EURASIP Journal on Applied Signal Processing 2002:11, 1202–1212


# Statistical Lip-Appearance Models Trained Automatically Using Audio Information

**Philippe Daubias**

*Laboratoire d'Informatique de l'Université du Maine (LIUM), Institut d'Informatique Claude Chappe, F-72085 Le Mans Cedex 9, France*

*Laboratoire d'Informatique Graphique Image et Modélisation (LIGIM), Bâtiment 710, 8, bd Niels Bohr, F-69622 Villeurbanne Cedex, France*
*Email: philippe.daubias@lium.univ-lemans.fr*

**Paul Deléglise**

*Laboratoire d'Informatique de l'Université du Maine (LIUM), Institut d'Informatique Claude Chappe, F-72085 Le Mans Cedex 9, France*
*Email: paul.deleglise@lium.univ-lemans.fr*

*Received 1 November 2001 and in revised form 19 June 2002*

We aim at modeling the appearance of the lower face region to assist visual feature extraction for audio-visual speech processing applications. In this paper, we present a neural network based statistical appearance model of the lips which classifies pixels as belonging to the *lips*, *skin*, or *inner mouth* classes. This model requires labeled examples to be trained, and we propose to label images automatically by employing a lip-shape model and a red-hue energy function. To improve the performance of lip-tracking, we propose to use *blue* marked-up image sequences of the same subject uttering the identical sentences as *natural* nonmarked-up ones. The easily extracted lip shapes from *blue* images are then mapped to the *natural* ones using acoustic information. The lip-shape estimates obtained simplify lip-tracking on the *natural* images, as they reduce the parameter space dimensionality in the red-hue energy minimization, thus yielding better contour shape and location estimates. We applied the proposed method to a small audio-visual database of three subjects, achieving errors in pixel classification around 6%, compared to 3% for hand-placed contours and 20% for filtered red-hue.

**Keywords and phrases:** lip-appearance model, lip-shape model, automatic lip-region labeling, artificial neural networks, dynamic time warping, audio-visual corpora.

## 1. INTRODUCTION

Today, automatic speech recognition (ASR) works well for several applications, but performance depends highly on the specificity of the task, and on the type and level of surrounding noise. To strengthen ASR systems against noise, one may, for example, use multiband systems [1], higher level (linguistic) information [2], or visual information which is complementary to the audio information. Since McGurk's experiments [3] which have proven the importance of visual information in human speech perception, visual modality has been successfully used for improving performance and robustness of ASR [4, 5, 6, 7, 8, 9], speaker recognition [10, 11], and other speech applications [12]. Using visual information in unconstrained conditions requires having accurate visual feature extraction, regardless of the visual features used:

(i) *pixel-based (data-driven) features:* images are fed directly into a speech recognition system [4, 5, 8, 13], after applying a few transformations or normalizations to the images (fixed-size ROI (region of interest) cropping, histogram normalization, for example);

(ii) *model-based features:* a model is located on images, and parameters to be used for ASR are deduced form the location and shape of the model.

In both cases, accurate lip-region detection is required. As this is difficult to achieve in all possible visual conditions, intrusive acquisition devices [14], or a specific blue markup for the subject [6, 10, 15] are sometimes used. We will use the term *blue* images or sequences to refer to images showing such colored lips in the following. On the opposite, images with bare lips, acquired without intrusive devices, will be referred to as *natural* images. In visual speech, most (about 2/3 for French [16]) relevant information is carried by the lips, and visual information may be used in data-driven or model-based ways. It was shown in [17] that data-driven methods lead constantly to higher ASR performance than that for the model-based ones. But even if the

performance is lower, a model of the lips allows easier comparison with other results on audio-visual speech (in production for instance). Most research of the model-based category has only focussed on lip shape. Lip-shape models are often a priori parametric models: a set of parabolic curves [18, 19, 20], elliptic or portions of elliptic curves [19, 21], or geometric templates (polygons) [22]. Revéret and Benoît [23], and Basu et al. [24], use a 3D parametric shape model of the lips. The shape model might also be an a posteriori geometrical template (polygon), with its shape and deformations learned statistically from corpora [7, 11]. Finally, Dupont and Luettin [25], and Matthews et al. [26] also use a lip grey-level appearance model learned from a corpus.

We are also interested in modeling the appearance, but of the whole lower face region. Contrary to other existing approaches relying on red-hue [14, 27, 28, 29], we aim at modeling appearance statistically using feed-forward artificial neural networks (ANNs) also known as multiple layer perceptrons (MLPs). We present here an MLP-based statistical appearance model of the lips which classifies pixels as belonging to the *lips*, *skin*, or *inner mouth* classes. Such an ANN requires labeled examples to be trained and these may only be found on *natural* images. Therefore, training this model requires a precise labeling of such images. One way to obtain examples is to hand-label images, but annotating lips is a hard and time-consuming task for humans. Moreover, using such a statistical appearance model on a large database would quickly make the labeling task become intractable, as shown in [8]: the necessary labeling phase must be partially or fully automated. In this paper, we propose to achieve this, avoiding labor-intensive hand labeling, by employing automatic lip contour tracking based on a lip-shape model and a red-hue energy function. The lip-shape model is a polygon, which represents outer and inner lip contours of a subject, and whose shape and deformations are learned statistically. The energy function uses hue information and a filtering function to be more robust. As nevertheless contour tracking lacks robustness, to improve its performance, we propose to use *blue* image sequences of the same subject uttering the identical sentences as *natural* ones. We then use the acoustic channel of both sequences to perform a dynamic time warping alignment. The easily extracted lip shapes from *blue* images are then mapped to the corresponding *natural* ones using the audio alignment of the two sequences. The obtained lip-shape estimates simplify lip-tracking on the *natural* images, as they reduce the parameter space dimensionality in the red-hue energy minimization, thus yielding better contour shape and location estimates. Such lip contours can then be used to automatically label image blocks as belonging to one of the three classes of interest.

In Section 2, we describe the statistical appearance model of the lower face we intend to use. Then, in Section 4 we explain how to train it automatically using a statistical lip-shape model (described in Section 3) and how to use acoustic information to obtain a more accurate automatic labeling. Finally, detailed results are presented in Section 5.

Figure 1: A sample *natural* image from the corpus: color image (left) and filtered red-hue image (right).

