EURASIP Journal on Applied Signal Processing 2002:11, 1228–1247

Automatic Speechreading with Applications to Human-Computer Interfaces

Xiaozheng Zhang
Center for Signal and Image Processing, Georgia Institute of Technology, Atlanta, GA 30332-0250, USA
Email: xzhang@ee.gatech.edu

Charles C. Broun
Motorola Human Interface Lab, Tempe, AZ 85284, USA
Email: charles.broun@motorola.com

Russell M. Mersereau
Center for Signal and Image Processing, Georgia Institute of Technology, Atlanta, GA 30332-0250, USA
Email: rmm@ee.gatech.edu

Mark A. Clements
Center for Signal and Image Processing, Georgia Institute of Technology, Atlanta, GA 30332-0250, USA
Email: clements@ee.gatech.edu

Received 30 October 2001 and in revised form 31 July 2002

There has been growing interest in introducing speech as a new modality into the human-computer interface (HCI). Motivated by the multimodal nature of speech, the visual component is considered to yield information that is not always present in the acoustic signal and enables improved system performance over acoustic-only methods, especially in noisy environments. In this paper, we investigate the usefulness of visual speech information in HCI related applications. We first introduce a new algorithm for automatically locating the mouth region by using color and motion information and segmenting the lip region by making use of both color and edge information based on Markov random fields. We then derive a relevant set of visual speech parameters and incorporate them into a recognition engine. We present various visual feature performance comparisons to explore their impact on the recognition accuracy, including the lip inner contour and the visibility of the tongue and teeth. By using a common visual feature set, we demonstrate two applications that exploit speechreading in a joint audio-visual speech signal processing task: speech recognition and speaker verification. The experimental results based on two databases demonstrate that the visual information is highly effective for improving recognition performance over a variety of acoustic noise levels.

Keywords and phrases: automatic speechreading, visual feature extraction, Markov random fields, hidden Markov models, polynomial classifier, speech recognition, speaker verification.

1. INTRODUCTION

In recent years there has been growing interest in introducing new modalities into human-computer interfaces (HCIs). Natural means of communicating between humans and computers using speech instead of a mouse and keyboard provide an attractive alternative for HCI.

With this motivation much research has been carried out in automatic speech recognition (ASR). Mainstream speech recognition has focused almost exclusively on the acoustic signal. Although purely acoustic-based ASR systems yield excellent results in the laboratory environment, the recognition error can increase dramatically in the real world in the presence of noise such as in a typical office environment with ringing telephones and noise from fans and human conversations. Noise robust methods using feature-normalization algorithms, microphone arrays, representations based on human hearing, and other approaches [1, 2, 3] have limited success. Besides, multiple speakers are very hard to separate acoustically [4].

To overcome this limitation, automatic speechreading systems, through their use of visual information to augment acoustic information, have been considered. This is motivated by the ability of hearing-impaired people to lipread. Most human listeners who are not hearing impaired also make use of visual information to improve speech perception

especially in acoustically hostile environments. In human speechreading, many of the sounds that tend to be difficult for people to distinguish orally are easier to see (e.g., /p/, /t/, /k/), and those sounds that are more difficult to distinguish visually are easier to hear (e.g., /p/, /b/, /m/). Therefore, visual and audio information can be considered to be complementary to each other [5, 6].

The first automatic speechreading system was developed by Petajan in 1984 [7]. He showed that an audio-visual system outperforms either modality alone. During the following years various automatic speechreading systems were developed [8, 9] that demonstrated that visual speech yields information that is not always present in the acoustic signal and enables improved recognition accuracy over audio-only ASR systems, especially in environments corrupted by acoustic noise and multiple talkers. The two modalities serve complementary functions in speechreading. While the audio speech signal is represented by the acoustic waveform, the visual speech signal usually refers to the accompanying lip movement, tongue and teeth visibility, and other relevant facial features.

An area related to HCI is personal authentication. The traditional way of using a password and PIN is cumbersome since they are difficult to remember, must be changed frequently, and are subject to "tampering." One solution is the use of biometrics, such as voice, which have the advantage of requiring little custom hardware and are nonintrusive. However, there are two significant problems in current generation speaker verification systems using speech. One is the difficulty in acquiring clean audio signals in an unconstrained environment. The other is that unimodal biometric models do not always work well for a certain group of the population. To combat these issues, systems incorporating the visual modality are being investigated to improve system robustness to environmental conditions, as well as to improve overall accuracy across the population. Face recognition has been an active research area during the past few years [10, 11]. However, face recognition is often based on static face images assuming a neutral facial expression and requires that the speaker does not have significant appearance changes. Lip movement is a natural by-product of speech production, and it does not only reflect speaker-dependent static and dynamic features, but also provides "liveness" testing (in case an imposter attempts to fool the system by using the photograph of a client or pre-recorded speech). Previous work on speaker recognition using visual lip features includes the studies in [12, 13].

To summarize, speech is an attractive means for a user-friendly human-computer interface. Speech not only conveys the linguistic information, but also characterizes the talker's identity. Therefore, it can be used for both speech and speaker recognition tasks. While most of the speech information is contained in the acoustic channel, the lip movement during speech production also provides useful information. These two modalities have different strengths and weaknesses and to a large extent they complement each other. By incorporating visual speech information we can improve a purely acoustic-based system.

To enable a computer to perform speechreading or speaker identification, two issues need to be addressed. First, an accurate and robust visual speech feature extraction algorithm needs to be designed. Second, effective strategies to integrate the two separate information sources need to be developed. In this paper, we will examine both these aspects.

We report an algorithm developed to extract visual speech features. The algorithm consists of two stages of visual analysis: lip region detection and lip segmentation. In the lip region detection stage, the speaker's mouth in the video sequence is located based on color and motion information. The lip segmentation phase segments the lip region from its surroundings by making use of both color and edge information, combined within a Markov random field framework. The key locations that define the lip position are detected and a relevant set of visual speech parameters are derived. By enabling extraction of an expanded set of visual speech features, including the lip inner contour and the visibility of the tongue and teeth, this visual front end achieves an increased accuracy in an ASR task when compared with previous approaches. Besides ASR, it is also demonstrated that the visual speech information is highly effective over acoustic information alone in a speaker verification task.

This paper is organized as follows. Section 2 gives a review of previous work on extraction of visual speech features. We point out advantages and drawbacks of the various approaches and illuminate the direction of our work. Section 3 presents our visual front end for lip feature extraction. The problems of speech and speaker recognition using visual and audio speech features are examined in Sections 4 and 5, respectively. Finally, Section 6 offers our conclusions.

2. PREVIOUS WORK ON VISUAL FEATURE EXTRACTION

It is generally agreed that most visual speech information is contained in the lips. Thus, visual analysis mainly focuses on lip feature extraction. The choice for a visual representation of lip movement has led to different approaches. At one extreme, the entire image of the talker's mouth is used as a feature. With other approaches, only a small set of parameters describing the relevant information of the lip movement is used.

In the image-based approach, the whole image containing the mouth area is used as a feature either directly [14, 15], or after some preprocessing such as a principal components analysis [16] or vector quantization [17]. Recently, more sophisticated data preprocessing has been used, such as a linear discriminant analysis projection and maximum likelihood linear transform feature rotation [18]. The advantage of the image-based approach is that no information is lost, but it is left to the recognition engine to determine the relevant features in the image. A common criticism of this approach is that it tends to be very sensitive to changes in illumination, position, camera distance, rotation, and speaker [17].

Contrary to the image-based approach, others aim at explicitly extracting relevant visual speech features. For example in [19], descriptors of the mouth derived from optical

flow data were used as visual features. In [20], oral cavity features including width, height, area, perimeter, and their ratios and derivatives were used as inputs for the recognizer.

In more standard approaches, model-based methods are considered, where a geometric model of the lip contour is applied. The typical examples are deformable templates [21], "snakes" [22], and active shape models [23]. Either the model parameters or the geometric features derived from the shape such as the height and width of the mouth are used as features for recognition. For all three approaches, an image search is performed by fitting a model to the edges of the lips, where intensity values are commonly used. The difficulty with these approaches usually arises when the contrast is poor along the lip contours, which occurs quite often under normal lighting conditions. In particular, edges on the lower lip are hard to distinguish because of shading and reflection. The algorithm is usually hard to extend to various lighting conditions, people with different skin colors, or people with facial hair. In addition, the teeth and tongue are not easy to detect using intensity-only information. The skin-lip and lip-teeth edges are highly confusable.

An obvious way of overcoming the inherent limitations of the intensity-based approach is to use color, which can greatly simplify lip identification and extraction. Lip feature extraction using color information has gained interest with the increasing processing power and storage of hardware making color image analysis more affordable. However, certain restrictions and assumptions are required in existing systems. They either require individual chromaticity models [24], or manually determined lookup tables [25]. More importantly, most of the methods only extract outer lip contours [26, 27]. No methods have been able to explicitly detect the visibility of the tongue and teeth so far.

Human perceptual studies [28, 29] show that more visual speech information is contained within the inner lip contours. The visibility of the teeth and tongue inside the mouth is also important to human lipreaders [30, 31, 32]. We, therefore, aim to extract both outer and inner lip parameters, as well as to detect the presence/absence of the teeth and tongue.

