snake control point coordinates are iteratively updated, converging towards a minimum of an energy function defined on basis of curve smoothness constraints and a matching criterion to desired features of an image.

A kind of a snake, *Jumping Snake* originally proposed by Eveno et al. [34] and improved by Gómez-Mendoza et al. [37], is specially designed to detect lip boundaries. The *Jumping Snake* is a simplified form of active contour that properly approximates the outer lip contour in color images. Pseudo-Hue (ph) and luminance (L) color components are used to compute the gradient flow that controls the snake evolution. N Points are added on each iteration at both left and right side of the seed incrementally, preserving a horizontal distance (Δ) and aiming to maximize the normalized gradient flow (φ) of $ph - L$ passing through each generated line segment. At the end of the iteration, vertical seed position is re-computed as the mean vertical position of the N added points that led the creation of the line segments with highest gradient flows.

Associated gradient flow for each upper and lower lip snake point ($\mathbf{p}_i$) are computed as given in Equations 2-4 and 2-5, respectively, where $\mathbf{dn}_-$ and $\mathbf{dn}_+$ are the normalized gradient vectors perpendicular to line segments conformed by the points located at left or right side of the seed with the seed itself, and $\nabla\{\cdot\}$ stands for gradient operator.

$$\varphi_i = \int_{\mathbf{p}_{i-1}\mathbf{p}_i} \frac{\nabla\{ph - L\}\cdot\mathbf{dn}_-}{|\mathbf{p}_{i-1}\mathbf{p}_i|} + \int_{\mathbf{p}_i\mathbf{p}_{i+1}} \frac{\nabla\{ph - L\}\cdot\mathbf{dn}_+}{|\mathbf{p}_i\mathbf{p}_{i+1}|} \tag{2-4}$$

$$\varphi_i = \begin{cases} \displaystyle\int_{\mathbf{p}_i\mathbf{p}_{i+1}} \frac{\nabla\{ph\}\cdot \mathbf{dn}_+}{|\mathbf{p}_i\mathbf{p}_{i+1}|} & i \in [1, N] \\ \displaystyle\int_{\mathbf{p}_{i-1}\mathbf{p}_i} \frac{\nabla\{ph\}\cdot \mathbf{dn}_-}{|\mathbf{p}_{i-1}\mathbf{p}_i|} & i \in [N+2, 2N+1] \end{cases} \tag{2-5}$$

Image gradient estimation is computed by convolving the image with the Scharr operators. The 3×3 horizontal ∇_x and vertical ∇_y Scharr operators used are given by:

$$\nabla_x = \begin{bmatrix} -3 & 0 & 3 \\ -10 & 0 & 10 \\ -3 & 0 & 3 \end{bmatrix} \qquad \nabla_y = \begin{bmatrix} -3 & -10 & -3 \\ 0 & 0 & 0 \\ 3 & 10 & 3 \end{bmatrix} \tag{2-6}$$

2.2 Visual Feature Extraction

The most important issue in speech/speaker recognition using visual signals is the feature extraction stage. Visual features have to be informative, discriminative and robust, aiming to model the speech dynamic, to make each class distinguishable, and to be accurately extracted under different scene conditions, respectively.

Since most of the visual speech information is located in mouth region, the feature extraction techniques aim to model the mouth shape and/or appearance as the RoI. There are approaches which model the mouth appearance, known as low-level or appearance-based techniques. A second group of approaches introduce a way to extract and/or model the mouth contour, these features are known as high-level or shape-based features. Other approaches have proposed the use of both appearance and shape features to model the mouth. Finally, there is a set of approaches which add temporal information to the feature space by extracting features from the partial derivative of the video sequence respect the time.

2.2.1 Low-level Features

These approaches assume that each pixel on the ROI can be used as a mouth modeling feature. Bearing this in mind, the first approach should be the use of the same ROI as the feature set. But, the drawback on this technique is *the curse of dimensionality*, which means, the high feature space dimensionality makes mouth's dynamic statistically unable of being modeled, for instance, by a Hidden Markov Model. As an example, if the ROI is 32×32 size, the feature space dimension is $D = 1024$. To overcome the *curse of dimensionality*, the authors propose dimension reduction by means of image transformations, mainly Principal Component Analysis (PCA) [11, 23, 38], Discrete Cosine Transform (DCT) [6, 23, 38], Linear Discriminant Analysis [1, 23], Discrete Wavelet Transform (DWT) [38, 39]. The next sections introduce two representative transformations, PCA and DCT, the later ones have similar structures.

EigenLips

EigenLips is the given name to the PCA-based image transformation for mouth RoI, introduced by [11]. Let an image be reshaped as a vector $\mathbf{x} = \{x_\rho \in \mathbb{R} : \rho \in [1, p]\}, p = n \times m$, and an input training sequence $X_r = \left[\mathbf{x}^1|\cdots|\mathbf{x}^t|\cdots|\mathbf{x}^{T_r}\right], X_r \in \mathbb{R}^{p\times T_r}$. The assembly of image sequences can be rewritten as the centered matrix $\mathbf{X} = [X_1|\cdots|X_r|\cdots|X_R], \mathbf{X} \in \mathbb{R}^{p\times T}, T = \sum_r T_r$, i.e., $\mathscr{E}\{\mathbf{X}\} = \mathbf{0}$.

Conventional projection by PCA states that there will be a couple of ortonormal matrices, $\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}_p$, plus a diagonal matrix $\varSigma$, such as the simple linear decomposition takes place, that is, $\mathbf{X} = \mathbf{U}\varSigma\mathbf{V}^\mathsf{T}$, $\mathbf{V} \in \mathbb{R}^{T\times T}, \mathbf{U} \in \mathbb{R}^{p\times p}$, where matrix $\varSigma \in \mathbb{R}^{T\times p}$ holds first descend–ordered $q \leq p$ as the largest eigenvalues of matrix $\mathbf{X}$, $\nu_1 \geqslant \nu_2, \ldots, \geqslant \nu_q \geqslant \nu_{q+1}, \ldots, \geqslant \nu_p \geqslant 0$, $p = \operatorname{rank}(\mathbf{X})$.

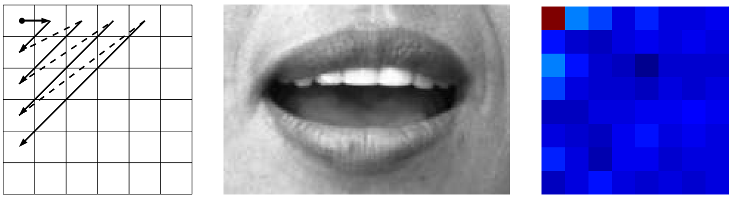
The matrix $\mathbf{V} = [\mathbf{v}_1|\cdots|\mathbf{v}_k|\cdots|\mathbf{v}_p]$ defines the basis vectors of a new space, such that each “eigenlip” component of an image $\mathbf{x}_r^t$ is determined by:

$$y_{kr}^t = \mathbf{v}_k^\mathsf{T}\mathbf{x}_r^t \tag{2-7}$$

Discrete Cosine Transform

The DCT has been widely used in image processing due to its energy compaction properties higher than DFT. Given an image $\mathbf{x}(i,j)$, its DCT is given by the Equation 2-8. As most of the energy is located on the first coefficients, the feature vector is assembled by a zig-zag scan of the DCT, as shown in Figure 2.3(a). Usually the DCT is estimated is as large as the image, but only a subset of coefficients is selected as the final feature set. This selection is performed by means a relevance criterion or the classification performance.

$$\mathbf{D}(u,v) = \sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{x}(i,j)\cos\left(\frac{(2i+1)u\pi}{n}\right)\cos\left(\frac{(2j+1)v\pi}{m}\right) \tag{2-8}$$



(a) DCT coefficients zig-zag scan

(b) Original Image

(c) 8×8 first DCT coefficients

Figure 2-3: Discrete Cosine Transform Representation

2.2.2 High-level Features

High-level feature extraction assumes that most of the speech information can be extracted from speaker lip contour, either the outer and/or inner one. The above implies a lip shape modeling and its tracking over a video sequence, this is why these set of features are so called *shape-based* features. There are two main ways of lip modeling, geometric-based and model-based.

Geometric Features

Geometric features are composed by a set of measurements made on the extracted lip contour. Some of those measurements have a shape meaning which can be easily read, such as contour perimeter, area, length and width. Other measurements extracted from lip contour are image moments and Fourier descriptors, which are invariant to affine transformations, although they have no a direct reading on the image. Such features contain significant visual speech information and their properties made them useful in speech/speaker identification tasks [7, 13, 40].

Invariant image moments Let an image $\mathbf{I} \in \mathbb{R}^{N \times M}$. The two-dimensional $(p+q)$–th order moment is defined as

$$m_{pq} = \sum_{x=1}^{N} \sum_{y=1}^{M} x^p y^q I(x,y) \qquad \forall p,q \in \mathbb{Z}^+ \tag{2-9}$$

The *central moments* are defined in the Equation 2-11, where $\bar{x} = \frac{m_{10}}{m_{00}}$ and $\bar{y} = \frac{m_{01}}{m_{00}}$ are the image centroid coordinates.

$$\mu_{pq} = \sum_{x=1}^{N} \sum_{y=1}^{M} (x - \bar{x})^p \, (y - \bar{y})^q \, I(x,y) \qquad \forall p,q \in \mathbb{Z}^+ \tag{2-10}$$

$$\tag{2-11}$$

By dividing each central moment by its correspondent scaled (00)–th moment (Equation 2-12), the scale invariant moment version can be constructed.

$$\eta_{pq} = \frac{\mu_{ij}}{\mu_{00}^{1+\frac{i+j}{2}}} \qquad \forall p,q \in \mathbb{Z}^+ \tag{2-12}$$

The non-orthogonal centralized moments are translation invariant and can be normalized with respect to changes in scale. However, to enable invariance to rotation they require reformulation. Hu [41] derived a set of nonlinear centralized moment expressions from algebraic invariants applied to the moment generating function under a rotation transformation. The proposed set is absolute orthogonal and can be used for scale, position, and rotation invariant pattern identification. These were used in a simple pattern recognition experiment to successfully identify various typed characters. They are computed from normalized centralized moments up to order three and are shown below,

$$M_1 = \eta_{20} + \eta_{02} \tag{2-13}$$

$$M_2 = (\eta_{20} - \eta_{02})^2 + (2\eta_{11})^2 \tag{2-14}$$

$$M_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \tag{2-15}$$

$$M_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \tag{2-16}$$

$$M_5 = (\eta_{30} - \eta_{12})(\eta_{30} + \eta_{12}) \left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] + \tag{2-17}$$

$$(3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] \tag{2-18}$$

$$M_6 = (\eta_{20} - \eta_{02}) \left[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] + 4\eta_{11}(n_{30} + \eta_{12})(n_{21} + \eta_{03}) \tag{2-19}$$

$$M_7 = (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12}) \left[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right] + \tag{2-20}$$

$$(\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03}) \left[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right] \tag{2-21}$$

Model-based Features

In the model-based approach, feature vectors are extracted from a parametric or statistical model of the lip contour, which can be obtained as discussed in Section 2.1.3. In the parametric approach, the snake's points or its radial vectors, can be directly employed as visual speech features. Similarly, Active Shape Models (ASMs) can be used as visual features by applying the model PCA on the vector of point coordinates of the estimated lip contour [28].