## 2. LIP APPEARANCE MODELING

We intend to model the lower face appearance to assist visual feature extraction. More precisely, we aim at classifying image pixels as belonging to the three classes of interest of this region: *skin*, *lips*, and *inner mouth*. In this section, we first briefly summarize existing approaches, and then propose how to model appearance statistically and how to train such a statistical model.

### 2.1. Literature approaches

Work towards automatic lip segmentation in natural conditions was reported by Coianiz et al. [27], Liévin and Luthon [14], Zhang et al. [29], and Wojdel and Rothkrantz [28]. They all make use of hue information, either filtering it with a parabolic filter [27, 28], or combining it with edge [29] or movement information [14]. Although reported results are always good, the extension to other corpora recorded in different conditions does not seem obvious. We tested the color transformation proposed by [27, 28, 29], but the results obtained were not as good as expected. The images obtained were similar to the one on the right of Figure 1, for which correct lip classification only reaches 80.5%, according to manually labeled contours, whereas non-lip classification is 98.1%. Contrary to others, Liévin uses a logarithmic color transformation to compute hue. This is justified by the low quality of the camera used and the noisy images produced by it. We have also tested this color transformation on our corpus, but the results were not any better. Wojdel [28] also proposes to use a very simple ANN instead of hue. Training is done using a closed mouth image roughly hand-labeled by the user: pixels inside a rectangular area are labeled as lips and the rest as non-lips, and we will discuss this type of model in more detail in the next subsection.

A crucial issue is the portability to different skin colors (for different ethnicities), and red-hue based models may not always be accurate in such contexts. We believe that it is possible to obtain more accurate models relying only on appearance and that the best way to achieve this, is to build, like other researchers [25, 26], an a posteriori statistically learned appearance model of the lips. In the following subsections, we present more precisely the appearance model we wish to build and how to train it.

### 2.2. Statistical modeling of lip appearance

Contrary to previous work [25], where local appearance is used (grey-level profiles perpendicular to the contour at each

contour point), we use global appearance similarly to [30], by applying a single classifier to each image pixel. We consider three classes of interest, and we train the classifier to assign image points into the *lips*, *skin*, or *inner mouth* (teeth, dark area, and tongue) classes. We tried both Gaussian mixture model (GMM) and neural network (NN) classifiers, obtaining very similar results. For example, correct classification is around 87.5% for a mixture of 4 Gaussians per class and 87.2% for a neural network with 10 hidden units.

We have also varied the image block size for classification, and results improved with increased block size, from 87.2% for $1 \times 1$ pixel blocks (as used by Wojdel [28]) to 98% for $5 \times 5$ pixel blocks. Classification results obtained using NNs are somewhat superior to GMMs, when image blocks are used, for example, 2 mixtures per class lead to a 94% recognition accuracy, compared to an NN classifier with 6 hidden units which achieves 96.5% accuracy. Similarly, a 10-mixture per class GMM reaches 97.4%, whereas a 30-hidden unit NN achieves 98.1% recognition and is less computationally expensive to use. There is a certain trade-off between the size of blocks and the number of training samples, and with regard to the size of our images, we chose to use $5 \times 5$ color blocks.

Our chosen appearance model is a three-layer feed-forward ANN. Its entry layer contains 75 units (one for each red, green, and blue value of each $5 \times 5$ block pixel), whereas its output consists of three units, one for each class of interest. The three output values of the network correspond to the probability of the image block to belong to the *lips*, *skin*, or *inner mouth* classes. Between the 75 input and the 3 output units, there exists one layer of hidden units. We have evaluated two different network architectures, one having 10 hidden units, and one having 15. Both networks are fully connected and are trained using back propagation (supervised learning).

### 2.3. Training of the lip appearance model

To train the neural networks, labeled $5 \times 5$ color blocks are required (blocks for which it is known whether they belong to the *skin*, *lips*, or *inner mouth* classes). As a lip contour, once located in an image, indicates the frontier between skin and lips (outer lip contour), and between lips and the inside of the mouth (inner lip contour), it is possible to use an image with its corresponding contour to automatically label the blocks. To obtain these contours, manual labeling may be considered on small data sets, but would become intractable on large databases, as shown in [8]. We assume that automatic labeling is possible, we will explain how it can be achieved later.

For the supervised network training, we will work on a region about 15 pixels wider than the lips on all sides, and centered around them, because this is the region where most of the confusion lies. This region will be scanned pixel by pixel, and for each pixel, we will consider the $5 \times 5$ pixel color image block centered around this position. For training, we will take only homogenous blocks into account (blocks for which all pixels belong to the same class). We will pseudo-randomly select 45 000 blocks (15 000 per subject) of each class and train the networks with them (1000 iterations). In Section 5, the networks obtained will be referred to as auto-10 and auto-15, where the number 10, or 15, corresponds to the number of hidden layer units. We will also train the two neural networks with "certified" blocks. These blocks will be a subset of the previous ones, less subject to mislabeling. As automatic location is less certain than manual location, to obtain this subset, we will reject all blocks that are near the lip contours representing the boundaries between the 3 different classes. In Section 5, these networks will be referred to as cert-10 and cert-15.

In Section 3, we will discuss how to automatically extract lip boundaries on *blue* images, however the algorithm performs poorly on *natural* images. The blue color has high contrast against flesh tones, but without cosmetic assistance, hue information does not allow to separate lips from skin efficiently (see Figure 1), and red-hue alone cannot be used to reliably estimate lip shape. To achieve accurate lip shape and location estimation on *natural* images, we propose to combine red-hue with a lip-shape model. The design and training of this shape model will be described in Section 3.

## 3. LIP SHAPE MODELING

To build a statistical shape model, the most natural way is to hand-label images, as done in [7, 26]. But, even with the help of a specialized tool we designed for it, it is a hard and time-consuming task, especially when building a precise model with numerous points. Time is an important issue considering that the larger the corpus, the better the model is. To build an accurate shape model, an automatic procedure is necessary.

### 3.1. Lip-contour extraction in "blue" images

To automatically build the shape model, we use *blue* video sequences. The blue color of the lipstick corresponds approximately to 220 degrees on the color circle and was chosen to make shape extraction of the lips easier: as blue does not exist in the skin, colored lips present high contrast against the rest, with respect to hue information. Due to variation in lighting recording conditions, on some images, the teeth reflect the blue color from the lips, thus appearing blue. To discriminate the blue lips from the blue teeth, we use saturation information, which corresponds roughly to the purity of the color (the quantity of white added in it). Lips and teeth have high and low saturation values, respectively, and we use a combination of saturation and hue.