One of the major challenges of any lip tracking system is its robustness over a large sample of the population. We include two databases in our study. One is the audio-visual database from Carnegie Mellon University [33, 34] including ten test subjects, the other is the XM2VTS database [35, 36], which includes 295 test subjects. In the next section, we present an approach that extracts geometric lip features using color video sequences.

3. VISUAL SPEECH FEATURE EXTRACTION

3.1. Lip region/feature detection

3.1.1 Color analysis

The RGB color model is most widely used in computer vision because color CRTs use red, green, and blue phosphors to create the desired color. However, its inability to separate the luminance and chromatic components of a color hinders the effectiveness of color in image recognition. Previous studies

(a) (b)

Figure 1: (a) Original image. (b) Manually extracted lip ROI.

[37, 38] have shown that even though different people have different skin colors, the major difference lies in the intensity rather than the color itself. To separate the chromatic and luminance components, various transformed color spaces can be employed, such as the normalized RGB space (which we denote as rgb in the following), YCbCr, and HSV. Many transformations from RGB to HSV are presented in the literature. Here the transformation is implemented after [39].

The choice of an appropriate color space is of great importance for successful feature extraction. To analyze the statistics of each color model, we build histograms of the three color components in each color space by discretizing the image colors and counting the number of times each discrete color occurs in the image. We construct histograms for the entire image and for the manually extracted lip regions of interest (ROI) bounded within the contour, as shown in Figure 1.

Typical histograms of the color components in the RGB, rgb, HSV, and YCbCr color spaces are shown in Figures 2, 3, 4, and 5, where two cases are given: (a) those for the entire image and (b) those for the extracted lip region only.

Based on the histograms obtained from video sequences taken under various test conditions and for different test subjects, we can make the following observations. (i) The color components (r, g, b), (Cb, Cr), and (H) exhibit peaks in the histograms of the lip region. This indicates that the color distribution of the lip region is narrow and implies that the color for the lip region is fairly uniform. On the other hand, color distributions of the R/G/B components (Figure 2) are wide spread since they contain luminance components. The RGB color space is therefore not suitable for object identification and is discarded in the following analysis. (ii) The color histogram of (r, g, b) and (Cb, Cr) in the lip region more or less overlaps with that of the whole image (Figures 3 and 5), while the hue component has the least similarity between the entire image and the isolated lip region (Figure 4). This shows that hue has high discriminative power. (iii) The distributions of (r, g, b) and (Cb, Cr) vary for different test subjects, while hue is relatively constant under varying lighting conditions, and for different speakers. We therefore conclude that hue is an appropriate measure for our application.

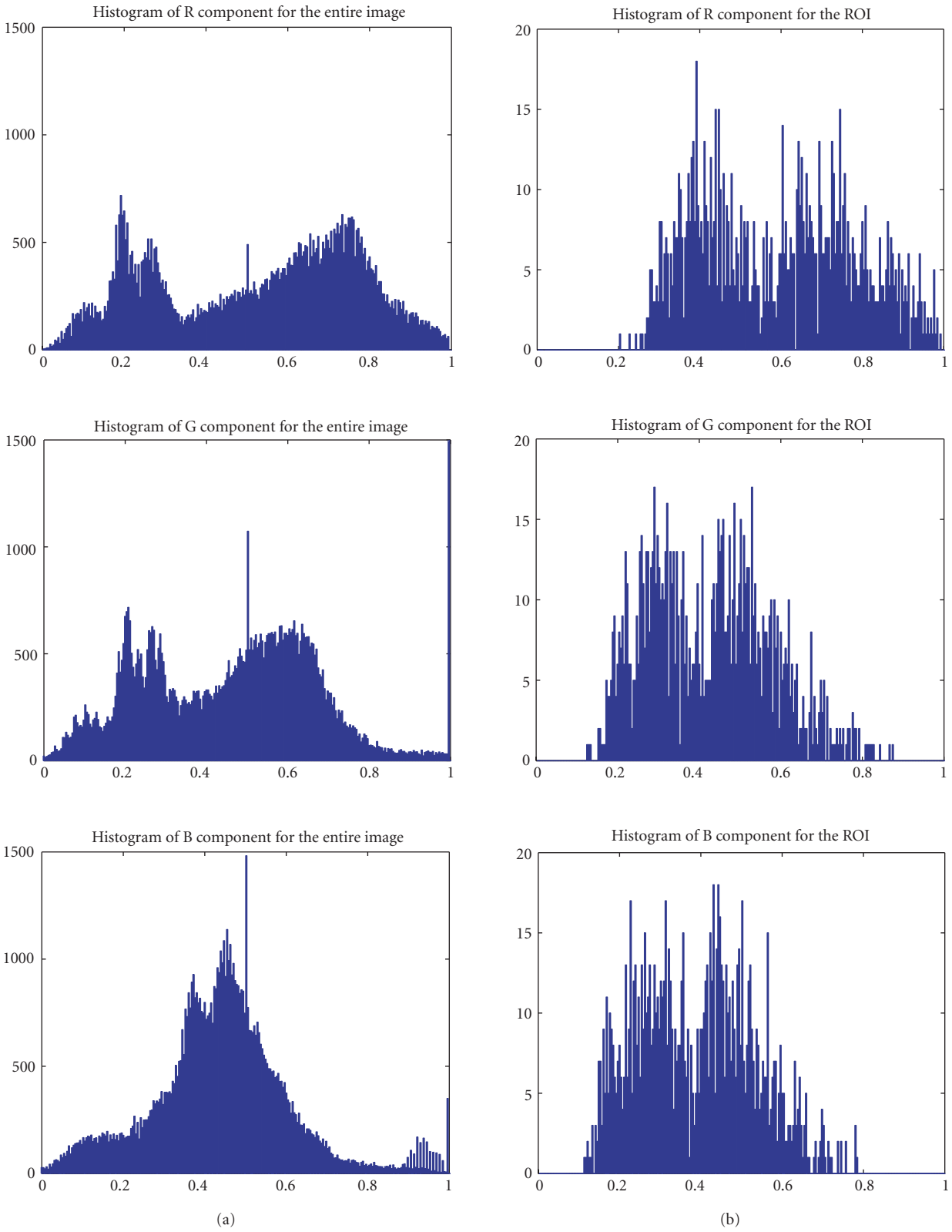


Figure 2: Histograms of R/G/B components. (a) Entire image. (b) Lip ROI.

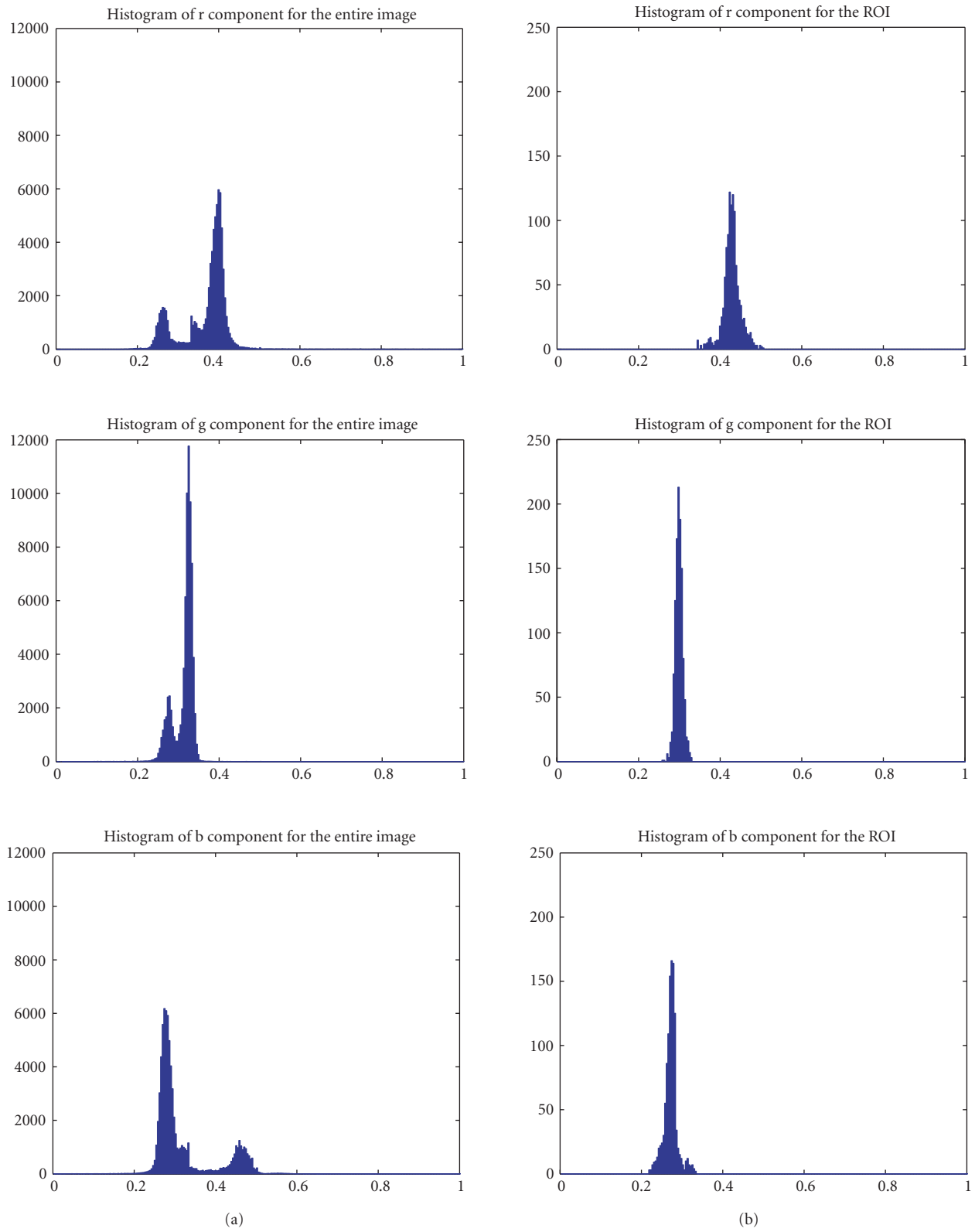


Figure 3: Histograms of r/g/b components. (a) Entire image. (b) Lip ROI.