2.3 Hidden Markov Models

A Hidden Markov Model (HMM) is a model for double random processes, which is composed of two layers. The first layer models the temporal feature evolution, i.e., the inner dynamic of the process; this layer is known as *hidden layer* and has a finite number of states. The second one, known as *observable layer*, models the occurrence of a random event on each state. HMMs have been widely used on applications with signals with a non-evident dynamic, since it is able to disconnect the time evolution process from the event occurrence process.

2.3.1 HMM Parameter Description

A HMM consists of a set of N nodes (states), each of which is associated with an observation model. The parameters of the model include [42]:

1. An initial state probability distribution $\boldsymbol{\pi}$ with elements $\{\pi_i\}$, to represent the likelihood of i–th state in the first state of the sequence s_1, $\pi_i = P(s_1 = i); i \in [1, N]$, such that the Equation 2-22 is satisfied.

$$\sum_{n=1}^{N} \pi_n = 1 \tag{2-22}$$

2. A state transition probability distribution $\boldsymbol{A} = \{a_{ij}\}; i, j \in [1, N]$ for the transition probability to node j given that the HMM is currently in state i, $a_{ij} = P(s_k + 1 = j|s_k = i)$. As a probability distribution, the matrix A has to meet the restrictions in Equation 2-23.

$$a_{ij} \geq 0 \qquad \sum_{j=1}^{N} a_{ij} = 1; \forall i \tag{2-23}$$

3. An observation probability set for each state in the model $\boldsymbol{B} = \{b_j(\cdot)\}$. According to how $\boldsymbol{B}$ is chosen, the HMM will be either discrete or continuous.

   Former case allows only discrete observations of a fixed codebook, where the output distribution in each emitting state i consists of a separate discrete probability b_{im} for each observation symbol m. For discrete HMM, the observation probability set can be written as a matrix $\boldsymbol{B} \in \mathbb{R}^{N \times M}$, constrained to the Equation 2-24.

$$\sum_{m=1}^{M} b_{im} = 1; \forall i \tag{2-24}$$

   Latter case uses parametric distributions of a predetermined form that usually are based on weighted sums (mixtures) of multivariate Gaussian densities. The probability density function (pdf) of a mixture with M components is given by the Equation 2-25, where $P_{\boldsymbol{B}}(\boldsymbol{x}|m)$ is the conditional probability function of the mixture given the component m.

$$P_{\boldsymbol{B}}(\boldsymbol{x}) = \sum_{m=1}^{M} p_m P_{\boldsymbol{B}}(\boldsymbol{x}|m) \qquad \sum_{m=1}^{M} p_m = 1 \tag{2-25}$$

   For a multivariate Gaussian the pdf is defined by the Equation 2-26.

$$\begin{aligned} P_{\boldsymbol{B}}(\boldsymbol{x}|m) &= \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m); \forall m \in [1, M] \\ \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) &= (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}_m|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_m)^\mathsf{T} \Sigma_m^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_m)\right] \end{aligned} \tag{2-26}$$

   And $\boldsymbol{B}$ comprises all combination weights, mean vectors and covariance matrices, as the parameter set of the Gaussian mixture,

$$\boldsymbol{B} = \{\boldsymbol{p}; \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_M; \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M\}$$

Finally, the parameter set of a HMM will be denoted as $\lambda = \{\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B}\}$.

2.3.2 HMM Training

Let the input training data set,

$$\mathscr{X} = \{(\boldsymbol{X}_r, c_r) : \boldsymbol{X}_r \in \mathbb{R}^{p \times T_r}, c_r \in \mathbb{Z}, r \in [1, R]\}$$

composed of R observations, where c_r is the class label of the sample $\boldsymbol{X}_r$. Each sample r is represented by a feature vector sequence of length T_r, $\boldsymbol{X}_r = \{x^t_{\rho r} : t \in [1, T_r], \rho \in [1, p]\}$. Since there is no an analytical solution for HMM parameter estimation, iterative and gradient training techniques have been used to estimate a suboptimal parameter set. Among the former, the predominant training technique is Maximum Likelihood Estimation (MLE), which maximizes the likelihood of the training data observations:

$$f_{MLE}(\boldsymbol{\Lambda}) = \sum_{r=1}^{R} \log P(\boldsymbol{X}_r | \lambda_{c_r}) \tag{2-27}$$

where P is the likelihood of the observation sequence $\boldsymbol{X}_r$ given the model λ_{c_r} of the correct transcription class c_r.

2.3.3 HMM-based Classifier

Given a sequence of observations $\boldsymbol{X} = [\boldsymbol{x}_1 | \ldots | \boldsymbol{x}_t | \ldots | \boldsymbol{x}_T]$, a label $\hat{c}$ has to be assigned. Since each class is modeled by a HMM, the whole set of HMM parameters comprises C models, $\Lambda = \{\lambda_c : c \in [1, C]\}$, with λ_c denoting the parameter set of the c–th class. The classification rule is given by the Equation 2-28.

$$\hat{c} = \mathrm{argmax}_c(P_\Lambda(\boldsymbol{X}|c)) \tag{2-28}$$

The direct calculation of $P_\Lambda(\boldsymbol{X}|c)$ implies a likelihood sum of all possible paths, which represents a high computational cost. Due to the above, the *forward-backward* algorithm is used as a more efficient estimation procedure.

3 Experimental Setup

The general methodology for VSR follows these stages:

1. *Preprocessing.* Initial step on VSR systems. In this stage the RoI search area is reduced and the image is “cleaned” by filtering.

2. *Lip segmentation.* In this stage, RoI is extracted from the detected face image. Then, the segmentation result is used as the initialization of lip contour detection.

3. *Feature extraction.* Each feature extraction methodology introduced in 2.2 is used to characterize the RoI. Also, a dimension reduction is incorporated in some strategies, aiming to avoid the *curse of dimension* on the classifier training.

4. *Recognition system.* A conventional HMM-based classifier is implemented as the recognition system. One HMM is trained to model one class dynamic. HMM parameter tunning is performed by classification performance.

Figure **3-1** shows the experimental outline of the methodology used in this work.

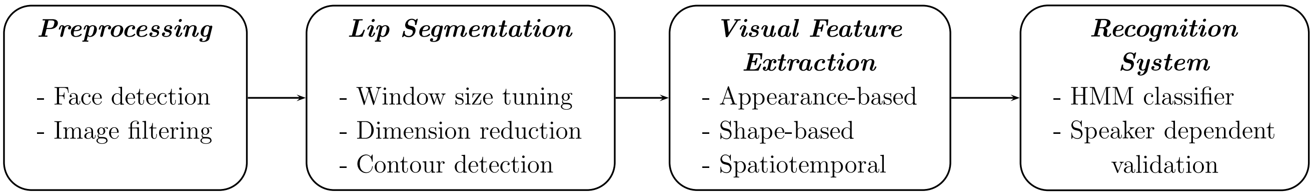


Figure 3-1: Methodology outline

3.1 Database Description

All feature extraction techniques were applied on a video database built at Universidad Nacional de Colombia (Manizales). The database is composed of 14 subjects, 8 male and 6 female, which have various skin complexions with no particular lipstick. Each recording include full frontal face color video sequences of each person, with the pronunciation of the Spanish alphabet (26 letters). Speaker’s utterances were recorded three times. Images were captured with a Basler Scout scA 640-70fc industrial camera. Acquisition parameters, such as ISO, exposure, aperture and white balance, were kept along the recoding session. Also, three light sources were used to control illumination conditions, two of them located at left and right side of the camera and the third one was a ceil lamp. Sequences were recorded as ppm-formated images with a resolution of 658 × 492 pixels and a frame rate of 50 frame/second. Some image samples from database are shown in Figure **3-2**.