To robustly extract the outer and inner lip contours, we first determine a region of interest (ROI). For this, we accumulate the grey-level values of the blue-hue image pixels along vertical and horizontal axes. After calculating the mean grey-level value of all pixels, we choose the ROI where the vertical and horizontal accumulators are lower than the mean value. Starting from this ROI, we seek the outer lip corners (these points are located at the left most and right most parts of the lips in a usual horizontal view), by looking at the image column after column. The center of gravity of all pixels below the mean grey-level value on the left most column is

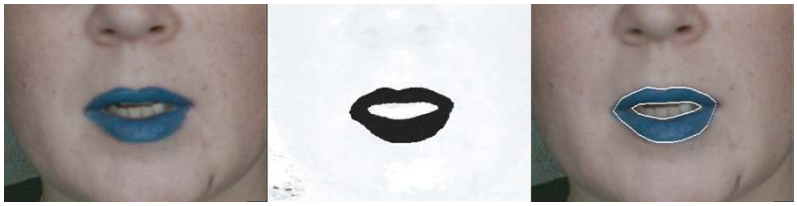

Figure 2: A sample image from the corpus: original image (left), blue-hue image (center), and the contour extracted (right).

considered as the left lip corner. Similarly, the center of gravity of all pixels below the mean grey-level value on the right most column is considered as the right lip corner.

Based upon these lip corners, we extract the outer lip contour by considering the interval between the two lip corners. To get an $n$ point contour, where $n$ is an even integer greater than 2 ($n > 2$), we cut this interval $(n-2)/2$ times and take as contour points the highest and lowest blue-hue points. We then detect whether the lips are open or closed (if there is a non-null inner lip contour). When appropriate, we extract the inner lip corners position and the inner lip contour in the same way. An example of contour extraction on a *blue* image can be seen in Figure 2.

### 3.2. Shape model building

The shape model of the lips is a 44-vertex polygon. More precisely, a 24-vertex polygon describes the outer-lip contour and a 20-vertex polygon describes the inner lip contour. For each picture, we extracted these 44 points, as described in Section 3.1. We then normalized all shapes for position, rotation, and scaling. From the 1190 shapes available in our corpus, we kept only those with a non-null inner contour (618). We then calculated the mean shape $\bar{\mathbf{c}}$ of all contours and subtracted it from all the normalized contours, thus obtaining 88-dimensional contour vectors. Finally, we performed a principal component analysis (PCA) on these vectors as in [7]. As a result, any shape of lips $\mathbf{c}$ found in the corpus can be expressed as

$$\mathbf{c} = \bar{\mathbf{c}} + \mathbf{V} \cdot \mathbf{p}, \tag{1}$$

where $\bar{\mathbf{c}}$ is the mean shape of the lips, $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{88})$ is the matrix for the column eigenvectors, and $\mathbf{p} = (p_1, p_2, \ldots, p_{88})$ is a column vector of dimension 88 containing the weights of each eigenvector to obtain the desired shape. The first coefficients of this vector account for most of the variation, and the projection of the original shape on the first few axes gives a good approximation of the contour. As a result, it is possible to use only the first few eigenvectors to approximate lip shapes. Here, the first 4 vectors represented about 94% of lip deformations, and are shown in Figure 3. Any shape of lips $\mathbf{c}$ can be approximated by

$$\mathbf{c} \approx \bar{\mathbf{c}} + p_1\mathbf{v_1} + p_2\mathbf{v_2} + p_3\mathbf{v_3} + p_4\mathbf{v_4}. \tag{2}$$

### 3.3. Shape model evaluation

To evaluate the lip-shape model, we filled the polygons corresponding to lip contours (examples of filled shapes can be seen in Figure 4) and measured the percentage of overlap between the original and the projected contour areas. Overlap can be defined as the number of intersection pixels divided by the number of union pixels of the two areas, that is,

$$\text{overlap} = \frac{\text{original\_contour} \cap \text{projected\_contour}}{\text{original\_contour} \cup \text{projected\_contour}}. \tag{3}$$

Table 1 shows the evolution of the quality of lip contour approximation, with regard to the number of eigenvectors used. For reference, the first line indicates the PCA global variance percentage. One noticeable point is the weak results obtained in the worst case. For this example, about ten eigenvectors would be required to obtain, for all contours, a shape very close to the extracted one.

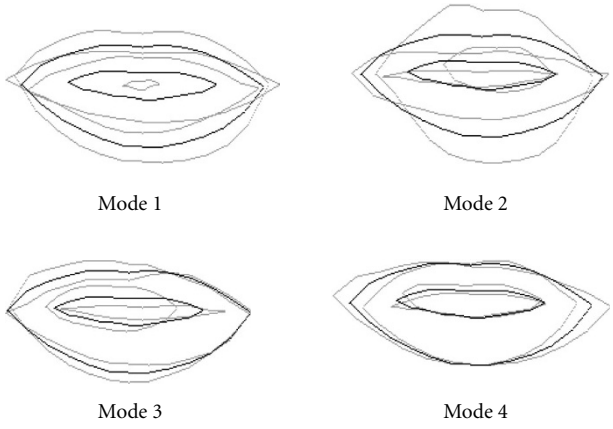


Figure 3: From the first to the fourth mode of deformation: mean shape in solid lines and maximum observed deviation on both sides of the mean (in grey).

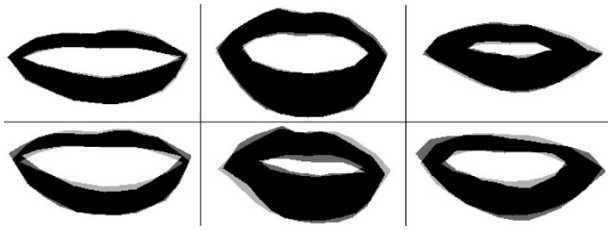

Figure 4: Examples of good (upper row) and bad (lower row) approximations of lip shape for the three subjects considered (2 males and one female (center)) using the first four eigenvectors.