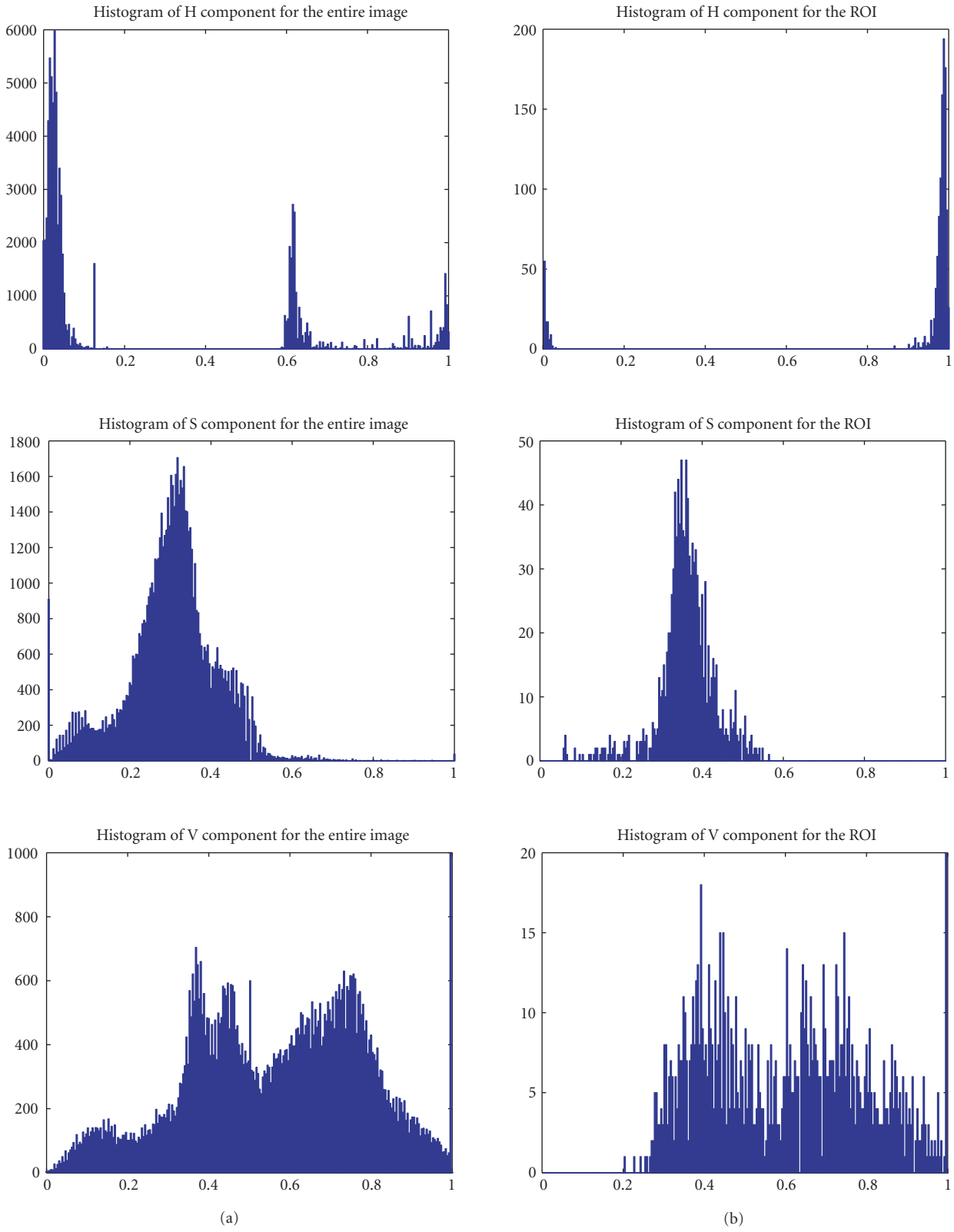


Figure 4: Histograms of H/S/V components. (a) Entire image. (b) Lip ROI.

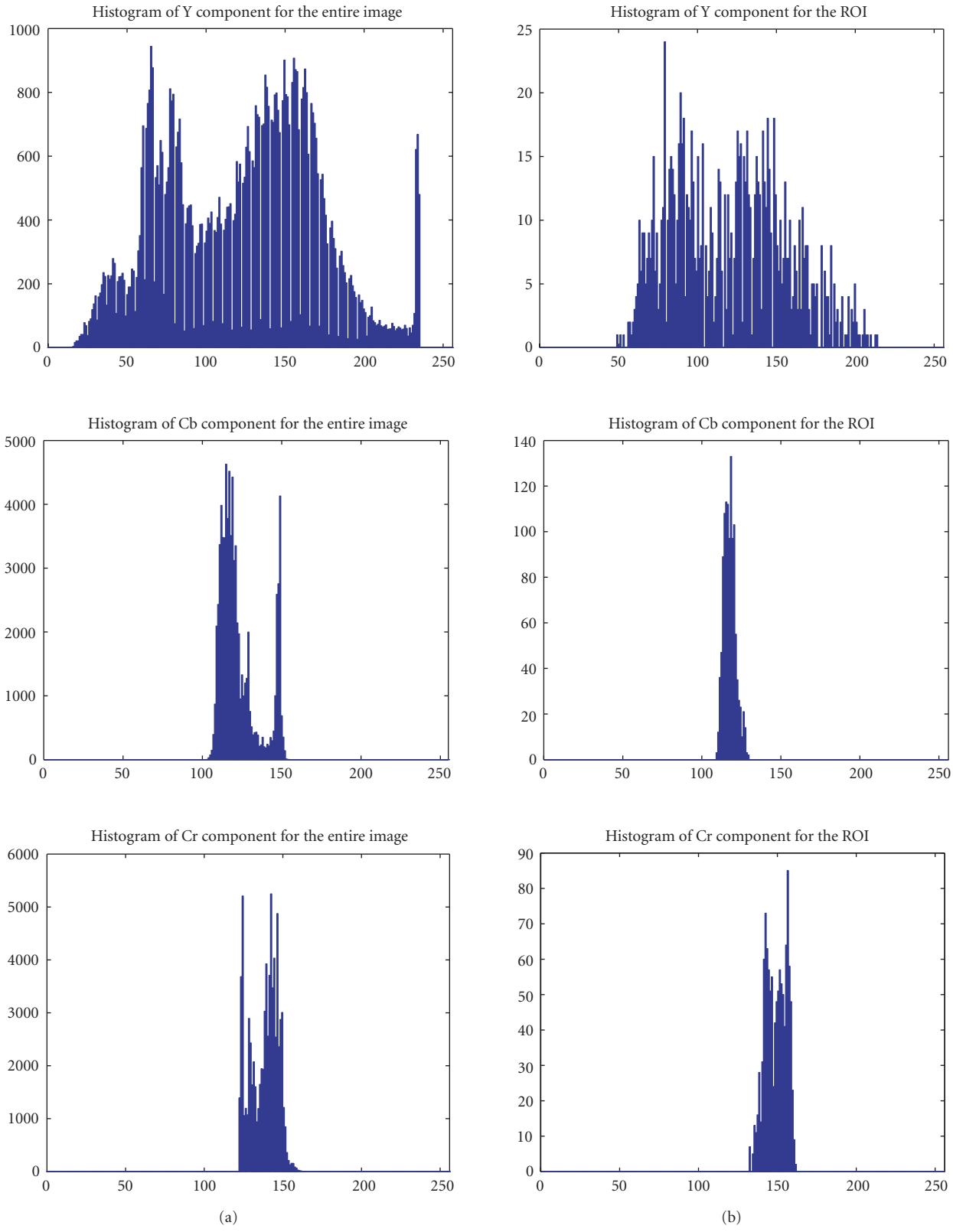


Figure 5: Histograms of Y/Cb/Cr components. (a) Entire image. (b) Lip ROI.