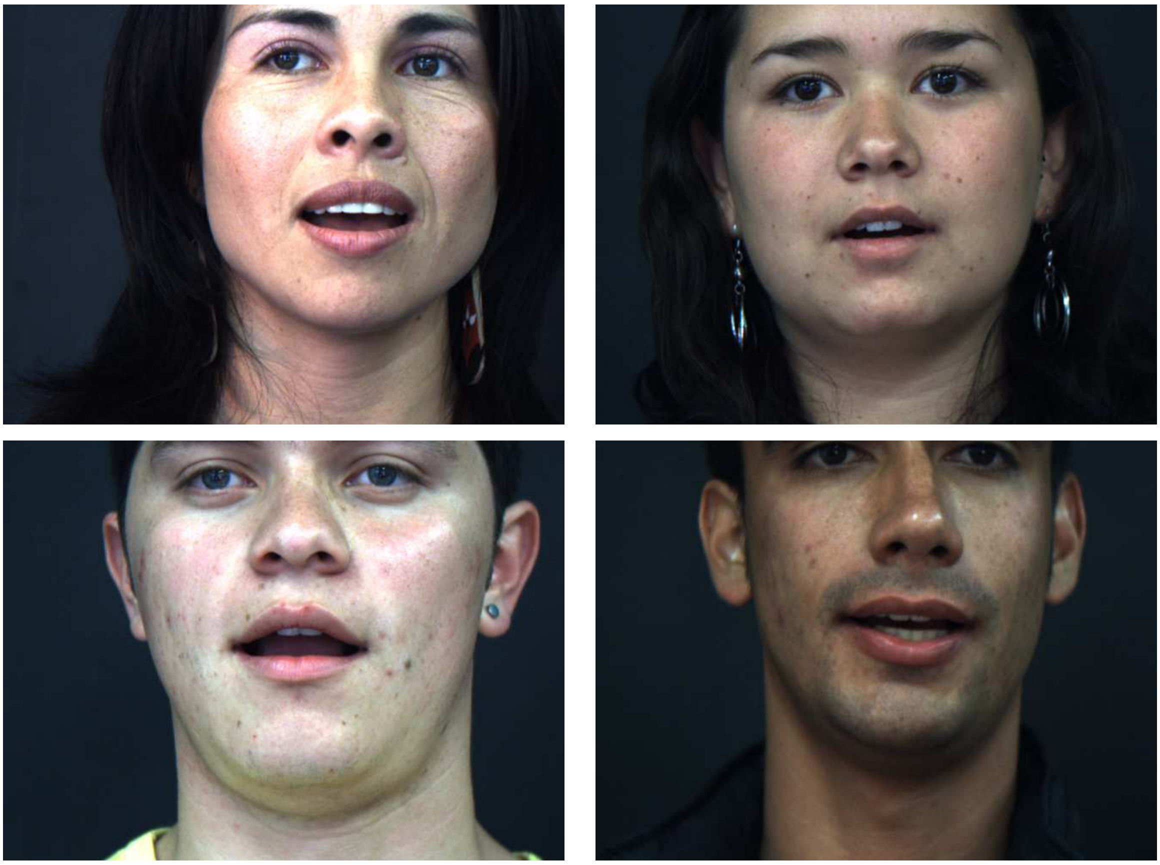


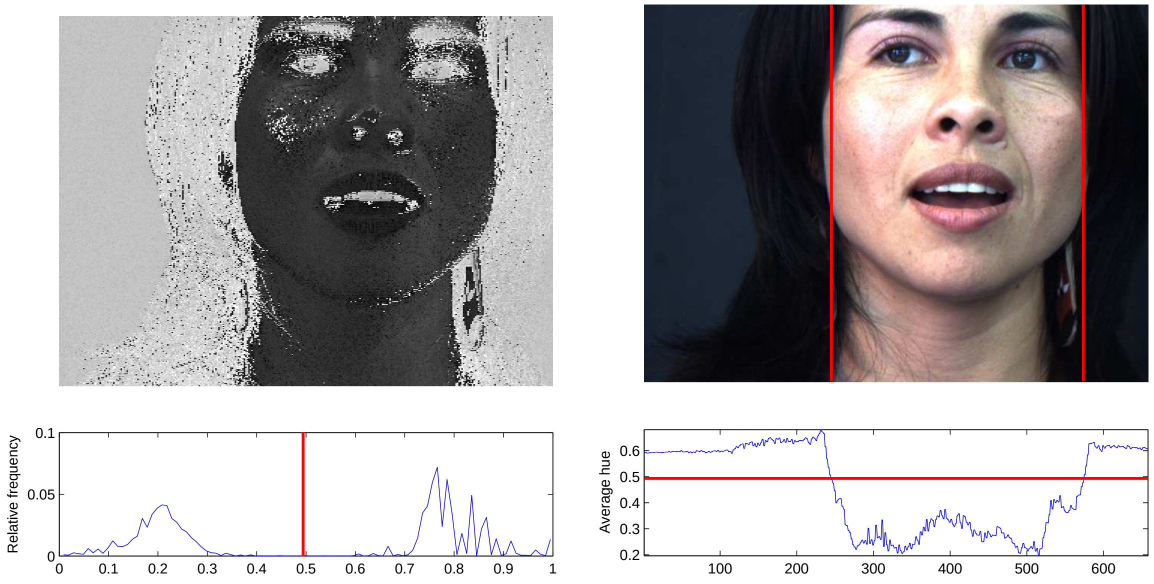
Figure 3-2: Database image samples of 4 subjects

3.2 Preprocessing

As said above, two task are performed on the preprocessing stage. First, the face is located to reduce the mouth search area. Second, the image is filtered to reduce noise and smooth the face. This stage aims to improve the segmentation and feature extraction stages.

3.2.1 Face Detection

Since speaker's face and background have a high contrast, a hue-based segmentation strategy is suitable as a face detector. First the image hue map is estimated. Then, a threshold is calculated by Otsu's method [43], to segment the pixels in two groups, namely, skin and background. Finally, the column-wise average hue is thresholded to find the skin region on the image. Figure 3.3(a) shows the hue map of a test image (top) and its histogram (bottom). The red line is located on the Otsu's threshold level. Figure 3.3(b) shows the original image and the boundaries detected for skin region (top), and the average hue for each column (bottom).



(a) Hue map

(b) Face boundaries

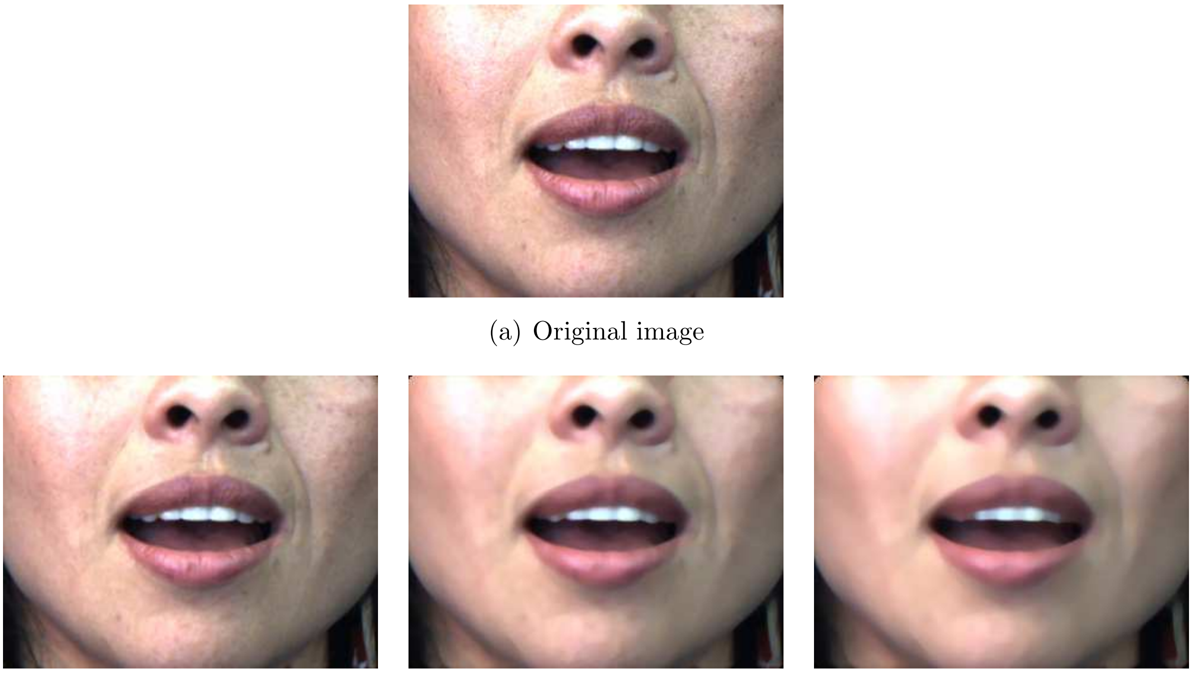
Figure 3-3: Hue-thresholding-based face detection

3.2.2 Image Filtering

Image filtering procedure is used as a "cleaning" strategy. To smooth the face region an median filter is used, since it has the behavior of a low pass filter and, additionally, it keeps the boundaries location well defined. Median filter in given by the Equation 3-1, where W is the size of the analysis window.

$$\boldsymbol{I}_{Med}(x,y) = \text{Median}\{\boldsymbol{I}(x+w_x, y+w_y)\} \qquad w_x, w_y \in \left[-\frac{W-1}{2}, \frac{W-1}{2}\right] \tag{3-1}$$

Different window sizes were tested to tune W. The appropriate window was selected by visual inspection. From Figure **3-4**, it can be seen that for a size of $W = 11$ each face structure is enough smooth and boundary definition level is kept.



(a) Original image

(b) $W = 3 \times 3$

(c) $W = 7 \times 7$

(d) $W = 11 \times 11$

Figure 3-4: Image smoothing by median filter for different window sizes

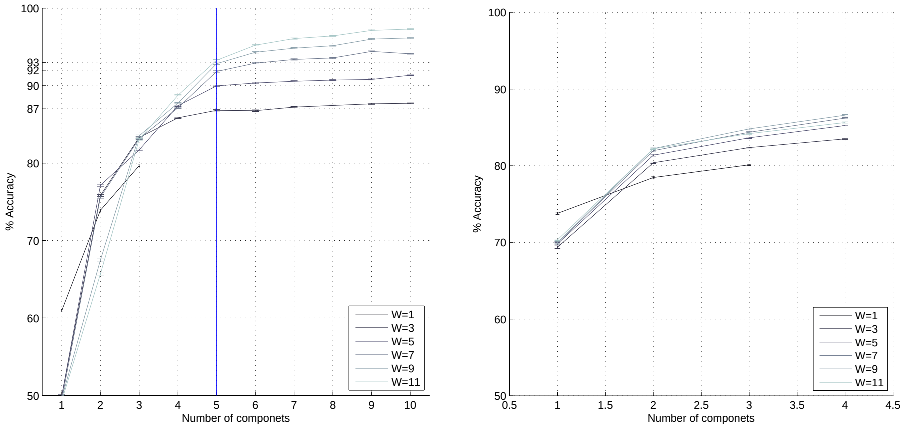
3.3 Lip Segmentation

Pixel classification and contour extraction procedures are performed on this stage. To train the pixel classifier a subset of 14 images from the image database is selected, one for each speaker. A 300 × 200 pixels window centered on the mouth is extracted from each image. Five structures are manually segmented on each window, lip, teeth, tongue, skin and dark (low illumination) pixels. The image database was split in two sets, 16 images for tuning and two images for testing. From the former, a set of 100.000 labeled pixels was randomly selected.

3.3.1 Pixel Lip Segmentation

The cross-validation strategy used to tune the segmentation methodology. Cross-validation consists on the division of each database into ten folds containing different records. nine of these folds are used for training and the remaining one for validation purposes. Transformation matrices and a classifier is obtained from training set. Then, an incremental training, from the first to the last component, is used to classify validation pixels. The procedure is repeated changing the training and validation folds, until the ten folds are used to validate the classifier. The parameter tuning is performed by means of the overall accuracy of a k-nn classifier.