Table 1: Overlap between real lip shapes and projected shapes for different numbers of eigenvectors used (percentages, %).

| Number of vectors | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| PCA variance % | 60.1 | 84.7 | 92.0 | 94.0 | 95.3 | 96.6 | 97.5 |
| Worst case | 37.4 | 54.7 | 55.2 | 60.6 | 60.9 | 62.2 | 80.4 |
| Mean | 75.7 | 84.1 | 86.7 | 87.7 | 89.2 | 90.9 | 92.4 |
| Best case | 90.9 | 93.8 | 95.2 | 96.2 | 96.3 | 97.1 | 97.7 |

Figure 4 shows the overlap between extracted normalized contours and their projection on the first four PCA axes, for three corpus subjects. The extracted contours are filled in dark grey, whereas the projected ones are filled in light grey. Points appearing black are intersection pixels. This figure gives the opportunity to see that the quality of shape approximation varies a lot, but also that a 4-parameter model may adapt to the specific lip shape of different subjects. Such a 4-parameter model is suitable for lip location as will be shown in Section 4.

## 4. LIP CONTOUR LOCATION ON *NATURAL* IMAGES

In Section 3, we have shown how to automatically build a lip-shape model. We will now use it for lip location on *natural* images using first a joint lip-shape and location estimation scheme and then a more efficient cascade lip-shape estimation and location scheme, made possible through the use of complementary acoustic information.

### 4.1. Joint lip-shape and location estimation

Locating the subject's lips in an image with our shape model can be accomplished by applying the right deformation to the mean shape $\bar{\mathbf{c}}$ and then placing the obtained shape at the right position. We must search a 4-dimensional space for lip contour position $(x, y)$, scaling $(r)$, and rotation $(\theta)$, as well as a 4-dimensional space $(p_1, p_2, p_3, p_4)$ for lip shape (see (2)). The best fit of the shape model on the image is obtained by minimizing the $f$ function which measures the distance from a red-centered (about 0 degree on the color circle) hue image (img) to a model image (mod), computed with each set of parameters $(x, y, r, \theta, p_1, p_2, p_3, p_4)$

$$f(x, y, r, \theta, p_1, p_2, p_3, p_4) = \sqrt{\frac{1}{\text{size}}\left[\sum_{i=0}^{\text{size}} (\text{img}_i - \text{mod}_i)^2\right]}, \tag{4}$$

where size is the number of pixels in the image (width $\times$ height), $\text{mod}_i$ and $\text{img}_i$ represent the model and red-hue, respectively for each $i$ pixel and will be detailed next. To reinforce red-hue's discrimination, we use the following robust function, proposed by Odobez [31] in the energy criterion

$$\text{img}_i = \frac{1}{\pi} * \arctan\left[0.4\pi * (\text{hue}_i - \text{hue}_t)\right] + \frac{1}{2}, \tag{5}$$

where $\text{hue}_i$ is the hue value and $\text{img}_i$ the output filtered hue value for each pixel. The transition hue value was set to $\text{hue}_t = 4$ degrees. Model image mod is computed as follows: the mean shape $\bar{\mathbf{c}}$ is deformed into $\mathbf{c}$ with (2), then it is scaled and rotated according to $r$ and $\theta$. The center of gravity of the obtained contour is translated from $(0, 0)$ to $(x, y)$, and this latter contour is drawn, then filled (as in Figure 4) on the mod image. The $\text{mod}_i$ values from each pixel $i$ of this image are

$$\text{mod}_i = \begin{cases} 0 & \text{between outer and inner lip contour,} \\ 1 & \text{elsewhere.} \end{cases} \tag{6}$$

We tested lip location on *natural* images using filtered red-hue, and the downhill simplex method (DSM) [32] as a minimization procedure, but the results were poor. Comparing the contours obtained by this method to the reference hand-labeled ones, the overlap was only about 69.0% on average with a minimum overlap of 31.5%. When looking only at the shapes (not at location), the overlap was on average about 72.2%, with a minimum of 34.2%. By a visual inspection, we noticed that the worst cases were obtained with wrong shapes: the DSM method searches all dimensions simultaneously to find the minimum. It may happen that a wrong and badly placed shape has so little energy that it attracts the simplex into a local minimum.

These poor results do not mean that our shape model is inappropriate, but rather indicate that appearance based on hue information, which we implicitly used in energy minimization, is not sufficient. The dominant skin color of the corpus subjects is rather similar to the dominant lip color, thus red-hue images are very noisy (see Figure 1). As a consequence, there are numerous local minima at which the minimization procedure may stop. To locate lips with our shape model more reliably in unconstrained conditions, a more accurate lip-appearance model like the statistical one presented in Section 2.2 would be necessary. However, to build this appearance model, we need accurate lip location on *natural* images and must improve location in any other way. For this, we propose to estimate lip shape before locating it on the image, instead of trying to jointly estimate shape and location. How this can be achieved will be explained next.

### 4.2. Cascade lip shape and location estimation using acoustic information

As jointly using a lip-shape model and a red-hue appearance function fails, we must add more constraints for location to work accurately. We propose to divide lip location on *natural* images into two subtasks: estimating lip shape and then locating this shape on the image. For that purpose, we have recorded our subjects uttering twice the same sentences: once with blue lipstick and once without, and propose to use acoustic information to align the *natural* sequences on the *blue* ones. This should allow to use the lip shapes extracted on *blue* images to deduce the presumable shape of the lips on *natural* images and to use well-known minimization techniques to locate these shapes on the images.

#### 4.2.1 Use of acoustic information for lip shape estimation

To train the lip-appearance model described in Section 2.2, we need to precisely locate lip boundaries on *natural* images, but want to avoid labor intensive contour hand-labeling. However, our shape model does not work well enough to locate lips on *natural* images accurately. To improve lip contour estimation, we propose to take advantage of speech bimodality: by considering *blue* and *natural* images of the same person uttering identical sounds, we can reliably obtain lip shape information from the *blue* images, and use this to simplify the energy minimization problem (4). We propose to find such image correspondence using dynamic time warping (DTW) on the audio channels of the two sequences. Note that on the literature, the visual channel is mostly used to bring complementary cues when acoustic information is unreliable or missing [12]. By contrast, here we propose to use acoustic information to simplify a visual problem.

DTW is efficient for small vocabulary, isolated word, single speaker automatic speech recognition [33]. It computes the distance from a test word to all words in the vocabulary ($q$ reference words). More precisely, the acoustic signal corresponding to any word is parameterized into acoustic vectors, in this paper, 6 or 12 mel-frequency cepstral coefficients (MFCCs), plus energy. The distances between the vectors from the test word and the ones from each $q$ reference word are computed. This produces $n \times m_k$ cost matrices, $n$ being the number of acoustic vectors associated with the test word, and $m_k$ the number of vectors associated with the $k$th reference word ($1 \leq k \leq q$). The lowest cost path between vector pairs $(1;1)$ and $(n;m_k)$ is computed to obtain the distance from the test word to the $k$th reference word (some constraints are also applied on the path). The lowest cost path computation is repeated $q$ times for each reference word, and the lowest overall distance corresponds to the recognized word. Usually, only this result is used, but DTW brings richer information: the lowest cost path indicates how to best align the vectors from the test and the reference words.