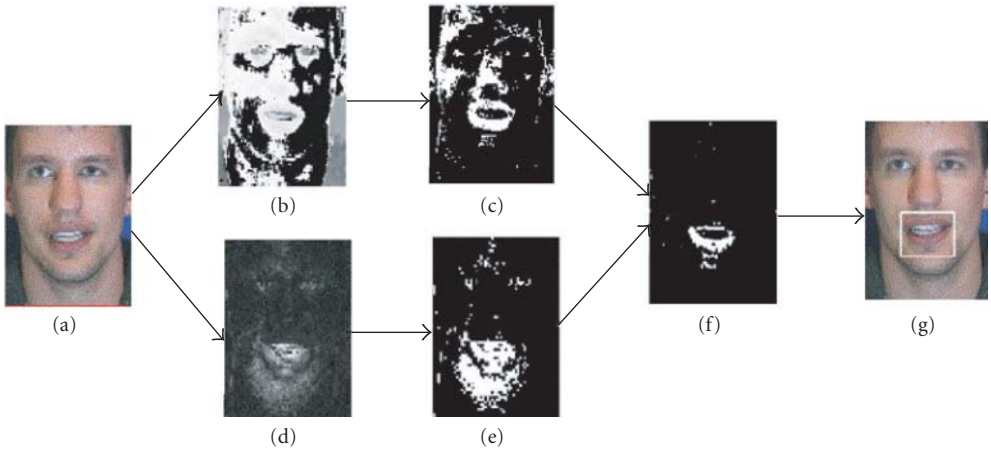


Figure 6: Lip region detection. (a) Gray level representation of the original RGB color image. (b) Hue image. (c) Binary image after H/S thresholding. (d) Accumulated difference image. (e) Binary image (d) after thresholding. (f) Result from AND operation on (c) and (e). (g) Original image with the identified lip region.

The first figure in Figure 4b shows the histogram of hue for the lip region. We observe that the red hue falls into two separate subsets at the low and high ends of the whole color range, as a result of the wrap-around nature of hue (hue is defined on a ring). For easy use of the hue component, we rotate the hue to the left, so that the red color falls in a connected region that lies at the high end of the hue range close to 1 (we scale the hue by a factor of 360 so that it is defined over the range $[0, 1]$). The modified RGB to HSV conversion is shown in the following:

$M = \max(\mathrm{R}, \mathrm{G}, \mathrm{B})$
$m = \min(\mathrm{R}, \mathrm{G}, \mathrm{B})$
$d = M - m$
Value calculation: $V = M$
Saturation calculation: $S = (M == 0)?0 : d/M$
Hue calculation:
if $(S == 0)$
    $H = 0$
else
    if $(d == 0)$ $d = 1$
    $H = (\mathrm{R} == M)?((\mathrm{G}-\mathrm{B})/d) : (\mathrm{G} == M)?(2+(\mathrm{B}-\mathrm{R})/d) : (4 + (\mathrm{R} - \mathrm{G})/d)$
    $H- = .2$
    $H/ = 6$
    if $(H < 0)$ $H+ = 1$

3.1.2 Lip region detection

The problem of visual feature extraction consists of two parts: lip region detection and lip feature extraction. In the first stage of the visual analysis, the speaker's mouth in the video sequence is located. We utilize hue for this purpose. Given an RGB image of the frontal view of a talker, as shown in Figure 6a, a modified hue color image can be derived (Figure 6b). Since the modified red hue value lies at the high end, the lips appear to be the brightest region, but there is considerable noise in the hue image. Part of the noise is related to the unfortunate singularity property of RGB to HSV conversion, which occurs when R = G = B (saturation = 0) [40]. To remove this type of noise, we require that S exceed a certain preset value. For segmenting the lips, we label a pixel as a lip pixel if and only if $H(i, j) > H_0$, $S(i, j) > S_0$, where $H_0 = 0.8$, $S_0 = 0.25$ for $H, S \in [0, 1]$. The accuracies of those two values are not very critical, and they proved to generalize well for other talkers. The resulting binary image is shown in Figure 6c.

Another component of the noise is caused by the non-lip red blobs in the image, for example when there are distracting red blobs in the clothing, or if the person has a ruddy complexion, as is the case for the person shown in Figure 6. In this case, we exploit motion cues to increase the robustness of detecting the lips. In this approach, we search for the moving lips in the image if an audio signal is present in the acoustic channel. To detect the moving object, we build difference images between subsequent frames and sum over a series of frames. The accumulated difference image (ADI) is defined as follows:

$$\begin{aligned} &\mathrm{ADI}_0(i, j) = 0, \\ &\mathrm{ADI}_k(i, j) = \mathrm{ADI}_{k-1}(i, j) + \Delta R_k(i, j), \quad k \in 1, \ldots, T, \end{aligned} \tag{1}$$

where the difference image $\Delta R_k(i, j)$ is calculated by pixel-wise absolute subtraction between adjacent frames $\Delta R_k(i, j) = |R_k(i, j) - R_{k-1}(i, j)|$. Note that we use the R component for our lip detection. T is set to 100 in our work, that is, we sum the difference images over 100 frames. An example of an accumulated difference image is shown in Figure 6d. To separate the moving lips from the background, we use two subsequent thresholding operations. The first threshold is applied to the entire image, where threshold t_1 is derived by using Otsu's method [41]. This operation separates the speaker from the background. A subsequent threshold is then applied to the image with all pixel values $> t_1$, and $t_2 > t_1$ is

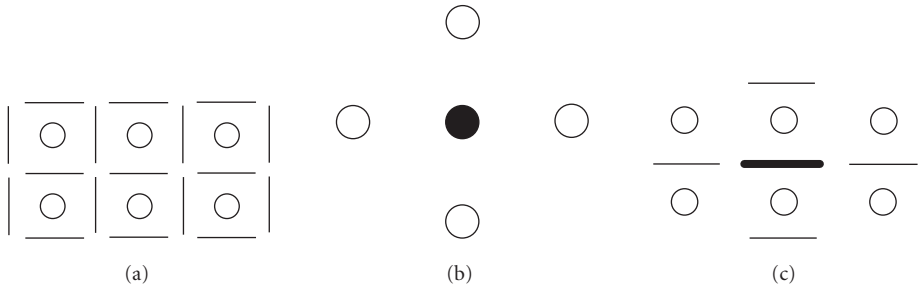


Figure 7: (a) Configuration of lip sites (∘) and edge sites (—). (b) Neighborhood system for lip sites. The filled circle represents the site, and unfilled circles represent the neighbors of the site. (c) Neighborhood system for horizontal edge sites. The thick line represents the site, and thin lines represent the neighbors of the site.

derived. The binary image based on t_2 is shown in Figure 6e with moving mouth being highlighted. When this is combined with the binary image from the hue/saturation thresholding, shown in Figure 6c, the binary image, Figure 6f, is obtained by combining the two binary images using an AND operation. Based on the resulting image, we extract the lip region from its surroundings by finding the largest connected region. The identified lip area is shown as a white bounding box in Figure 6g.

There exist many other sophisticated classifiers in the literature such as in [42, 43]. The effectiveness of this rather simple algorithm lies in the fact that the hue color is very efficient in identifying the lips due to its color constancy and high discriminative power. It should be noted, however, that it is assumed here that the video sequence contains the frontal view of a speaker without significant head motion.

The lip location algorithm described above needs to be done only once for the first image of the sequence. For the succeeding frames, we estimate the lip region from the detected lip features of the previous frame based on the assumption that the mouth does not move abruptly from frame to frame. Subsequent processing is restricted to the identified lip region.

3.1.3 MRF-based lip segmentation

Since hue in [39] is defined on a ring (see Section 3.1.1) rather than on an interval **R**, standard arithmetic operations do not work well with it. In [44] another hue definition was suggested, $H = \mathrm{R}/(\mathrm{R}+\mathrm{G})$, where R, G denote the red and green components. It is defined on **R**, and achieves nearly as good a reduction of intensity dependence as the conventional hue definition.

In addition to the color information, edges characterize object boundaries and provide additional useful information. We perform edge detection by using a Canny detection on the hue image. In the Canny detector, the input image H is convolved with the first derivative of a Gaussian function $G(i,j) = \sigma^2 e^{-(i^2+j^2)/2\sigma^2}$ (we set σ to 1.0 in our implementation) to obtain an image with enhanced edges. The convolution with the two-dimensional Gaussian can be separated into two convolutions with one-dimensional Gaussians in directions i and j. The magnitude of the result is computed at each pixel (i, j) as

$$e(i,j) = \sqrt{c_1 H_i'(i,j)^2 + c_2 H_j'(i,j)^2}, \tag{2}$$

where H_i' and H_j' are results of the convolutions between the first derivatives of the Gaussian and the image H in the two separate directions. Based on this magnitude, a nonmaximum suppression and double thresholding algorithm are performed and the edge map is derived. In expression (2), c_1 and c_2 are normally set to be equal. Since the lips contain mainly horizontal edges, we assign $c_1 = 10c_2$ to accentuate the importance of horizontal edges. This modification results in an improved edge map for lip images.

To combine the edge and hue color information, we have chosen to use the machinery of the Markov random field (MRF), which has been shown to be suitable for the problem of image segmentation. An MRF is a probabilistic model defined over a lattice of sites. The sites are related to each other through a neighborhood system. In MRFs, only neighboring sites have direct interaction with each other. Due to the Hammersley-Clifford theorem, the joint distribution of an MRF is equivalent to a Gibbs distribution, which takes the form

$$p(\mathbf{x}) = \frac{1}{Z}\exp\left(-\frac{1}{T}U(\mathbf{x})\right), \tag{3}$$

where Z is the normalizing constant, T the temperature constant, and $U(\mathbf{x})$ the Gibbs potential

$$U(\mathbf{x}) = \sum_{c\in C} V_c(\mathbf{x}), \tag{4}$$

which is the sum of clique potentials $V_c(\mathbf{x})$ over all possible cliques C.