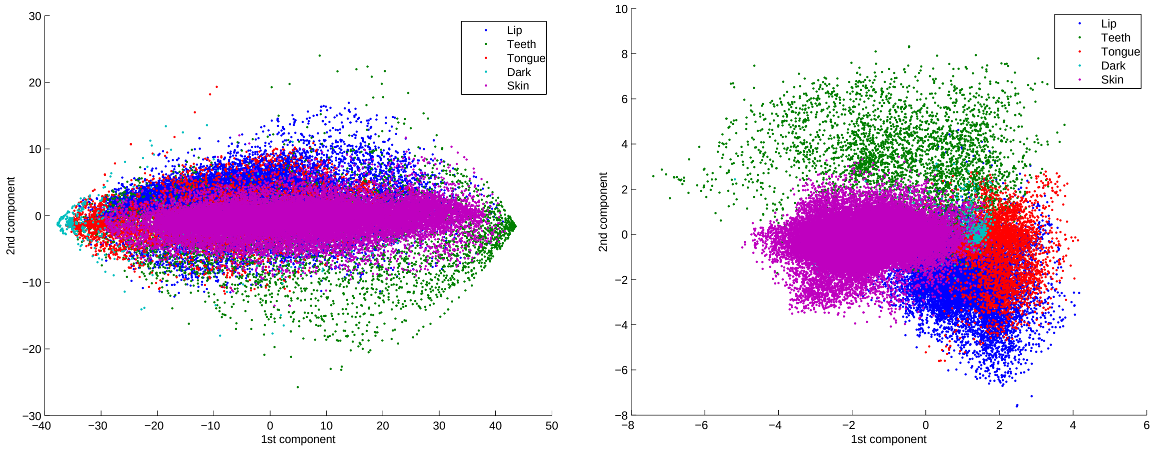
In this stage, a dimension reduction technique has to be selected. Additionally, the neighborhood size and the number of components is tuned. Figures 3.5(a) and 3.5(b) show the performance of the k–nn ($k = 1$) classifier for PCA and LDA. Window sizes of $[1, 3, 5, 7, 9, 11]$ were tested on the experiment. Average classification accuracy and its standard deviation is reported for each configuration.



(a) Principal Component Analysis (b) Linear Discriminant Analysis

Figure 3-5: k–nn classifier accuracy vs. the number of representation components.

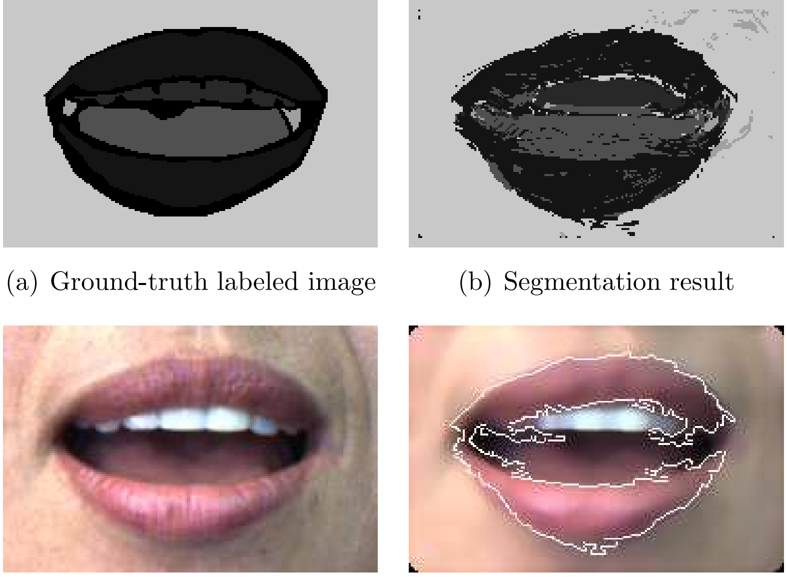
From results reported, the first 5 components of PCA representation for a window size of $W = 9$ is chosen as the configuration for the segmentation algorithm. Figures 3.6(a) and 3.6(b) show the representation of the tunning set on the 2 first components for both transformations, PCA and LDA, respectively.



(a) Principal Component Analysis (b) Linear Discriminant Analysis

Figure 3-6: Representation of tuning set on the 2 first components for $W = 9$.

To test the algorithm, an image from the test was automatically segmented. Results are shown in Figure 3-7.



(a) Ground-truth labeled image

(b) Segmentation result

(c) Original image

(d) Segmented lip contour

Figure 3-7: Methodology segmentation results for a test image.

3.3.2 Lip Contour Extraction

Lip contour extraction was performed by using the methodology introduced in 2.1.3. Gradient maps were estimated from the filtered face image. Seeds for upper and lower snakes were calculated as follows:

1. Lip binary map is selected from segmentation image.
2. Image centroid is calculated.
3. The upper and lower seeds are located marginally above and under the first and last lip pixel along the centroid y axis, respectively.

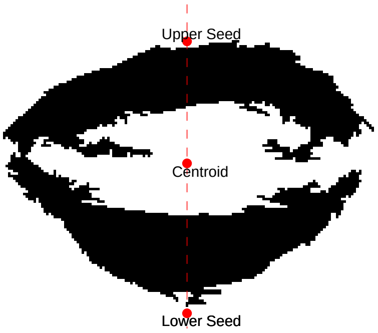
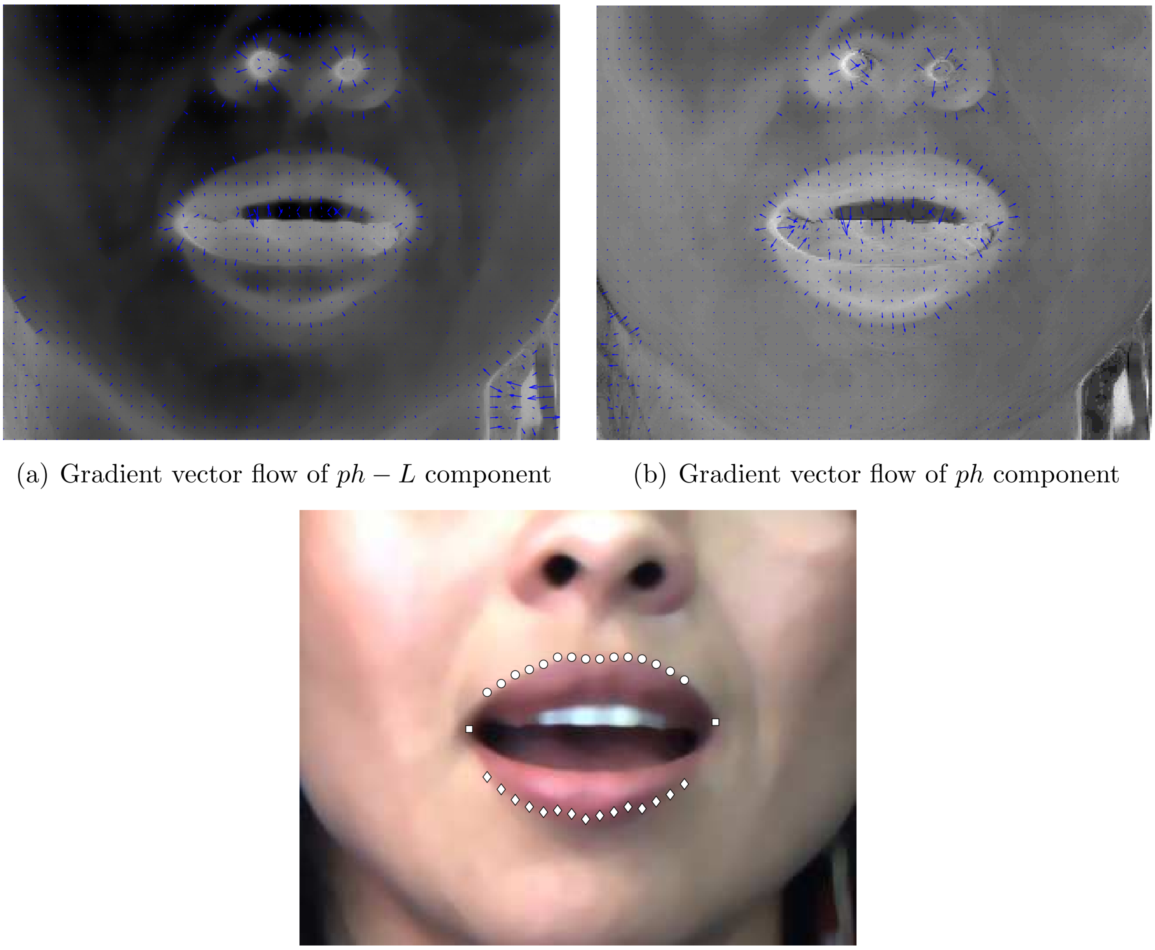


Figure 3-8: Lip contour initialization seeds

Figures 3.9(a) and 3.9(b) show the gradient vector flow for $ph - L$ and ph components, which are used as objective function by the *Jumping-Snake* algorithm. 3.9(c) show the gradient vector flow and the snake adjustment on a test image, respectively.



(a) Gradient vector flow of $ph - L$ component

(b) Gradient vector flow of ph component

(c) *Jumping-Snake* adjustment

Figure 3-9: Lip contour extraction

3.4 Visual Feature Extraction

In this section, each feature extraction methodology is used to characterize the image sequence database. Additionally, parameters of the methodologies are tuned and results of them, as image features, are shown.

3.4.1 EigenLips Representation

EigenLips representation is the result of a dimension reduction of the gray level image pixels. To estimate the transformation matrix $\boldsymbol{V}$ a set of N images is randomly selected from the database and arranged on a matrix $\mathbf{X} \in \mathbb{R}^{p \times N}$, where p is the number of pixels per image and N is the number of images. In this work, $N = 1000$ images were selected and each image was scaled to $p = 48 \times 59 = 2832$. Figure **3-10** shows the normalized accumulated eigenvalues of the matrix $\mathbf{X}$. To reduce the space dimension, a certain amount of variance has to be selected. For selected dataset, 90% of the total variance is achieved at 20 components.

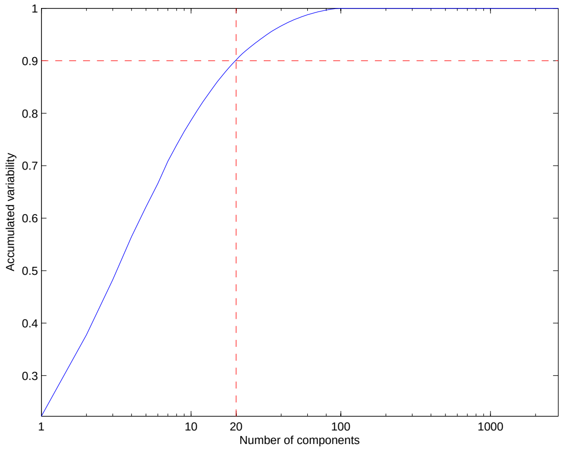
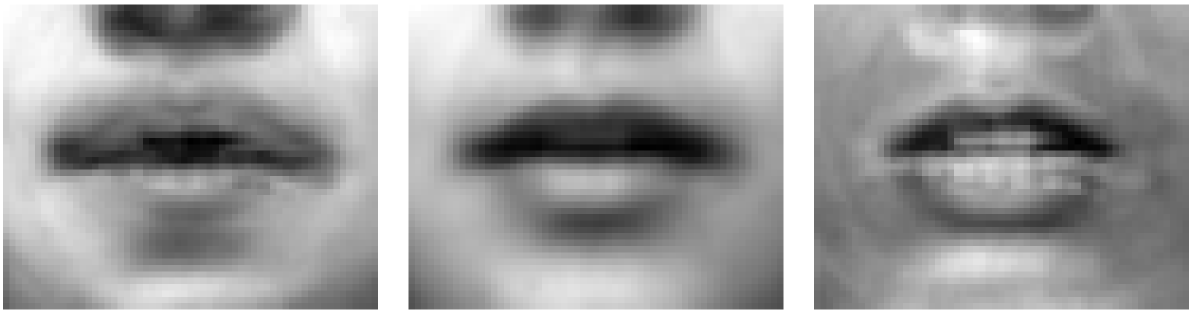


Figure 3-10: Normalized accumulated eigenvalues of dataset for eigenlips representation

The average image of the reconstruction using 20 PCA components is shown in Figure 3.11(b), and the images of ±3 standard deviation are shown in Figures 3.11(a) and 3.11(c). It can be seen that each of the three images attempt to explain a mouth pose.