In our case, we use DTW to align the acoustic features of a test sentence, corresponding to a *natural* image sequence, to the acoustic features of a reference sentence, corresponding to a *blue* image sequence. The sentences are pronounced as continuous speech and can be seen as two utterances of the same word, both corresponding to an identical phonemical content uttered by the same subject in very similar conditions. The alignment (path) produced using DTW gives, for any acoustic vector from the test sentence, the corresponding acoustic vector in the reference sentence. Using this association and the synchronicity between acoustic and visual signals, it is possible to deduce to which reference (*blue*) images (and contours) corresponds a test (*natural*) image. Due to the different audio and video frame rates (here, 100, or 200 Hz vs. 25 Hz, respectively), a few (4, or 8) acoustic vectors will correspond to one test image. These acoustic vectors will correspond to a similar number of reference sentence acoustic vectors, which may map to different reference images. Therefore, one test image might be associated with several different contours. Thus, computing which contour corresponds to the test image is not straightforward. We studied three different ways to choose the contour in [34], using either the first, the last, or an interpolation between all possible contours, and the best results were obtained by the latter method.

Two utterances of the same sentence will not always lead to identical lip shapes, especially considering large corpora with different subjects. Nevertheless, this was true to a high degree for any given subject on the corpus used, and the DTW alignment method worked well even with differences in speaking rate. We do not believe that this method can easily be extended to align any speaker with any other, but more experiments should be carried, to see if classes of speakers can be found. Anyway, if only appearance modeling is intended, just a few representative subjects (with different skin colors and facial hair) are needed to build a speaker-independent model, and cross-speaker use of DTW alignment is not necessary.

#### 4.2.2 Lip contour location estimation

Knowing the presumable lip shape corresponding to a picture facilitates the automatic lip extraction process noticeably. We just have to locate a shape on an image instead of estimating a shape before placing it, as it was required in Section 4.1. This reduces the risk of wrong location due to wrong shape, and can be done by minimizing the energy function of Section 4.1. The minimization is now over a 4-dimensional space $(x, y, r, \theta)$ instead of the original 8-dimensional one $(x, y, r, \theta, p_1, p_2, p_3, p_4)$, and is therefore more robust. Once again, we use filtered red hue (as [14, 27, 28, 29]) in the energy function calculation. In fact, we calculate for each color image the corresponding filtered red hue (mod) image (see the right part of Figure 1) with (5) and locate the lip shapes on the latter images. We tried two different types of minimization for locating the lips using:

(i) coordinate pair minimization: combinatory exploration of scale ($r$) and rotation ($\theta$) and, for each value of these parameters, energy function calculation at every possible position $(x, y)$;

(ii) DSM: minimization of all parameters $(x, y, r, \theta)$ by a downhill simplex method [32], initialized at the center of image with a scaling factor of one, and no rotation.

For each case, we kept the best contour found according to the energy function.

## 5. DATABASE AND EXPERIMENTS

In this section, we first describe the corpus used to statistically train and test both parts of the lip model, and then the evaluation principles that we propose to follow. Subsequently, we present our experimental results.

### 5.1. The audio-visual database

Due to the nature of our proposed method, that requires both *blue* and *natural* videos of subjects uttering identical

Table 2: Characteristics from the subset of our corpus.

| | |
|---|---|
| Task | Phonetically balanced sentences in French |
| Image contents | Mouth area and base of nose |
| Subjects | 3 |
| Utterances | 8 per subject (4 *blue*, 4 *natural*) |
| Frames | 2400 images |
| Mouth size | Approximately $200 \times 100$ pixels |
| Lighting | Ambient sun light |

sentences, we had to build our own audio-visual database. In total, we recorded nine subjects (8 males, 1 female), seven of whom were recorded wearing blue lipstick (6 males, 1 female) to allow training and testing of different shape models, and six with no make-up (5 males, 1 female) for training and testing different appearance models. Videos were captured uncompressed at a resolution of $384 \times 288$ pixels in 24 bit RGB color, at a frame rate of 25 noninterleaved images per second. For video acquisition, a PAL analog camcorder was used, connected to a SUN ULTRA2 workstation with an analog card for digitization. Audio was recorded using a standard SUN microphone at a sample rate of 16 kHz with 16 bit samples.

In our experiments, we mainly used three subjects (2 males, 1 female) from the corpus. A brief description of the corresponding corpus subset is shown in Table 2. The subjects were recorded twice on four phonetically balanced sentences in French [35]: first with blue lipstick (as, for example, in [10]), and then without any make-up. These 24 utterances were recorded in natural sun light conditions, which resulted in shadows on some images under the subject's nose and mouth. We have filmed the subjects, placing the lips near the center of the picture at the beginning of the acquisition, but some subjects moved either before or during recording. The camera-subject distance was variable (about one meter), and so was the zooming factor.

### 5.2. The evaluation paradigm

Usually, in automatic speech recognition, only global evaluations which consist of measuring the recognition rate at the system's output are made. Although we have recently published this type of evaluation for our models elsewhere [36], a more detailed evaluation of each step is necessary. Before we report our lip contour estimation results and lip appearance model accuracy on the audio-visual database, we briefly discuss the evaluation paradigm for our algorithms.

#### 5.2.1 Lip contour evaluation

We chose to hand-label a subset of the *natural* images from the corpus, and to use them as a reference for all evaluations. We will compare the automatically extracted contours with the hand-labeled ones, using two metrics:

(i) the average distance (in pixels) between each point from the test contour and the corresponding point in the reference (hand-labeled) contour;
(ii) the percentage of filled shape overlap (see (3)).

#### 5.2.2 Appearance model evaluation

To evaluate the neural network based appearance models, we also use the hand-labeled contours. We want to see if our automatic labeling method allows to build appearance models of the mouth area comparable to manually obtained ones. More precisely, we compare the classification results of the two neural networks described in Section 2.2 after training, as well as of reference networks trained with blocks obtained using hand-labeled contours.