In our problem, each site $s = (i, j)$ is assigned a label $x_s^l = 1$ (for lips) or 0 (for non-lips), and $x_s^e = 1$ (for edge) or 0 (for non-edge). Figure 7a shows configuration of lip sites and edge sites. Figures 7b and 7c show neighborhood systems for lip and horizontal edge sites, respectively. Here we use a first-order neighborhood system. A very similar two-label scheme can be found in [45]. The maximum a posteriori (MAP) criterion is used to formulate what the best labeling should be.

The MAP estimate is equivalent to that found by minimizing the posterior energy term

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} U(\mathbf{x}|\mathbf{y}), \tag{5}$$

where $\mathbf{x} = \{\mathbf{x}^l, \mathbf{x}^e\}$ denotes the configuration of the labeling, and $\mathbf{y}$ the observed image data.

Using Bayes' rule, the posterior probability is expressed as

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x}^l)\, p(\mathbf{y}|\mathbf{x}^e)\, p(\mathbf{x}), \tag{6}$$

where $p(\mathbf{y}|\mathbf{x}^l)$ and $p(\mathbf{y}|\mathbf{x}^e)$ represent the conditional probability distribution of the observed image color and edge data given the true interpretation of the images $\mathbf{x}^l$ and $\mathbf{x}^e$. They are modeled as follows:

$$p(\mathbf{y}|\mathbf{x}^l) \propto \exp\left[-\sum_s \frac{(y_s - \mu_{x_s^l})^2}{2\sigma_{x_s^l}^2}\right], \tag{7}$$

where $\mu_{x_s^l}$ and $\sigma_{x_s^l}$ are the mean and variance of all pixels in the image with the lip label x_s^l. They are obtained by using Otsu's methods [41] based on the histogram. The observed color data is represented by the hue color $y_s = \mathrm{R}/(\mathrm{R}+\mathrm{G})$ at site $s = (i, j)$.

In addition,

$$p(\mathbf{y}|\mathbf{x}^e) \propto \exp\left[-\sum_s e_s(1 - x_s^e)\right], \tag{8}$$

where e_s represents the strength of the edge at site s and is the magnitude derived from the Canny detector described in (2). The label x_s^e is the edge label at site s. It is 1 if there is an edge, and 0 otherwise. Since the edge map is defined for each pixel, we shift the edge map by 1/2 pixel downwards against the original image, so that x_s^e at $s = (i, j)$ indicates the edge between pixels (i, j) and $(i, j+1)$. For simplicity, we only consider horizontal edges.

By combining the above equations, it is clear that the MAP solution is equivalent to minimizing the following energy function:

$$U(\mathbf{x}|\mathbf{y}) = \sum_{c\in C} V_c(\mathbf{x}) + \lambda_1 \sum_s \frac{(y_s - \mu_{x_s^l})^2}{2\sigma_{x_s^l}^2} + \lambda_2 \sum_s e_s(1 - x_s^e). \tag{9}$$

In (9), the first term expresses the prior expectation and the second and third terms bind the solution to the color and edge data, respectively. We use $\lambda_1 = 2$ and $\lambda_2 = 1$. The V_c are the clique potentials describing the interactions between neighbors. They encode a priori knowledge about the spatial dependence of labels at neighboring sites. They are composed of three parts

$$V_c = k_1 V_c^l + k_2 V_c^e + k_3 V_c^{le}, \tag{10}$$

where $k_1 = 10$, $k_2 = 1$, and $k_3 = 1$. The first term in (10), V_c^l, imposes smoothness and continuity of color regions over an entire image, the second term, V_c^e, is responsible for boundary organization for the edges, and the third term, V_c^{le}, is the coupling term between the color and edge labels. There has been some work on applying statistical methods to estimate parameters for the clique potentials, such as in [46, 47]. However, choosing the clique potentials on an ad hoc basis has been reported to produce promising results [48, 49]. In this paper, we define these terms as follows:

$$V_c^l(i, j; i+1, j) = \begin{cases} -1 & \text{if } x^l(i, j) = x^l(i+1, j), \\ +1 & \text{otherwise;} \end{cases}$$

$$V_c^l(i, j; i, j+1) = \begin{cases} -1 & \text{if } x^l(i, j) = x^l(i, j+1), \\ +1 & \text{otherwise;} \end{cases}$$

$$V_c^e(i, j; i+1, j) = \begin{cases} -1 & \text{if } x^e(i, j) = x^e(i+1, j), \\ +1 & \text{otherwise;} \end{cases}$$

$$V_c^{le}(i, j; i, j+1) = \begin{cases} -1 & \text{if } x^l(i, j) \neq x^l(i, j+1),\ x^e(i, j) = 1, \\ -1 & \text{if } x^l(i, j) = x^l(i, j+1),\ x^e(i, j) = 0, \\ +1 & \text{otherwise.} \end{cases} \tag{11}$$

For the optimization strategy, a stochastic relaxation technique, such as simulated annealing, can be used to find the globally optimal interpretation for the image [45]. However, an exhaustive search for a global optimum imposes a large computational burden because the labels for all pixels need to be estimated simultaneously. Therefore, alternative estimates have been suggested, including using a Monte Carlo method [50], mean field technique [51], iterated conditional modes (ICM) [52], and high confidence first (HCF) algorithm [53]. We chose to use the HCF, because it is deterministic, computationally attractive, and achieves good performance. HCF differs from the other methods in the order of sites which are visited. Instead of updating the pixels sequentially, HCF requires that the site that is visited next be the one that causes the largest energy reduction. This procedure converges to a local minimum of the Gibbs potential within a relatively small number of cycles. The current lip feature extraction algorithm runs at a speed of 5 seconds per frame with an original image resolution of 720×480. The algorithm is designed to be scalable and can work in near-real time at lower image resolution with decreased tracking accuracy.

3.2. Visual speech features

Segmentation results with different persons and different lip opening situations are demonstrated in Figure 8. We observe that the highlighted pixels fairly well match the true lip area. Based on the segmented lip image, we are able to extract the key feature points on the lips [54]. We detect four feature points along the vertical lip line—the upper/lower outer/inner lip. To increase the accuracy of the identified feature points, we incorporate intensity gradient information. If the gradient of the detected point is below a preset value, we start searching for the largest gradient in its vicinity, and

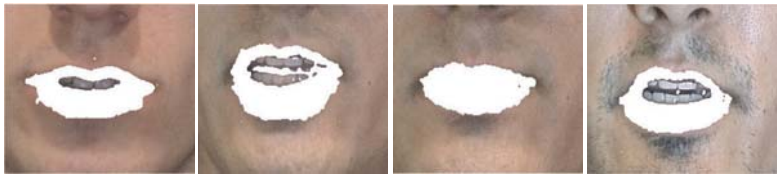


Figure 8: Segmented lips overlayed on the original image.

Figure 9: Measured feature points on the lips.

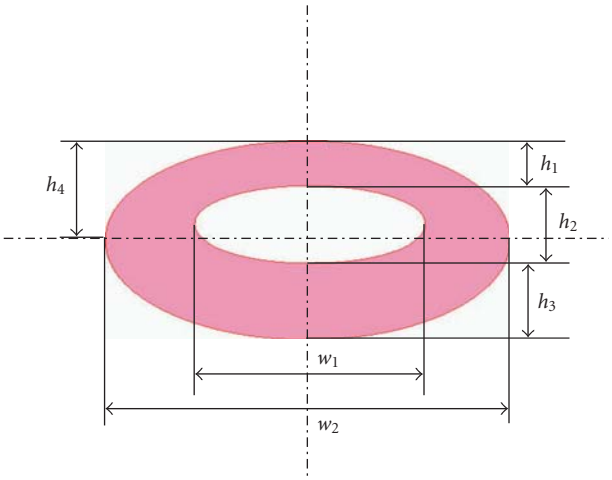
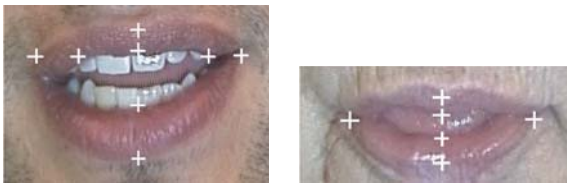


Figure 10: Illustration of the extracted geometric features of the lips.



(a) (b)

Figure 11: (a) Tongue is separated from the lips. (b) Tongue merges with the lips.

replace the old one with it. Finally, given the constraints of the outer corners and the upper/lower inner lip, we locate the inner lip corners. Examples of extracted feature points are shown in Figure 9.

Based on the extracted key feature points, we can derive the geometric dimensions of the lips. The following features are used in our study: mouth width (w_2), upper/lower lip width (h_1, h_3), lip opening height/width (h_2, w_1), and the distance between the horizontal lip line and the upper lip (h_4). An illustration of the geometry is shown in Figure 10.