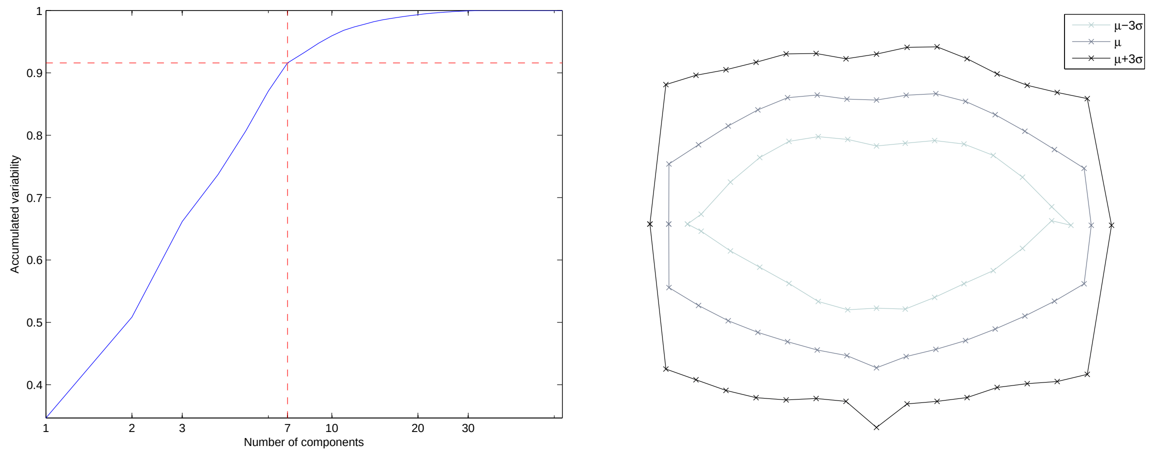


(a) $\mu - 3\sigma$ (b) μ (c) $\mu + 3\sigma$

Figure 3-11: Image reconstruction using 20 PCA components

3.4.2 Active Shape Model

This feature set is built from the PCA transformation of the coordinates of the contour points. As EigenLips representation, a reduced number of components is chosen based on the amount of explained variability. The first seven components are chosen since they explain 90% of the full variability (Figure 3.12(a)), they are chosen to build the new space. Figure 3.12(b) shows the mean shape ±3 standard deviation of the dataset reconstructed from the ASM representation.



(a) Normalized accumulated eigenvalues for lip contour points

(b) Average contour represented on 7 components and its standard deviation

Figure 3-12: ASM representation of lip contour

3.4.3 Spatiotemporal Representation

Final feature extraction methodology is composed of appearance, shape and temporal information. Feature vector of an input sequence on the original space, at time instant k is built as:

$$\boldsymbol{x}(k) = \left[\mathbf{x}(k)|C_x(k)|C_y(k)|\frac{\partial \mathbf{x}}{\partial t}(k)\right] \tag{3-2}$$

where, $\mathbf{x}$ is the image RoI reshaped as introduced in 2.2.1, C_x and C_y are the coordinates x and y of the lip contour, and $\frac{\partial \mathbf{x}}{\partial t}(k)$ is the partial time derivative of the sequence at time k. As in methodologies above, a space dimension reduction has to be performed. Figure **3-13** shows the eigenvalues for PCA decomposition of the training dataset. It can be seen that a high number of features is required to explain 90% of the total variance. Hence, in classification tasks only 70% of variance will be assumed, which is achieved at 33 components.

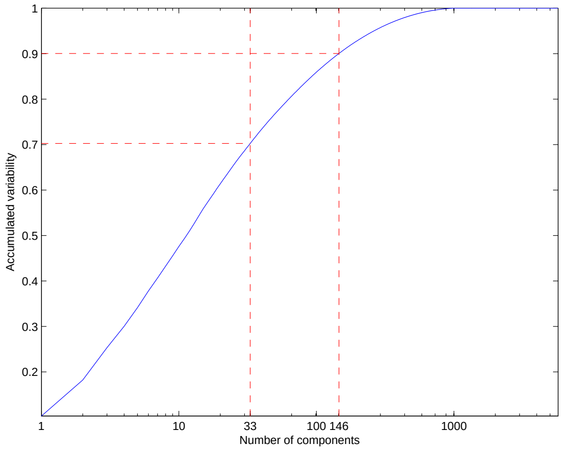


Figure 3-13: Normalized accumulated eigenvalues of dataset for spatiotemporal representation

Figure 3-14 shows the mean shape and appearance ±3 standard deviation of the dataset reconstructed from the spatiotemporal representation.

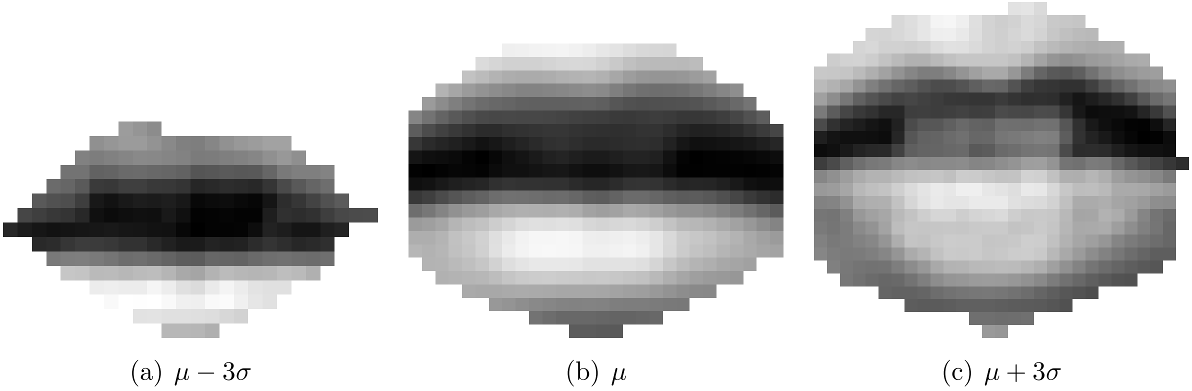

(a) $\mu - 3\sigma$ (b) μ (c) $\mu + 3\sigma$

Figure 3-14: Active appearance representation for selected components

3.5 Recognition System

On a speech recognition system, speaker dependent and speaker independent validation methodologies are commonly used. In the former, the speaker is included in training and validation sets, while in the latter, each speaker belongs to just one set. In this work, speaker dependent tests are performed to tune the system and the speaker independent test is used to evaluate the performance of the selected topology.

After obtaining the visual feature vector for each feature extraction methodology, a HMM-based classifier is trained and its performance is evaluated against the validation data. Since the database contains three repetitions for each speaker, system validation is carried out by 3-fold cross-validation methodology. That is, the experiments are repeated 3 times, on each fold 2 repetitions are used as train set and the remaining one as validation set. From accuracy results, the best feature extraction methodology and the optimum HMM parameter configuration are selected.

A flexible left to right HMM topology has been adopted, which allows to have different number of states for each class. This number has to be proportional to the average class length $\bar{T}_c$, that is, $n_S = f\bar{T}_c$, being f a scalar factor. HMM topologies were tested for $f = [1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}]$ and $n_G = [1, 2, 4, 8]$.

3.5.1 Low Level Features

Results for HMM tested topologies using 2D-DCT and EigenLips feature extraction methodologies are shown in Tables **3-1** and **3-2**. On all topologies, EigenLips has an accuracy around 5% higher than 2D-DCT results, that makes the variance representation more informative on classification task than the spectral content of each image.

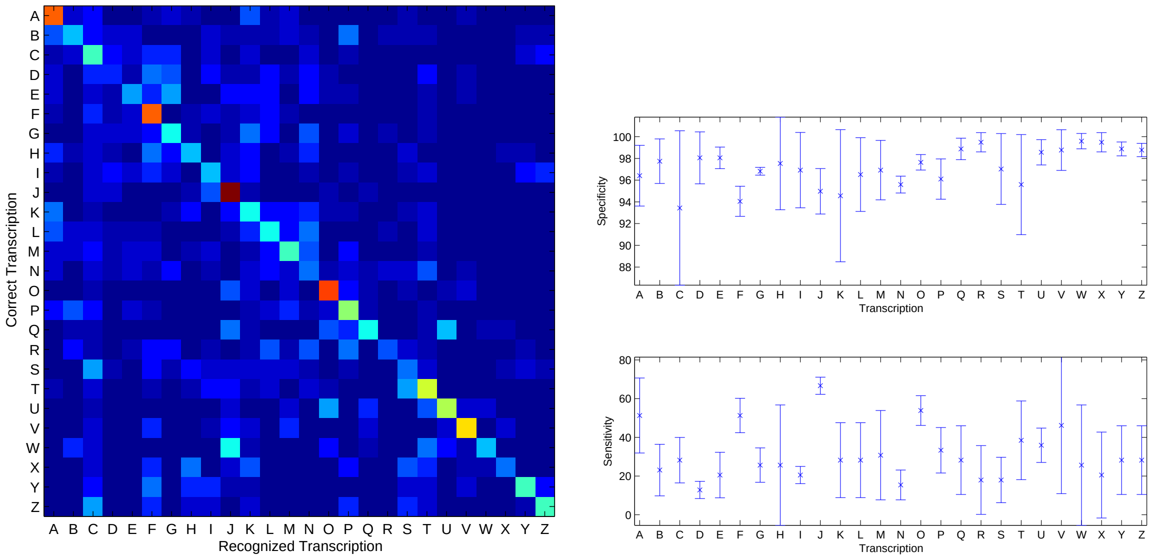
		n_G			
		1	2	4	8
f	$\frac{1}{8}$	17.00 ± 0.74	17.30 ± 0.90	17.79 ± 1.63	16.61 ± 1.12
	$\frac{1}{4}$	17.79 ± 0.45	17.50 ± 1.40	17.30 ± 0.62	17.40 ± 1.29
	$\frac{1}{2}$	17.99 ± 0.51	17.40 ± 1.29	16.51 ± 1.80	18.09 ± 2.81
	1	16.71 ± 1.68	16.71 ± 0.95	17.79 ± 3.08	17.40 ± 3.33

Table **3-1**: HMM parameter tuning for 2D-DCT

		n_G			
		1	2	4	8
f	$\frac{1}{8}$	21.40 ± 1.46	25.84 ± 0.17	28.21 ± 4.49	24.65 ± 4.31
	$\frac{1}{4}$	21.01 ± 2.23	28.60 ± 3.75	29.19 ± 3.91	24.95 ± 4.88
	$\frac{1}{2}$	22.58 ± 3.27	26.43 ± 3.75	25.64 ± 3.57	22.09 ± 2.24
	1	23.37 ± 2.42	28.80 ± 2.80	28.21 ± 3.57	23.37 ± 2.82

Table **3-2**: HMM parameter tuning for EigenLips

Bearing in mind the results above, the EigenLips methodology is chosen as the appearance descriptor for further experiments. Confusion matrix (Figure 3.15(a)) was computed as the sum of each fold confusion matrix. Class specificity and sensitivity are shown in Figure 3.15(b), mean and standard deviation of each fold are plotted.