To build these reference models, we need a labeled subset of the corpus: for each of the three subjects considered, we hand-placed the contour on a tenth of the *natural* images (40 out of 400). To compare all models, we use the same images in all cases, and as acoustic alignment is possible only when there is speech, we do not use many closed lips images. The 120 remaining images produce 1.71 million *skin* blocks, 0.72 million *lip* blocks, and 0.11 million *inner mouth* blocks. As in Section 2.3, we retain 45 000 pseudo-randomly chosen blocks of each class to train the networks. The whole set, including the blocks used for training (approximatively 2.5 million blocks), is used for testing all the networks. In Section 5.3.3, the reference networks obtained will be referred to as ref-10 and ref-15.

Finally, we normalized (in position, rotation, and scaling) the hand-labeled contours and used our two energy minimization algorithms (Section 4.2.2) to replace them on their original images. These re-located contours give the opportunity to extract a new set of blocks to train the neural networks. As the contours are the reference ones, they should be 100% accurate in shape and only location may be inaccurate. This allows to see to what extent location and shape estimation are responsible for performance loss of the automatically obtained models over the reference ones. The trained models will be referred to as loc-10 and loc-15 in the results section.

### 5.3. Experimental results

We first present quality of lip-shape estimation results, obtained using DTW for alignment, followed by location results of these shapes on *natural* images. Finally, we describe classification results obtained with the two neural networks (10 and 15 hidden layer units) trained with four different sets of image blocks (hand-labeled blocks (ref), automatically labeled ones (auto), "certified" ones (cert), and relocated hand-labeled ones, (loc)).

#### 5.3.1 Lip-shape estimation

We present here an evaluation of lip shapes obtained using DTW alignment, compared to reference hand-labeled ones. Both contours are normalized before the comparison. Notice that we only compare here the contour shapes, since the quality of location will be evaluated in the next subsection. Table 3 presents the results obtained, depending on the type of acoustic parameters used for DTW-based alignment between the *blue* and *natural* sequence audios. In particular, we vary the acoustic feature extraction rate (100 vs. 200 Hz) and the number of MFCCs used (6 or 12). Both pixel distance and

Table 3: Distance to the reference of the shapes obtained using DTW.

| Acoustic features | | Distance (pixels) | | | Overlap (%) | | |
|---|---|---|---|---|---|---|---|
| Rate | MFCCs | Worst | Mean | Best | Worst | Mean | Best |
| 100 | 12 | 17.06 | 6.67 | 1.81 | 45.63 | 77.90 | 90.01 |
| 100 | 6 | 17.06 | 6.57 | 1.77 | 45.63 | 77.77 | 91.67 |
| 200 | 12 | 18.60 | 6.66 | 1.89 | 44.87 | 77.94 | 91.85 |
| 200 | 6 | 17.06 | 6.57 | 1.93 | 45.63 | 77.96 | 91.01 |

Table 4: Location results by coordinate pair minimization, DSM, and best of both.

| Minimization | Distance (pixels) | | | Overlap (%) | | |
|---|---|---|---|---|---|---|
| Algorithm | Worst | Mean | Best | Worst | Mean | Best |
| Coord. pair min. | 19.46 | 7.07 | 2.93 | 59.58 | 77.98 | 89.73 |
| DSM | 23.86 | 7.41 | 3.24 | 60.23 | 77.72 | 86.94 |
| Best of both | 23.86 | 7.39 | 2.93 | 60.23 | 77.96 | 89.73 |

percentage of overlap between the lip contours are depicted (see also Section 4.2.1).

One may notice that results are very similar, whatever the type of coefficients used, and that they are approximately 78% for overlap and 6.6 pixels for average distance between the two contours. A number of facts influence these results. First, hand labeling of contours is not perfectly accurate, and the shapes may differ even if the alignment is right, as the reference shape is hand-labeled and the estimated shapes are interpolated from automatically extracted shapes. In addition, one of the database subjects painted a little more than his lips with the blue lipstick. As a result, the automatically extracted lip shapes differ in shape from the manually labeled ones. Not using this speaker, results become slightly higher than 80% for overlap and reach an average 6 pixels for distance. Finally, we used relatively high resolution images compared to other existing corpora. A 6 pixel average error must be compared to the 200 pixel average width of the lips.

#### 5.3.2 Quality of location

Table 4 shows the results obtained after energy minimization using the methods described in Section 4.2.2 (coordinate pair minimization, DSM, and best of both, according to the energy criterion used). Notice that the distance results degrade compared to Table 3. Several factors may help explaining this phenomenon: first of all, note that even a one degree variation in rotation alone yields an important error in the results for the distance in pixels between the contours. Also remember that, as explained in the previous subsection, the shapes themselves may differ, which has consequences on both measures. However, contour overlap results remain good, and are almost identical to those of Table 3, which demonstrates that location estimation is successful. As a comparison, the use of our minimization procedures for locating contours extracted on blue-hue images (see Figure 2), and normalized back on their source image, yielded results about 3 pixels in distance and 86% in overlap for coordinate pair minimization and 4.3 pixels and 82%, respectively for DSM.

Table 5: Correct classification rates (%) with each neural network.

| Network | Skin | Inner mouth | Lips | Global |
|---|---|---|---|---|
| ref-10 | 96.91 | 97.86 | 98.48 | 97.40 |
| loc-10 | 95.16 | 98.25 | 97.62 | 96.00 |
| auto-10 | 93.68 | 98.31 | 93.52 | 93.84 |
| cert-10 | 93.53 | 97.19 | 95.65 | 94.30 |
| ref-15 | 97.61 | 97.77 | 98.19 | 97.78 |
| loc-15 | 95.56 | 97.81 | 98.49 | 96.49 |
| auto-15 | 95.39 | 97.17 | 91.07 | 94.24 |
| cert-15 | 94.94 | 97.75 | 94.30 | 94.88 |

#### 5.3.3 Appearance model accuracy

In our work, the main objective is to build an appearance model of the lips, not to evaluate the quality of location of a contour on an image. Nevertheless, the quality of *alignment*, *lip shape*, and *location estimation* may also be evaluated by comparing the classification results obtained with the different ANNs. Classification results are shown in Table 5. They were calculated on a WTA (winner takes it all) basis, which means that for each block entered, the output unit having the highest value was selected as output of the network. These results were computed on a test set of available hand-labeled image blocks in which all classes are not equally represented (67.3% *skin*, 28.4% *lips*, and 4.3% *inner mouth*). The result for all blocks is presented in the Global column. The *lips* output of the network is also presented on the right of Figure 5.