Besides the geometric dimensions of the lips, we also detect the visibility of the tongue and teeth. For detecting the tongue, we search for the "lip" labels within the inner lip region. Two cases need to be differentiated, as shown in Figure 11. In the first case, the tongue is separated from the lips by the teeth. Tongue detection is trivial in this case. In the second case however, the tongue merges with the lips. From the segmented image, we have a lip closure case. Here, we use the gradient of the intensity to detect the inner upper/lower lip. In the case that $h_2 = 0$, we search for intensity gradient values along the vertical lip line. If the gradients of two points exceeding a preset value are found, they are identified as upper/lower inner lip. The parameter for the tongue is represented by the total number of lip-color pixels within the inner lip contour.

The teeth are also easy to detect since their H values are distinctly different from the hue of the lips. This is a big advantage compared with gray-level-based approaches that may confuse skin-lip and lip-teeth edges. Teeth are detected by forming a bounding box around the inner mouth area and testing pixels for white teeth color: $S < S_0$, where $S_0 = 0.35$. The parameter of the teeth is the total number of white pixels within the bounding box.

We applied the feature extraction algorithm on the Carnegie Mellon University database [33] with ten test subjects and the XM2VTS database [35] including 295 subjects. These two databases include head-shoulder full frontal face color video sequences of a person talking. Test subjects have various skin complexions with no particular lipstick. The first database was provided on DV tapes. We captured the sequences as AVI files with a resolution of 640×480 pixels and a frame rate of 30 frame/second. The second database was stored in DV encoded AVI format. The pixel resolution is 720×576 with a frame rate of 25 frame/second. The feature extraction algorithm works well for most of the sequences in the two data sets, which cover approximately seven hours and more than 300 individuals. In a few cases, a few pixels of inaccuracy are observed. The limitation of the color-based feature extraction occurs when the lip color and its surrounding skin color are very close to each other, which exists in a small percentage of the population. In these cases, the extraction of the key points on the upper and lower lips becomes unstable. We therefore attempt to control the errors by using the geometric constraint and time constraint

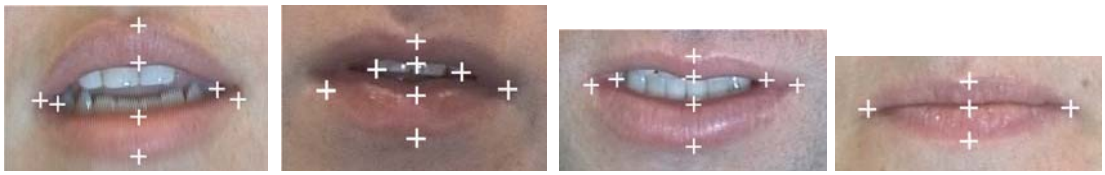


Figure 12: Examples of detected feature points.

methods. In the geometric constraint method the ratio between the lip opening height and the mouth width is less than a threshold, and in the time constraint method the variation of the measures between successive frames is within a limited range. Figure 12 shows examples of feature extraction results. Since the evaluation of feature extraction methods is often subjective, it is common that the direct evaluation is omitted at the visual feature level, and performance is evaluated based only on the final results of the system, which could be a speech recognition or speaker verification system. In our experiment, we evaluate the accuracy of the derived visual features for the tongue and teeth by randomly selecting a set of test sequences. These sequences are typically hundreds of frames long. We verify the computed results by visual inspection of the original images. The results show that the computed feature sets have an accuracy of 93% for the teeth and 91% for tongue detection, approximately.

4. AUTOMATIC SPEECH RECOGNITION

4.1. Visual speech recognition

In this section, we describe the modeling of the extracted lip features for speech recognition using hidden Markov models. HMMs [55] have been successfully used by the speech recognition community for many years. These models provide a mathematically convenient way of describing the evolution of time sequential data.

In speech recognition, we model the speech sequence by a first-order Markov state machine. The Markov property is encoded by a set of transition probabilities with $a_{ij} = P(q_t = j|q_{t-1} = i)$, the probability of moving to state j at time t given the state i at time $t-1$. The state at any given time is unknown or hidden. It can, however, be probabilistically inferred through the observations sequence $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T\}$, where $\mathbf{o}_t$ is the feature vector extracted at time frame t and T is the total number of observation vectors. The observation probabilities are commonly modeled as mixtures of Gaussian distributions

$$b_j(\mathbf{o}) = \sum_{k=1}^{M} c_{jk} \mathcal{N}(\mathbf{o}; \mu_{jk}, \mathbf{\Sigma}_{jk}), \tag{12}$$

where $\sum_{k=1}^{M} c_{jk} = 1$ and M is the total number of mixture components, μ_{jk} and $\mathbf{\Sigma}_{jk}$ are the mean vector and covariance matrix, respectively, for the kth mixture component in state j.

An HMM representing a particular word class is defined by a parameter set $\lambda = (A, B, \pi)$, where π is the vector of initial state probabilities, $A = \{a_{ij}\}$ the matrix of state transition probabilities, and $B = \{b_i(\mathbf{o}_t)\}$ the vector of state-dependent observation probabilities. Given a set of training data (segmented and labeled examples of speech sequences), the HMM parameters for each word class are estimated using a standard EM algorithm [56]. Recognition requires evaluating the probability that a given HMM would generate an observed input sequence. This can be approximated by using the Viterbi algorithm. For isolated word recognition considered in this paper, given a test token $\mathbf{O}$, we calculate $P(\mathbf{O}|\lambda_i)$ for each HMM, and select λ_c where $c = \arg\max_i P(\mathbf{O}|\lambda_i)$.

We perform the speech recognition task using the audio-visual database from Carnegie Mellon University [33]. This database includes ten test subjects (three females, seven males) speaking 78 isolated words repeated ten times. These words include numbers, days of the week, months, and others that are commonly used for scheduling applications. Figure 13 shows a snapshot of the database.

We conducted tests for both speaker-dependent and independent tasks using visual parameters only. The eight visual features used are: w_1, w_2, h_1, h_2, h_3, h_4 corresponding to Figure 10, and the parameters for the teeth/tongue. The visual feature vectors are preprocessed by normalizing against the average mouth width w_2 of each speaker to account for the difference in scale between different speakers and different record settings for the same person. For comparison, we also provide test results on partial feature sets. In particular, we limited the features to the geometric dimensions of the inner contour (w_1, h_2), and outer contour $(w_2, h_1 + h_2 + h_3)$. The role of the use of the tongue and teeth parameters was also evaluated. For the HMM, we use a left-right model and consider continuous density HMMs with diagonal observation covariance matrices, as is customary in acoustic ASR. We use ten states for each of the 78 HMM words and due to the training set size model the observation vectors using only two Gaussian mixtures for the speaker-independent task. Because of an even more limited training data available, we use only one Gaussian mixture in the speaker-dependent case. The recognition system was implemented using the Hidden Markov Model Toolkit (HTK) [57].

For the speaker-dependent task, the test was set up by using a leave-one-out procedure, that is, for each person, nine repetitions were used for training and the tenth for testing. This was repeated ten times. The recognition rate was averaged over the ten tests and again over all ten speakers. For the speaker-independent task, we use different speakers for training and testing, that is, nine subjects for training and the tenth for testing. The whole procedure was repeated ten

Table 1: Recognition rates for visual speech recognition using database [33]. The numbers represent the percentage of correct recognition.

Features	SD (static)	SD (static + Δ)	SI (static)	SI (static + Δ)
All (8)	45.51	45.59	18.17	21.08
All excl. tongue/teeth (6)	40.26	40.60	12.78	16.70
Outer/inner contour (4)	39.90	43.45	14.85	20.97
Outer contour (2)	28.72	35.16	7.9	12.55
Inner contour (2)	29.5	31.88	11.91	15.63

Figure 13: Examples of extracted lip ROI from the audio-visual database from Carnegie Mellon University [33].

times, each time leaving a different subject out for testing. The recognition rate was averaged over all ten speakers.

The experimental results for the two modes are shown in Table 1. Rows correspond to various combinations of visual features used. The numbers in brackets give the total number of features used in each test. The Δ refers to the delta features which are obtained by using a regression formula drawing in a few number of frames before and after the current frame. The second/third and forth/fifth columns give the average results in the speaker-dependent (SD) and speaker-independent (SI) mode, respectively. All recognition rates are given in percent.

We observe that the geometric dimensions of the lip outer contour, as used in many previous approaches [58, 59, 60], are not adequate for recovering the speech information. Comparing the case with a total of eight features, the rate drops by 16.79 percentage points for the SD and 10.27 percentage points for the SI case. While the use of the lip inner contour features achieves almost the same recognition rate as that of the lip outer contour in the SD mode, it outperforms the former by four percentage points in the SI task, and suggests it provides a better speaker-independent characteristic. The contribution of the use of tongue/teeth is about five percentage points in both tasks. The delta features yield additional improved accuracy by providing extra dynamic information. It is noted that while the contribution of the dynamic features in the eight features case is rather marginal for the speaker-dependent task, they are very important for the speaker-independent case. This suggests that the dynamic features are more robust across different talkers. Overall best results are obtained by using all relevant features, achieving 45.59% for the speaker-dependent case and 21.08% for the speaker-independent case.