(a) Confusion matrix of validation sets for EigenLips representation

(b) Specificity and Sensitivity of each class for EigenLips representation

Figure 3-15: HMM-based classifier performance on validation datasets for EigenLips representation

3.5.2 High Level Features

HMM tuning results for representation based on invariant moments and Active Shape Models are shown in Tables **3-3** and **3-4**, respectively. The former is a geometric-based representation, while the latter is a model-based feature extraction methodology. They both aim to model the shape of the lip region. From the two tested high-level feature extraction methodologies, the best performance was obtained by using ASM. Therefore, the selected shape descriptor employed is next experiments is the Active Shape Model.

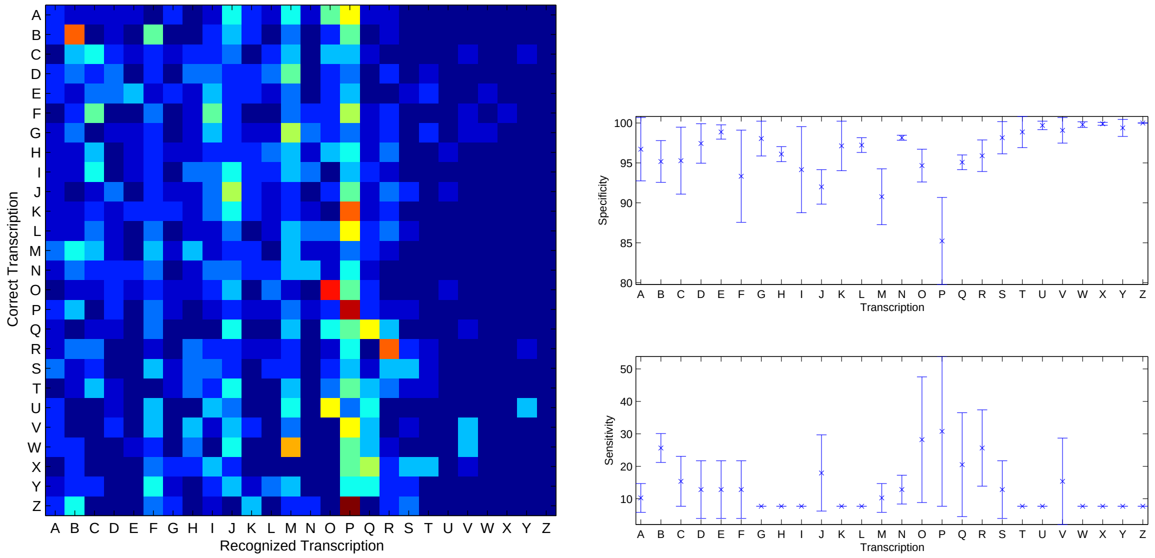
		n_G			
		1	2	4	8
f	$\frac{1}{8}$	10.78 ± 1.73	13.64 ± 1.12	14.43 ± 0.59	13.83 ± 1.07
	$\frac{1}{4}$	12.16 ± 2.10	15.02 ± 0.00	15.12 ± 1.49	14.43 ± 0.78
	$\frac{1}{2}$	10.68 ± 3.02	13.14 ± 0.62	13.83 ± 1.94	13.54 ± 2.23
	1	11.27 ± 2.14	14.62 ± 1.63	15.81 ± 0.90	16.30 ± 0.62

Table **3-3**: HMM parameter tuning for invariant moments

		n_G			
		1	2	4	8
f	$\frac{1}{8}$	18.09 ± 1.33	18.28 ± 3.09	21.74 ± 2.41	20.45 ± 3.25
	$\frac{1}{4}$	19.96 ± 3.77	18.88 ± 1.54	20.06 ± 1.18	19.96 ± 2.41
	$\frac{1}{2}$	19.47 ± 2.31	21.44 ± 4.02	20.55 ± 1.88	19.66 ± 0.45
	1	19.47 ± 2.42	20.36 ± 2.13	19.86 ± 1.23	19.76 ± 0.78

Table **3-4**: HMM parameter tuning for ASM

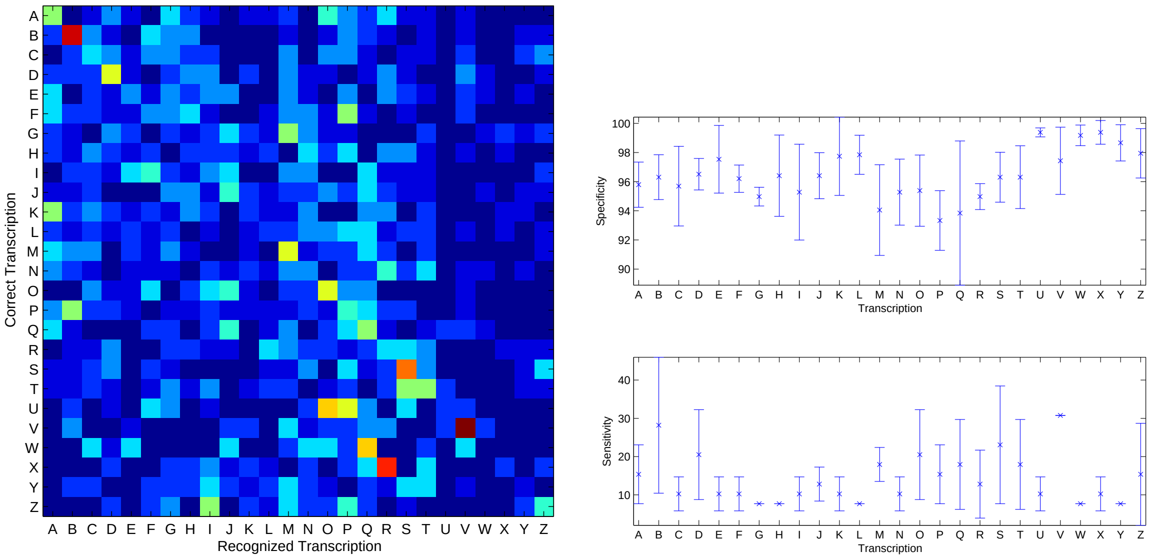
Figures 3.16(a) and 3.17(a) show the confusion matrix obtained from the evaluation of validation samples on the trained classifier, using invariant moment descriptor and ASM features, respectively. Also, average class specificity and sensitivity and its standard deviation are plotted for both representations (Figures 3.16(b) and 3.17(b)).



(a) Confusion matrix of validation sets for invariant moments descriptors

(b) Specificity and Sensitivity of each class for invariant moments descriptors

Figure 3-16: HMM-based classifier performance on validation datasets for invariant moments descriptors



(a) Confusion matrix of validation sets for ASM representation

(b) Specificity and Sensitivity of each class for ASM representation

Figure 3-17: HMM-based classifier performance on validation datasets for ASM representation

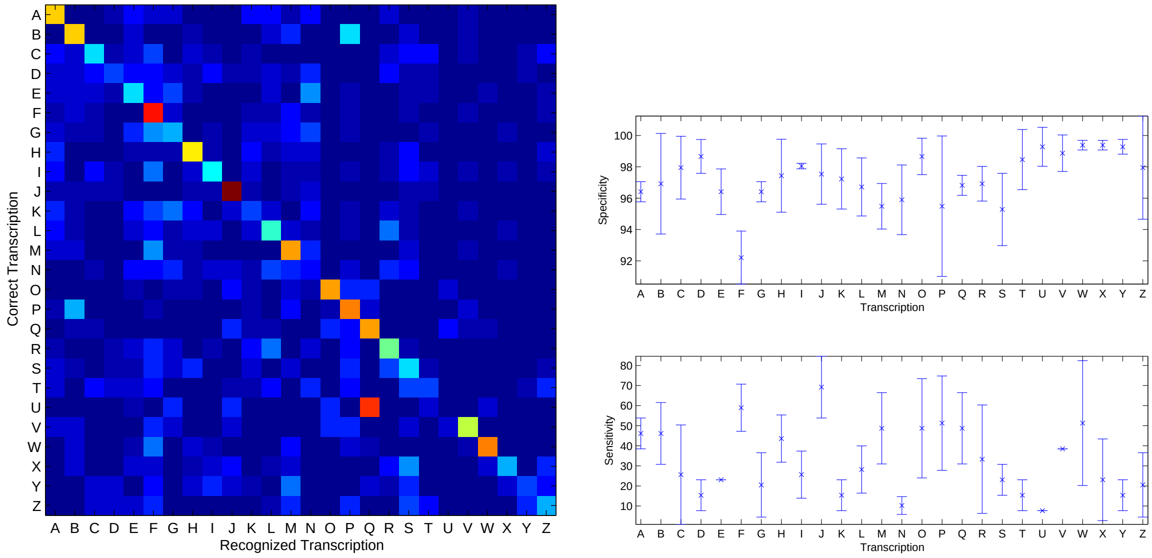
3.5.3 Spatiotemporal Representation

The spatiotemporal representation is built from the appearance and shape descriptors chosen in previous experiments, namely EigenLips and ASM. Additionally, time derivative of the image sequence is included to provide information about mouth dynamics. The mixture of features from different nature can improve system performance since each of them aims to model a different kind of information contained on the mouth and their combination can lead to a better RoI representation.