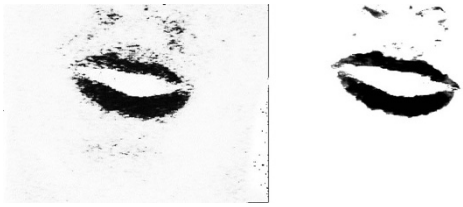

Figure 5: Filtered (with (5)) red-hue image (left) and output of the neural network cert-10 (right).

The correct classification rates obtained are very high (always well over 90%), whatever the ANN used. Table 5 shows a global improvement of classification using "certified" blocks for the training phase. The results obtained with the loc-10 and loc-15 networks seem to indicate that the difference between the reference models (ref-10 and ref-15) and the models built using "certified" blocks (cert-10 and cert-15) is half due to errors in location (Section 4.2.2) and half due to errors of lip-shape estimation. Finally, one can notice that the networks having 15 units in the hidden layer yield slightly higher global results than the ones with 10 units in the hidden layer.

## 6. SUMMARY AND DISCUSSION

In this paper, we have shown how to build statistical shape and appearance models of the lips without hand-labeling. The only drawback of the proposed method is the need for some of the subjects to utter at least part of the training corpus twice (once with a blue lip-enhancer). The shape model is easy to extend to more subjects, but may also be adapted to a specific recognition task (see [36]). The automatic appearance model is rather similar to the one obtained on the same corpus by manually labeling images and it is a lot more efficient than filtered red-hue. The shape and appearance models presented here were used in a parallel and sequential way (appearance, then shape) to compare the efficiency of both approaches for audio-visual ASR in [36].

An important advantage of a statistical model like MLP over a red-hue function, is its adaptation capability: it should be adaptable to all subjects after training. To build a general speaker-independent appearance model, one should introduce in the training corpus subjects with different skin colors and/or facial hair (beard or moustaches), leaving the adaptation work to the ANN. Important attention should then be paid to the architecture of the network, to study more precisely its impact on the adaptation capability.

ANN models can also adapt to previously unencountered data, and they may be used to extract visual speech information on speakers not present in the training corpus. We tested the three subject model described here on a fourth subject (the one studied in [36]), and the preliminary results are rather satisfying, although not entirely: ref-10 reaches 81.9% of correct *lip* classification, with only 63.2% for global classification. When looking more precisely at the confusion matrix, it appears that this is due to the fact that the *inner-mouth* class becomes highly confusable with *lips* and *skin*. For cert-10, results are 75.0% for *lips* and only 58.1% globally, for the same reason. Somewhat surprisingly, networks with 10 hidden units were adapting better to this subject than those with 15 hidden units. However, even if performance is a little lower, the networks still model appearance better than filtered red-hue: considering lip versus non-lip classification, red-hue reaches only 55.6% with the same settings as in Section 2, on the manually labeled images of this subject. More experiments on appearance modeling will be carried in the near future to study cross-speaker adaptation of the models.

Another important issue we have discovered with statistical lip appearance modeling is that pixel information is rather less discriminant than image blocks, and this should be taken into consideration by other researchers in the field. Combining a statistical model like ours with spatial (Zhang [29]) or temporal (Liévin [14]) gradient information should yield a fairly robust appearance model of the lips.

Finally, the use of blue lipstick is not mandatory in our method, as it is not very convenient for speakers and can cause imperfections to the extracted shapes. If one has a very reliable lip-contour extraction system using another intrusive method, it can be used instead. One may, for example, record the first set of sentences with an intrusive device such as a head-mounted camera, or with very intensive frontal lighting to avoid shadows, and then record the second repetition under more realistic conditions.

## ACKNOWLEDGMENTS

The authors would like to thank the EURASIP Journal on Applied Signal Processing reviewers and the editor in charge of the Joint Audio-Visual Speech Processing special issue for their helpful comments in improving this paper.

## REFERENCES

[1] A. Morris, A. Hagen, H. Glotin, and H. Bourlard, "Multi-stream adaptive evidence combination for noise robust ASR," *Speech Communication Journal*, vol. 34, no. 1-2, pp. 25–40, 2001.

[2] I. Zitouni, K. Smaïli, and J.-P. Haton, "Beyond the conventional statistical language models: the variable-length sequences approach," in *Proc. 6th International Conference on Spoken Language Processing (ICSLP)*, vol. 3, pp. 962–965, Beijing, China, October 2000.

[3] H. McGurk and J. McDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, December 1976.

[4] C. Bregler, H. Hild, S. Manke, and A. Waibel, "Improving connected letter recognition by lipreading," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. 557–560, Minneapolis, Minn, USA, April 1993.

[5] P. Duchnowski, M. Hunke, D. Büsching, U. Meier, and A. Waibel, "Toward movement-invariant automatic lip-reading and speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. 109–112, Detroit, Mich, USA, May 1995.

[6] R. Kaucic, B. Dalton, and A. Blake, "Real-time lip tracking for audio-visual speech recognition applications," in *Proc. 4th*

*European Conference on Computer Vision (ECCV)*, vol. 2, pp. 376–387, Cambridge, UK, April 1996.

[7] J. Luettin, N. A. Thacker, and S. W. Beet, "Active shape models for visual speech feature extraction," in *Speechreading by Humans and Machines: Models, Systems, and Applications*, D. G. Stork and M. E. Hennecke, Eds., vol. 150 of *NATO Advanced Science Institutes*, pp. 383–390, Springer-Verlag, New York, NY, USA, 1996.

[8] C. Neti, G. Potamianos, J. Luettin, et al., "Audio-visual speech recognition," Tech. Rep. Workshop 2000, Center for Language and Speech Processing (CLSP), Johns Hopkins University, Baltimore, Md, USA, October 2000.

[9] A. Rogozan and P. Deléglise, "Adaptive fusion of acoustic and visual sources for automatic speech recognition," *Speech Communication Journal*, vol. 26, no. 1-2, pp. 149–161, 1998.

[10] R. Auckenthaler, J. Brand, J. S. D. Mason, F. Deravi, and C. C. Chibelushi, "Lip signatures for automatic person recognition," in *Proc. 2nd International Conference on Audio- and Video-Based Biometric Person Authentification (AVBPA)*, pp. 142–147, Washington, DC, USA, March 1999.