4.2. Audio-visual integration

We consider speaker-dependent tasks in the following audio-visual speech recognition experiments. In our acoustic sub-system, we use 12 mel frequency cepstral coefficients (MFCCs) and their corresponding delta parameters as the features—a 24-dimensional feature vector. MFCCs are derived from FFT-based log spectra with a frame period of 11 milliseconds and a window size of 25 milliseconds. We employ a continuous HMM, where eight states and one mixture are used. The recognition system was implemented using the HTK Toolkit.

In the following, we examine three audio-visual integration models within the HMM based speech classification framework: early integration, late integration and multistream modeling [58, 60, 61, 62, 63]. The early integration model is based on a traditional HMM classifier on the concatenated vector of the audio and visual features

$$\mathbf{o}_t = \left[\mathbf{o}_t^{A\,T}, \mathbf{o}_t^{V\,T}\right]^T, \tag{13}$$

where $\mathbf{o}_t^A$ and $\mathbf{o}_t^V$ denote the audio- and visual-only feature vectors at time instant t. The video has a frame rate of 33 milliseconds. To match the audio frame rate of 11 ms, linear interpolation was used on the visual features to fit the data values between the existing feature data points.

The late integration model is built by applying separate acoustic and visual HMMs, and the combined scores take the following form: $\log P_{av} = \lambda \log P_a + (1 - \lambda) \log P_v$, where λ is

the weighting factor (0.7 in our experiments), and P_a and P_v are the probability scores of the audio and visual components.

In the expression of (13), early integration does not explicitly model the contribution and reliability of the audio and visual sources of information. To address this issue, we employ a multistream HMM model by introducing two stream exponents γ_A and γ_V in the formulation of the output distribution

$$b_j(\mathbf{o}_t^{AV}) = \left(\sum_{k=1}^{M_1} c_{1jk}\mathcal{N}(\mathbf{o}_t^A; \mu_{1jk}, \mathbf{\Sigma}_{1jk})\right)^{\gamma_A} \cdot \left(\sum_{k=1}^{M_2} c_{2jk}\mathcal{N}(\mathbf{o}_t^V; \mu_{2jk}, \mathbf{\Sigma}_{2jk})\right)^{\gamma_V}, \tag{14}$$

where M_1 and M_2 are the numbers of mixture components in audio and video streams. The exponents γ_A and γ_V are the weighting factors for each stream. We set $\gamma_A = 0.7$ and $\gamma_V = 0.3$ in our experiments, as was used in other similar implementations, such as in [62].

In the following, we present our experimental results on audio-visual speech recognition over a range of noise levels using these three models. We used the same database and data partition for the training and test as described in the last section for the visual speech recognition. Artificial white Gaussian noise was added to simulate various noise conditions. The experiment was conducted for speaker-dependent tasks under mismatched condition—the recognizers were trained at 30 dB SNR, and tested from 30 dB down to 0 dB in steps of 5 dB.

Figure 14 summarizes the performance of various recognizers. As can be seen, while the visual-only recognizer remains unaffected by acoustic noise, as must be the case since the signals were the same, the performance of the audio-only recognizer drops dramatically at high noise levels. A real-life experiment with actual noise might show variations in the visual only performance due to the Lombard effect [64, 65], but this aspect was not investigated (the Lombard effect was examined for example in study [66]).

In the speaker-dependent speech recognition, the multistream model performs the best among the three AV models at high SNR. Compared with the early integration model, the multistream model better explains the relations between the audio and video channels in this SNR range by emphasizing the reliability of the acoustic channel more. However at low SNR, the weighting factors of $\gamma_A = 0.7$ and $\gamma_V = 0.3$ are not appropriate any more, since the visual source of information becomes relatively more reliable.

Apart from the exception at high SNR for the late integration, all integrated models demonstrate improved recognition accuracy over audio-only results. However, the performance of the integrated systems drops below the performance of the visual-only system at very low SNRs, because the bad acoustic recognizer pulls down the total result. It is observed that the visual contribution is most distinct at low SNR. When the performance of the acoustic recognizer improves with increasing SNRs, the benefit of the addition of the visual component becomes less visible because there is less room for improvement. In total, when the best AV integration model is used, we obtain a performance gain of 27.97 percentage points at 0 dB and 8.05 percentage points at 30 dB over audio-only.

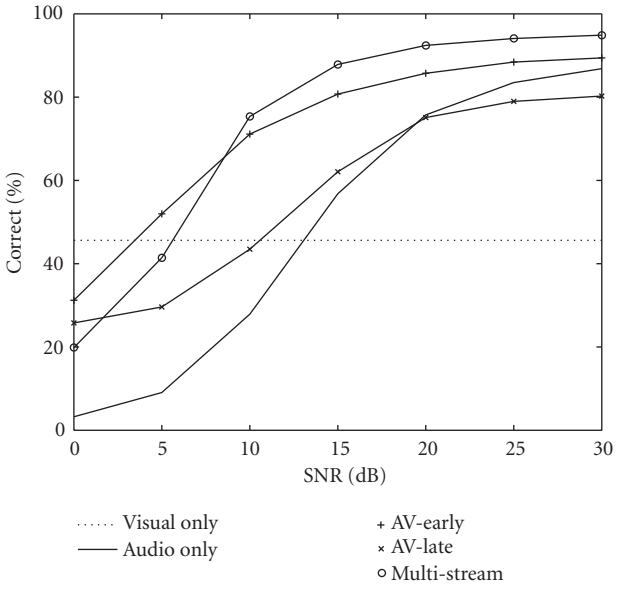


Figure 14: Performance comparison for various audio-visual speaker-dependent speech recognition systems under mismatched conditions. Recognition in speaker-dependent mode.

The CMU database [33] has been studied by several other groups [34, 67] for audio-visual speech recognition. However, only partial vocabulary and test subjects were used. To our knowledge, the results presented here are the first ones that evaluated the entire database.

5. SPEAKER VERIFICATION

The speaker verification task corresponds to an open test set scenario where persons who are unknown to the system might claim access. The world population is divided into two categories—a client who is known to the system, and imposters who falsely claim to have the identity of a client. Speaker verification is to validate a claimed identity: either to accept or reject an identity claim. Two types of error are possible: false acceptance of an imposter (FA), and false rejection of a client (FR).

For the speaker verification task, we use the polynomial-based approach [68]. Polynomial-based classification requires low computation while maintaining high accuracy. Because of the Weierstrass approximation theorem, polynomials are universal approximators for the Bayes classifier [69].

The classifier consists of several parts as shown in Figure 15. The extracted feature vectors $\mathbf{o}_1, \ldots, \mathbf{o}_N$ are introduced to the classifier. For each feature vector $\mathbf{o}_i$, a score is produced by using the polynomial discriminant

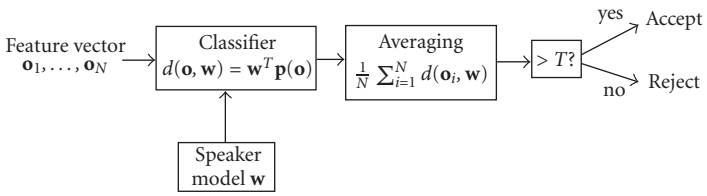


Figure 15: Structure of a polynomial classifier.

function $d(\mathbf{o}, \mathbf{w}) = \mathbf{w}^T \mathbf{p}(\mathbf{o})$, where $\mathbf{p}(\mathbf{o})$ is the polynomial basis vector constructed from the input vector $\mathbf{o}$, $\mathbf{p}(\mathbf{o}) = [1 \;\; o_1 \;\; o_2 \;\; o_1^2 \;\; o_1 o_2 \;\; o_2^2]^T$ for a two-dimensional feature vector $\mathbf{o} = [o_1 \;\; o_2]^T$ and for polynomial order two, and $\mathbf{w}$ is the class model. The polynomial discriminant function approximates the a posteriori probability of the client/impostor identity given the observation [69]. In [70, 71], a statistical interpretation of scoring was developed. The final score is computed by averaging over all feature vectors

$$\text{Score} = \frac{1}{N} \mathbf{w}^T \sum_{i=1}^{N} \mathbf{p}(\mathbf{o}_i). \tag{15}$$

The accept/reject decision is performed by comparing the output score to a threshold. If Score $< T$, then reject the claim, otherwise, accept the claim.

The verification system requires discriminative training in order to maximize its accuracy. For a speaker's features, an output value of 1 is desired. For impostors' features, an output of 0 is desired. The optimization problem can be formulated using a mean-squared error criterion

$$\mathbf{w}_{\text{spk}} = \arg\min_{\mathbf{w}} \left[\frac{1}{N_{\text{spk}}} \sum_{i=1}^{N_{\text{spk}}} \left| \mathbf{w}^T \mathbf{p}(\mathbf{o}_i) - 1 \right|^2 + \frac{1}{N_{\text{imp}}} \sum_{i=1}^{N_{\text{imp}}} \left| \mathbf{w}^T \mathbf{p}(\overline{\mathbf{o}}_i) \right|^2 \right], \tag{16}$$

where $\mathbf{o}_1, \ldots, \mathbf{o}_{N_{\text{spk}}}$ contain all training data for the user and $\overline{\mathbf{o}}_1, \ldots, \overline{\mathbf{o}}_{N_{\text{imp}}}$ are the data for the impostors. The reason to incorporate the weighting factors in (16) is to balance the number of vectors in the two classes, since normally there is a large amount of data for impostors and only a few values for the user. This equalization prevents overtraining on the impostor data set.