Classifier accuracy for tested topologies, using the spatiotemporal representation, is shown in Table 3-5. Full set of results for the best configuration is shown in Figure 3-18.

		n_G			
		1	2	4	8
f	$\frac{1}{8}$	21.85 ± 2.37	28.65 ± 2.75	34.57 ± 2.67	31.61 ± 2.68
	$\frac{1}{4}$	21.16 ± 2.42	30.63 ± 4.17	32.99 ± 4.19	31.02 ± 3.70
	$\frac{1}{2}$	24.41 ± 3.98	29.74 ± 1.36	34.36 ± 4.27	30.04 ± 1.36
	1	25.20 ± 0.45	30.92 ± 2.35	31.85 ± 4.04	28.75 ± 3.30

Table 3-5: HMM parameter tuning for spatiotemporal representation



(a) Confusion matrix of validation sets for spatiotemporal representation

(b) Specificity and Sensitivity of each class for spatiotemporal representation

Figure 3-18: HMM-based classifier performance on validation datasets for spatiotemporal representation

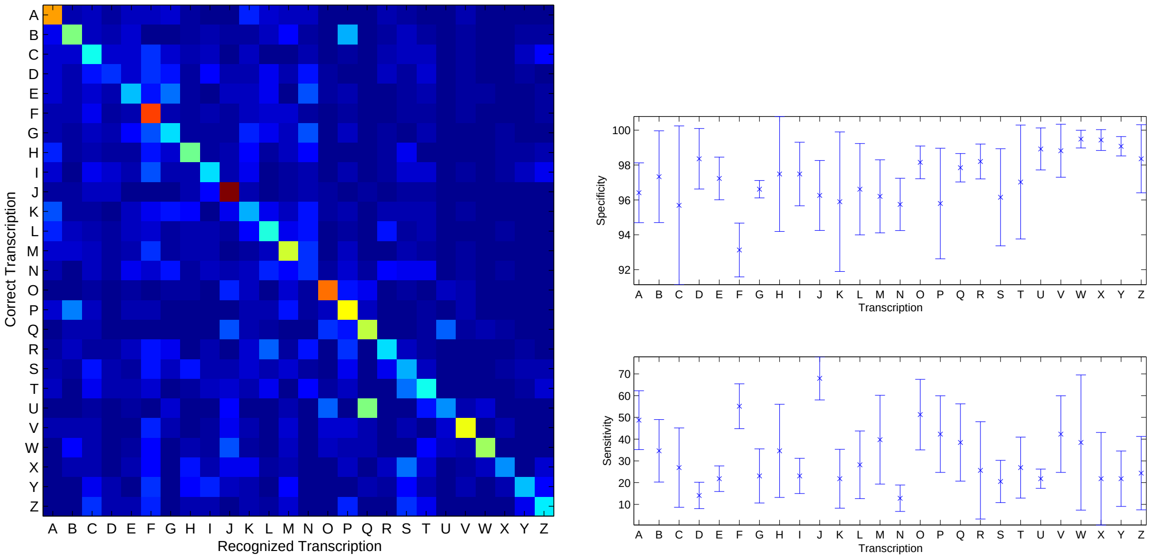
3.5.4 Comparative results

First, the best results from each methodology (Figure **3-6**) are analyzed based on its average accuracy, specificity and sensitivity.

		Accuracy	Specificity	Sensitivity
Feature Set	EigenLips	29.19 ± 3.91	97.16 ± 2.87	30.96 ± 19.54
	ASM	21.74 ± 2.41	96.46 ± 2.40	14.20 ± 9.11
	Spatiotemporal	34.36 ± 4.27	97.27 ± 2.23	32.84 ± 20.99

Table **3-6**: Average performance of visual feature extraction methodologies

As a final result, the best lip representation and the best HMM topology are used to train a classifier using the speaker independent validation strategy, that is, a 14-fold cross-validation methodology. On each repetition a different speaker is leaved out to use it as validation data and the other 13 speakers are used to train the system. Average results of the system performance for the validation samples are shown in Figure **3-19** and a comparison of speaker dependent and speaker independent results is presented in Table **3-7**



(a) Confusion matrix for speaker independent test

(b) Class specificity and sensitivity for speaker independent test

Figure 3-19: Visual speech recognition system performance for speaker independent test

	Accuracy [%]	Specificity [%]	Sensitivity [%]
Speaker Dependent	34.36 ± 4.27	97.27 ± 2.23	32.84 ± 20.99
Speaker Independent	31.21 ± 2.07	97.22 ± 1.46	31.85 ± 13.43

Table **3-7**: Recognition system performance using two different training strategies

4 Discussion

About lip segmentation methodology, the texture-based feature extraction showed an accurate segmentation of facial structures. Since each window is transformed by a matrix estimated from annotated data, the need of a standard color map able to differentiate lip and skin regions, is overcome. Moreover, the use of a k-nn classifier improved classification accuracy. This is due to the better modeling skills on overlapped regions of the k-nn, which made it a proper classifier selection for this task. The main drawback about the use of the k-nn is the time required to perform a single pixel labeling.

About the windowing-based feature extraction, learning curves behavior (Figure **3-5**) shows that an optimal window size can be achieved, although the computational load is increased, due to the initial feature space dimension, that is why a suboptimal size was chosen. Additionally, the overall required segmentation allows the use of a smaller window, specially when accuracy rate for the optimal size is not higher than the chosen size accuracy.

Selection of space reduction methodology for lip segmentation is based on the pixel classification accuracy. Two techniques were tested, PCA and LDA. Results show that LDA (Figure 3.5(b)) have a better performance than PCA (Figure 3.5(a)) on lower dimensions. This fact is supported by the visual inspection of the train set mapping (Figure **3-6**), where classes on LDA representation are less overlapped than in PCA. However, PCA accuracy surpass LDA's not only on higher dimensions but also on larger neighborhood windows. This results make PCA a appropriate dimension reduction strategy for mouth structure segmentation task.

From the two low-level visual features studied, results on Tables **3-1** and**3-2** show that the EigenLips representation outperforms the 2D-DCT representation on the classification task. Since the main objective of using that kind of feature extraction methodologies is the space dimension reduction, it can be said that the EigenLips representation holds more information in a few components than 2D-DCT on the first coefficients. This is because the former space transformation is built from samples of the database and each component is the result of the maximization of the variability contained on the train sets, while the latter transformation is the same for any image. This assumption can be extended for other image transformations such as DWT. Additionally, Figure **3-11** shows that the modes of variation of the PCA version for lip images aim to explain a mouth pose, for example, “closed” and "open" mouth.

On the high level feature extraction methodologies, the best results were obtained for the ASM representation, which is an statistical model of the lip shape. The invariant moments have the lowest performance in terms of accuracy among the feature extraction methodologies tested. Although the moments have important invariance skills, the quality of the lip representation is not enough to perform classification tasks of sequences. It must be beard in mind that, the scale invariance of the moments implies that a shape and its scaled version have the same descriptor vector. But in some cases that is not suitable; for example, the pronunciation of */a/* and */o/* letters have the same shape. This fact can be clearly seen on the confusion matrix in Figure 3.16(a), where no main diagonal pattern can be highlighted, instead, there are a lot samples from all classes labeled as */p/*. About the ASM representation, results show that this shape descriptor can be used effectively for classification task. It must be highlighted that only seven components are used in the feature space.

When comparing results between the chosen low-level (EigenLips) and high-level (ASM) feature extraction methodologies, it can be seen that the former performs significantly better than the latter. The superiority of appearance based features is not surprising, since significant speech information lies within the oral cavity, and such information cannot be captured by the lip contour. Besides, lip contour estimation errors compromise the recognition accuracy.

On the last methodology, a feature set using the best low level and high level features plus the motion information was built. The aim of this combination is to feed the HMM with different kind of information to better model the lip dynamic of each class. Results reported in Table **3-6** show that the spatiotemporal feature set outperforms the other representations. This means that feature sets are complementary. Moreover, the main mistakes reported on the confusion matrix (Figure 3.18(a)) are related to letters belonging to same viseme. For instance, */b/* and */p/* are both produced by a bilabial movement, so it is common to confuse them. A similar analysis can be maid for */q/* and */u/*, which have a similar visual representation. The above introduces the idea of a clustering of the letters of the Spanish alphabet. The main consequence of this fact is the improve of the classifier performance by reducing the number of classes. Moreover, for an online speech recognition application language rules can be included as restriction for the system. That is, a viseme classification task is performed initially, then the most likely letter from the viseme set is chosen depending on the language corpus.

Finally, a speaker independent recognition task was used to test the system. As expected, this test exhibit classification results lower than the speaker dependent task (Table **3-7**). This is mainly because the inner speaker dynamic is not included in the training samples.

However, the system is still able to recognize the general class dynamic.

5 Conclusions

- In this thesis work, an analysis of image feature extraction methodologies was performed based on the quality of the representation and its performance on modeling lip dynamic.

- Multiple image feature extraction methodologies for lip modeling were implemented and tested on a common classification task. Obtained results allowed to sort the methodologies by its lip dynamic modeling skills.

- A visual speech recognition system was designed and tested on two different validation modalities, for speaker dependent and speaker independent speech recognition. As expected, on the former, the system has a better performance than the latter. Nevertheless, results are enough representative to let the system be used as complementary information on multimodal speech recognition tasks.

- Recognition results led the feature set selection to the spatiotemporal-based. This is because this methodology is based on the combination of complementary features, which individually have shown good performance on visual speech modeling.

- Additionally, the proposed preprocessing and lip segmentation stages showed satisfactory results, which eased extraction of lip features and reduced the recognition errors due low quality features.

5.1 Future work

Given that a visual feature extraction analysis was performed on this thesis work, the feature integration of a multimodal speech recognition system can be studied on further works, this stage is nowadays an important research field. Since, the recognition task performed on this work deals with isolated digits recognition, a continuous audio-visual speech recognition system is the next work to deal with aiming to design real-time application. To do it, new stages have to be implemented, namely, the event occurrence detection and segmentation. additionally, the work can be extended to other recognition tasks on other interest fields, such as, anthropometry, to identify reference patterns on visual speech, and medical, for instance, to analyze lip movement of patients with reconstructive surgeries. In that task, an assisted diagnose and therapy tool is useful to speech therapists and surgeons. Finally, the mouth structure segmentation technique can be improved by using faster algorithms to solve the *nearest neighbor* problem, such as kd-tree or quicksort.

Bibliography

[1] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, 91(9): 1306–1326, September 2003. ISSN 0018-9219. doi: 10.1109/JPROC.2003.817150. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1230212`.

[2] W H Sumby and Irwin Pollack. Visual Contribution to Speech Intelligibility in Noise. *Journal of the Acoustical Society of America*, 26 (2):212–215, 1954. ISSN 00014966. doi: 10.1121/1.1907309. URL `http://link.aip.org/link/JASMAN/v26/i2/p212/s1&Agg=doi`.