[11] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner, "Acoustic-labial speaker verification," in *Proc. 1st International Conference on Audio- and Video-Based Biometric Person Authentification (AVBPA)*, J. Bigün, G. Chollet, and G. Borgefors, Eds., pp. 319–326, Springer-Verlag, Crans-Montana, Switzerland, March 1997.

[12] L. Girin, L. Varin, G. Feng, and J.-L. Schwartz, "A signal processing system for having the sound "pop-out" in noise thanks to the image of the speaker's lips: New advances using multilayer perceptrons," in *Proc. 5th International Conference on Spoken Language Processing (ICSLP)*, vol. 4, pp. 1451–1454, Sydney, Australia, December 1998.

[13] U. Meier, R. Stiefelhagen, J. Yang, and A. Waibel, "Towards unrestricted lip reading," in *Proc. 2nd International Conference on Multimodal Interfaces (ICMI)*, Hong Kong, 1999.

[14] M. Liévin and F. Luthon, "Unsupervised lip segmentation under natural conditions," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 6, pp. 3065–3068, Phoenix, Ariz, USA, March 1999.

[15] C. Benoît, T. Guiard-Marigny, B. Le Goff, and A. Adjoudani, "Which components of the face do humans and machines best speechread?," in *Speechreading by Humans and Machines: Models, Systems, and Applications*, D. G. Stork and M. E. Hennecke, Eds., vol. 150 of *NATO Advanced Science Institutes*, pp. 315–328, Springer-Verlag, New York, NY, USA, 1996.

[16] C. Benoît, "On the production and perception of audio-visual speech by man and machine," in *Multimedia Communications and Video Coding*, Y. Wang, S. Panwar, S.-P. Kim, and H. L. Bertoni, Eds., Plenum, New York, NY, USA, October 1995.

[17] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," in *Proc. IEEE International Conference on Image Processing (ICIP)*, vol. 3, pp. 173–177, Chicago, Ill, USA, October 1998.

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

[19] S. Horbelt and J.-L. Dugelay, "Active contours for lipreading: combining snakes with templates," in *15th GRETSI Symposium Signal and Image Processing*, pp. 717–720, Juan-les-Pins, France, September 1995.

[20] R. Kober, J. Schiffers, and K. Schmidt, "Model-based versus knowledge-guided representation of non-rigid objects: A case study," in *Proc. IEEE International Conference on Image Processing*, vol. 1, pp. 973–977, IEEE Computer Society Press, Los Alamitos, Calif, USA, 1994.

[21] B. Leroy, I. L. Herlin, and L. D. Cohen, "Face identification by deformation measure," in *Proc. IEEE International Conference on Pattern Recognition (ICPR)*, vol. 3, pp. 633–637, Vienna, Austria, August 1996.

[22] K. F. Lai, C. W. Ngo, and S. Chan, "Tracking of deformable contours by synthesis and match," in *Proc. IEEE International Conference on Pattern Recognition (ICPR)*, vol. 1, pp. 657–661, Vienna, Austria, August 1996.

[23] L. Revéret and C. Benoît, "A new 3D lip model for analysis and synthesis of lip motion in speech production," in *Proc. Auditory-Visual Speech Processing (AVSP)*, pp. 207–212, Terrigal, Australia, December 1998.

[24] S. Basu, N. Oliver, and A. Pentland, "3D modeling and tracking of human lip motion," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 337–343, Bombay, India, January 1998.

[25] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[26] I. Matthews, T. Cootes, S. Cox, R. Harvey, and J. A. Bangham, "Lipreading using shape, shading and scale," in *Proc. Auditory-Visual Speech Processing (AVSP)*, pp. 73–78, Terrigal, Australia, December 1998.

[27] T. Coianiz, L. Torresani, and B. Caprile, "2D deformable models for visual speech analysis," in *Speechreading by Humans and Machines: Models, Systems, and Applications*, D. G. Stork and M. E. Hennecke, Eds., vol. 150 of *NATO Advanced Science Institutes*, pp. 391–398, Springer-Verlag, New York, NY, USA, 1996.

[28] J. C. Wojdel and L. J. M. Rothkrantz, "Using aerial and geometric features in automatic lip-reading," in *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech)*, vol. 4, pp. 2463–2466, Aalborg, Denmark, September 2001.

[29] X. Zhang, C. C. Broun, R. M. Mersereau, and M. Clements, "Automatic speechreading with applications to human-computer interfaces," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 11, pp. 1228–1247, 2002.

[30] R. Kaucic and A. Blake, "Accurate, real-time, unadorned lip tracking," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 370–375, Bombay, India, January 1998.

[31] J.-M. Odobez, *Estimation, détection et segmentation du mouvement: une approche robuste et Markovienne*, Ph.D. thesis, Université de Rennes I, December 1994.

[32] J. A. Nelder and R. Mead, "A simplex method for function minimization," *Computing Journal*, vol. 7, no. 4, pp. 308–313, 1965.

[33] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[34] P. Daubias, "Utilisation de l'information acoustique pour aligner deux séquences de parole audiovisuelle," in *Proc. 4th Rencontres Jeunes Chercheurs en Parole*, pp. 74–77, Mons, Belgium, September 2001.

[35] R. Descout, J.-F. Sérignat, O. Cervantes, and R. Carré, "BD-SONS: Une base de donnés des sons du français," in *Proc. 12th International Congress on Acoustics (ICA)*, Toronto, Canada, 1986.

[36] P. Daubias and P. Deléglise, "Lip-reading based on a fully automatic statistical model," in *Proc. 7th International Conference on Spoken Language Processing (ICSLP)*, vol. 1, pp. 209–212, Denver, Col, USA, September 2002.

**Philippe Daubias** was born in Le Mans (France) in 1972. He completed his Master's thesis in computer science in Université du Maine, France in 1996. He is now finishing his Ph.D. on visual feature extraction for audio-visual automatic speech recognition. His research interests include multimedia processing and statistical modeling.

**Paul Deléglise** received his Ph.D. in computer science from the Pierre & Marie Curie University (Paris VI, France) in 1983 and his Doctorat D'état in 1991. He worked in the Signal Laboratory of the École Nationale Supérieure des Télécommunications (ENST) on automatic speech recognition from 1985 to 1992. Since October 1992, he is full professor at Université du Maine (France) where he works in the LIUM Laboratory on data fusion applied to audio-visual speech recognition.