[3] Jeremy I Skipper, Virginie van Wassenhove, Howard C Nusbaum, and Steven L Small. Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral cortex (New York, N.Y. : 1991)*, 17 (10):2387–99, October 2007. ISSN 1047-3211. doi: 10.1093/cercor/bhl147. URL `http://www.ncbi.nlm.nih.gov/pubmed/17218482`.

[4] Q. Summerfield. Some preliminaries to a comprehensive account of audio-visual speech perception. *Hearing by Eye: The Psychology of Lip-Reading*, pages 3–51, 1987.

[5] E. Petajan. Automatic lipreading to enhance speech recognition. In *Global Telecommunications Conference*, pages 265–272, 1984.

[6] H Cetingul, E Erzin, Y Yemez, and A Tekalp. Multimodal speaker/speech recognition using lip motion, lip texture and audio. *Signal Processing*, 86(12):3549–3558, December 2006. ISSN 01651684. doi: 10.1016/j.sigpro.2006.02.045. URL `http://linkinghub.elsevier.com/retrieve/pii/S0165168406001344`.

[7] S. Gurbuz, Z. Tufekci, E. Patterson, and J.N. Gowdy. Application of affine-invariant Fourier descriptors to lipreading for audio-visual speech recognition. *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, pages 177–180, 2001. doi: 10.1109/ICASSP.2001.940796. URL `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=940796`.

[8] Ara V. Nefian, Luhong Liang, Xiaobo Pi, Xiaoxing Liu, and Kevin Murphy. Dynamic Bayesian networks for audio-visual speech recognition. *EURASIP Journal Ap-*

plications of Signal Processing, (1):1274—-1288, 2002. doi: http://dx.doi.org/10.1155/S1110865702206083. URL http://dx.doi.org/10.1155/S1110865702206083.

[9] G. Potamianos and H.P. Graf. Discriminative training of HMM stream exponents for audio-visual speech recognition. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, volume 1, pages 3733–3736. Ieee, 1998. ISBN 0-7803-4428-6. doi: 10.1109/ICASSP.1998.679695. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=679695.

[10] You Zhang, Stephen Levinson, and Thomas Huang. Speaker independent audio-visual speech recognition. In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, volume 00, pages 1073–1076, 2000. doi: 10.1109/ICME.2000.871546.

[11] C. Bregler and Y. Konig. "Eigenlips" for robust speech recognition. *Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages II/669–II/672, 1994. doi: 10.1109/ICASSP.1994.389567. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=389567.

[12] A. J. Goldschen, O. N. Garcia, and E. D. Petajan. Rationale for phoneme-viseme mapping and feature selection in visual speech recognition. *Speechreading by Humans and Machines, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer-Verlag*, pages 505—-515, 1996.

[13] Martin Heckmann, Frédéric Berthommier, and Kristian Kroschel. Noise Adaptive Stream Weighting in Audio-Visual Speech Recognition. *EURASIP Journal on Advances in Signal Processing*, 2002(11):1260–1273, 2002. ISSN 1687-6172. doi: 10.1155/S1110865702206150. URL http://www.hindawi.com/journals/asp/2002/720764.abs.html.

[14] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of Congnitive Neuroscience*, 3(1):71–86, 1991.

[15] H. Takimoto, Y. Mitsukura, M. Fukumi, and N. Akamatsu. Face detection and emotional extraction system using double structure neural network. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, volume 2, pages 1253–1257. Ieee, 2002. ISBN 0-7803-7898-9. doi: 10.1109/IJCNN.2003.1223873. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1223873.

[16] Pham The Bao, Jin Young Kim, and Seung You Na. Fast multi-face detection in color images using fuzzy logic. In *2005 International Symposium on Intelligent Signal Processing and Communication Systems*, pages 777–780.

Ieee, 2005. ISBN 0-7803-9266-3. doi: 10.1109/ISPACS.2005.1595525. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1595525.

[17] Zhang Jian, M.N. Kaynak, A.D. Cheok, and Ko Chi Chung. Real-time lip tracking for virtual lip implementation in virtual environments and computer games. In *10th IEEE International Conference on Fuzzy Systems. (Cat. No.01CH37297)*, pages 1359–1362. Ieee, 2001. ISBN 0-7803-7293-X. doi: 10.1109/FUZZ.2001.1008910. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1008910.

[18] L E I Xie, Xiu-li Cai, Zhong-wa Fu, Rong-cwn Zhao, and Dong-me Jiang. A robust hierarchical lip tracking approach for lipreading and audio visual speech recognition. In *Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.04EX826)*, number August, pages 3620–3624. Ieee, 2004. ISBN 0-7803-8403-2. doi: 10.1109/ICMLC.2004.1380425. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1380425.

[19] G I Chiou and J N Hwang. Lipreading from color video. In *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, volume 6, pages 1192–5, January 1997. doi: 10.1109/83.605417. URL http://www.ncbi.nlm.nih.gov/pubmed/18283008.

[20] Juergen Luettin. Towards speaker independent continuous speechreading. In *Eurospeech*, pages 1991–1994, 1997.

[21] Pierre Jourlin. WORD-DEPENDENT ACOUSTIC-LABIAL WEIGHTS IN HMM-BASED SPEECH RECOGNITION 1 INTRODUCTION 2 . 1 Position of the Problem 2 ACOUSTIC-LABIAL WEIGHTS 3 . 1 The AMIBE Database. In *AVSP*, pages 69–72, 1997.

[22] Gerasimos Potamianos, Eric Cosatto, Hans Peter Graf, David B Roe, Park Ave, Florham Park, Schulz Drive, and Red Bank. SPEAKER INDEPENDENT AUDIO-VISUAL DATABASE FOR BIMODAL ASR. In *European Tutorial and Research Workshop on Audio-Visual Speech Processing*, pages 65–68, 1997.

[23] Paul Duchnowski, Uwe Meier, and Alex Waibel. See me, hear me: Integrating automatic speech recognition and lip-reading. In *Proc. Int. Conf. Spoken Language Processing*, pages 1–4, 1994.

[24] I. Matthews, J.a. Bangham, and S. Cox. Audiovisual speech recognition using multiscale nonlinear image decomposition. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, pages 38–41. Ieee, 1996. ISBN 0-7803-3555-4. doi: 10.1109/ICSLP.1996.607019. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=607019.

[25] Michael S Gray, Javier R Movellan, and Terrence J Sejnowski. Dynamic features for visual speech- reading : A systematic comparison. *Advances in Neural Information Processing Systems*, pages 751–757, 1997.

[26] Gerasimos Potamianos, Chalapathy Neti, and Iain Matthews. *Audio-Visual Automatic Speech Recognition : An Overview.* 2004.

[27] P. Delmas, P.Y. Coulon, and V. Fristot. Automatic snakes for robust lip boundaries extraction. *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, pages 3069–3072 vol.6, 1999. doi: 10.1109/ICASSP.1999.757489. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=757489.

[28] Juergen Luettin, Neil A Thacker, and Steve W Beet. Active Shape Models for Visual Speech Feature Extraction. *Computer*, (95), 1996.

[29] Shu-hung Leung, Shi-lin Wang, and Wing-hong Lau. Incorporating an Elliptic Shape Function. *IEEE Transactions on Image Processing*, 13(1):51–62, 2004.

[30] S Wang, W Lau, A Liew, and S Leung. Robust lip region segmentation for lip images with complex background. *Pattern Recognition*, 40(12):3481–3491, December 2007. ISSN 00313203. doi: 10.1016/j.patcog.2007.03.016. URL http://linkinghub.elsevier.com/retrieve/pii/S0031320307001446.

[31] R. Rohani, S. Alizadeh, F. Sobhanmanesh, and R. Boostani. Lip segmentation in color images. *2008 International Conference on Innovations in Information Technology*, pages 747–750, December 2008. doi: 10.1109/INNOVATIONS.2008.4781689. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4781689.

[32] Christoph Mayer, Matthias Wimmer, and Bernd Radig. Adjusted pixel features for robust facial component classification. *Image and Vision Computing*, 28(5): 762–771, May 2010. ISSN 02628856. doi: 10.1016/j.imavis.2009.07.012. URL http://linkinghub.elsevier.com/retrieve/pii/S0262885609001565.

[33] Juan-Bernardo Gómez-Mendoza, Flavio Prieto, and Tanneguy Redarce. Automatic Lip-Contour Extraction and Mouth-Structure Segmentation in Images. *Computing in Science & Engineering*, 13(3):22–30, 2011.

[34] Nicolas Eveno, Alice Caplier, and Pierre-Yves Coulon. Jumping Snakes and Parametric Model for Lip Segmentation. In *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, pages 867–870, 2003.

[35] Anya C. Hurlbert and Tomaso A. Poggio. Synthesizing a color algorithm from examples. *Science*, 239:482–485, January 1988. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/3340834.

[36] M. Sadeghi, J. Kittler, and K. Messer. Modelling and segmentation of lip area in face images. *IEE Proceedings - Vision, Image, and Signal Processing*, 149(3):179, 2002. ISSN 1350245X. doi: 10.1049/ip-vis:20020378. URL http://link.aip.org/link/IVIPEK/v149/i3/p179/s1&Agg=doi.

[37] Juan-Bernardo Gómez-Mendoza, Flavio Prieto, and Tanneguy Redarce. Automatic Outer Lip Contour Extraction in Facial Images. *IWSSIP 2010 - 17th International Conference on Systems, Signals and Image Processing*, pages 336–339, 2010.

[38] G. Potamianos, H.P. Graf, and E. Cosatto. An image transform approach for HMM based automatic lipreading. *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No.98CB36269)*, pages 173–177, 1998. doi: 10.1109/ICIP.1998.999008. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=999008.

[39] N. Puviarasan and S. Palanivel. Lip reading of hearing impaired persons using HMM. *Expert Systems with Applications*, 38(4):4477–4481, April 2011. ISSN 09574174. doi: 10.1016/j.eswa.2010.09.119. URL http://linkinghub.elsevier.com/retrieve/pii/S0957417410010766.

[40] Xiaozheng Zhang, Charles C. Broun, Russell M. Mersereau, and Mark a. Clements. Automatic Speechreading with Applications to Human-Computer Interfaces. *EURASIP Journal on Advances in Signal Processing*, 2002(11):1228–1247, 2002. ISSN 1687-6172. doi: 10.1155/S1110865702206137. URL http://www.hindawi.com/journals/asp/2002/240192.abs.html.

[41] Ming-Kuei Hu. Visual Pattern Recognition by Moment Invariants. *IRE Transactions on Information Theory*, pages 66–70, 1962.

[42] Lawrence R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[43] Nobuyuki Otsu. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1): 62–66, 1979. ISSN 0018-9472. doi: 10.1109/TSMC.1979.4310076. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4310076.