When expressed in matrix form, (16) can be rewritten as

$$\mathbf{w}_{\text{spk}} = \arg\min_{\mathbf{w}} \|\mathbf{DMw} - \mathbf{Du}\|_2, \tag{17}$$

where $\mathbf{D}$ is a diagonal matrix, $\mathbf{u}$ is the vector consisting of N_{spk} ones followed by N_{imp} zeros, and

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_{\text{spk}} \\ \mathbf{M}_{\text{imp}} \end{pmatrix} \tag{18}$$

with

$$\mathbf{M}_{\text{spk}} = \begin{pmatrix} \mathbf{p}(\mathbf{o}_1)^T \\ \mathbf{p}(\mathbf{o}_2)^T \\ \vdots \\ \mathbf{p}(\mathbf{o}_{N_{\text{spk}}})^T \end{pmatrix}, \qquad \mathbf{M}_{\text{imp}} = \begin{pmatrix} \mathbf{p}(\overline{\mathbf{o}}_1)^T \\ \mathbf{p}(\overline{\mathbf{o}}_2)^T \\ \vdots \\ \mathbf{p}(\overline{\mathbf{o}}_{N_{\text{imp}}})^T \end{pmatrix}. \tag{19}$$

It can be shown that (17) can be solved [72] by using

$$\mathbf{R}_{\text{spk}} + \frac{N_{\text{spk}}}{N_{\text{imp}}} \mathbf{R}_{\text{imp}} \quad \mathbf{w}_{\text{spk}} = \mathbf{M}_{\text{spk}}^T \mathbf{1}, \tag{20}$$

where $\mathbf{1}$ is the vector of N_{spk} ones, $\mathbf{R}_{\text{spk}} \equiv \mathbf{M}_{\text{spk}}^T \mathbf{M}_{\text{spk}}$ and $\mathbf{R}_{\text{imp}} \equiv \mathbf{M}_{\text{imp}}^T \mathbf{M}_{\text{imp}}$. Note that both matrices $\mathbf{R}_{\text{spk}}$ and $\mathbf{R}_{\text{imp}}$ are of fixed size and $\mathbf{R}_{\text{imp}}$ can be precomputed and stored in advance.

We perform the speaker verification test on the XM2VTS database [35]. This database includes four recordings of 295 subjects taken at one month intervals. (However we were able to use only 261 of the 295 speakers because of corrupted audio or video sequences [73].) Each sequence is approximately 5 seconds long and contains the subject speaking the sentence "*Joe took father's green shoe bench out.*" The database covers a large population variation from various ethnic origins and with various appearances. The same person might attend the four sessions with a different appearance, including hairstyles, with/without glasses, with/without beard, with/without lipstick. A snapshot of one person attending four sessions is shown in Figure 16.

To evaluate the performance of the person authentication systems on the XM2VTS database, we adopt the protocol defined in [74]. We chose configuration II due to the audio-visual data we are using. For the data partition defined in the protocol, each subject appears only in one set. This ensures realistic evaluation of the imposter claims whose identity is unknown to the system.

The verification performance is characterized by two error rates computed during the tests: the false acceptance rate (FAR) and the false rejection rate (FRR). The FAR is the percentage of the trials that the system falsely accepts an

Figure 16: Snapshot of the XM2VTS database [35].

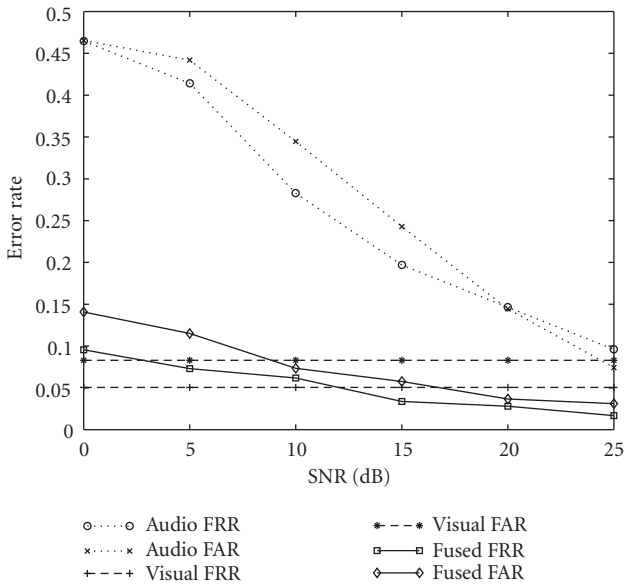


Figure 17: Performance of audio-visual speaker verification in noisy conditions. Speaker verification FRR and FAR at EER in varying noise conditions

imposter, and the FRR is the percentage of times access is denied to a valid claimant. The pooled equal error rate (EER) threshold at which FAR = FRR is determined from the evaluation set and used against the test population to determine the system performance. Both FAR and FRR are reported for this operating point. The test results for a visual-only speaker verification system are shown in Table 2.

In our experiment, polynomial orders two and three are used. The visual features included are the eight parameters derived in Section 3. Extra features are the corresponding delta features and the normalized time index i/M, where i is the current frame index, and M is the total number of frames. Since the score in a polynomial-based classifier (15) is an average of all feature vectors, the time index carries temporal information within the spoken sentence. As can be seen, by incorporating extra features, a lower error rate is achieved. At the same time, increasing the polynomial order also contributes to improved verification results.

To our knowledge, there were no other published results on using visual speech features for the speaker verification experiments based on the XM2VTS database

Table 2: Performance for the speaker verification tasks using database [35].

Features	Poly. order	FRR%	FAR%
All (8)	2	8.8	9.7
All + Δ (16)	2	6.1	9.3
All (8)	3	5.0	9.0
All + Δ (16)	3	4.4	8.2
All + time (9)	2	8.3	9.2
All + time (9)	3	4.8	8.5

(Studies [13, 75] performed speaker verification experiments on a smaller set of the M2VTS database). However, the XM2VTS database has been extensively used by the face recognition community. A face verification contest was organized at the *International Conference on Pattern Recognition*, 2000 to promote a competition for the best face verification algorithm. The tests were carried out using the static image shots of the XM2VTS database. All research groups participated in the contest used the same database and the same protocol for training and evaluation. A total of fourteen face verification methods were tested and compared [76]. For the same configuration as carried out in our speaker verification experiments, the published results of FAR/FRR range from 1.2/1.0 to 13.0/12.3. This suggests that our speaker verification approach that uses the lip modality is comparable to the state-of-the-art face-based personal authentication methods.

In the audio modality, each feature vector is composed of 12 cepstral coefficients and one normalized time index [68]. A third-order polynomial classifier is used. To fuse the two modalities, we use a late integration strategy. We combine the classifier outputs from the audio and visual modalities by averaging the class scores, $s = \alpha s_A + (1 - \alpha)s_V$, where $s_{A,V}$ are computed from (15) for the audio and visual channels. For the following experiments, the audio and visual modalities are weighted equally (i.e., $\alpha = 0.5$).

The performance of the bimodal speaker verification system is shown in Figure 17. Artificial white noise was added to clean speech to simulate various noise conditions. The performance was measured from 0 dB to 25 dB in steps of 5 dB. This figure shows the FRR and the FAR for each modality independently, as well as for the fused system. Both curves are of interest since the threshold is determined with an evaluation population separated from the test population. As can

be seen, the contribution of the visual modality is most distinct at low SNR. We observe an error rate drop of 36 percentage points for FRR and 32 percentage points for FAR at 0 dB over audio-only when visual modality is incorporated. As illustrated in the figure, the audio-visual fusion is shown to outperform both modalities at high signal-to-noise ratios. However, error rates over the low range of signal-to-noise ratios (SNR) are worse than the visual-only results and it indicates that a dynamic fusion strategy, for example, adjusting the weighting of the modalities as SNR degrades, may improve the overall system performance.

6. SUMMARY

In this paper, we described a method of automatic lip feature extraction and its applications to speech and speaker recognition. Our algorithm first reliably locates the mouth region by using hue/saturation and motion information from a color video sequence of a speaker's frontal view. The algorithm subsequently segments the lip from its surroundings by making use of both color and edge information, combined within a Markov random field framework. The lip key points that define the lip position are detected and the relevant visual speech parameters are derived and form the input to the recognition engine. We then demonstrated two applications by exploring these visual parameters. Experiments for automatic speech recognition involve discrimination of a set of 78 isolated words spoken by ten subjects [33]. It was found that by enabling extraction of an expanded set of visual speech features including the lip inner contour and the visibility of the tongue and teeth, the proposed visual front end achieves an increased accuracy when compared with previous studies that use only lip outer contour features. Three popular audio-visual integration schemes were considered and the visual information is shown to improve recognition performance over a variety of acoustic noise levels. In the speaker verification task, we employed a polynomial based approach. The speaker verification experiments on the database with 261 speakers achieve an FRR of 4.4% and an FAR of 8.2% with polynomial order 3, and suggest that visual information is highly effective in reducing both false acceptance and false rejection rates in such tasks.

ACKNOWLEDGMENTS

The research on which this paper is based acknowledges the use of the extended multimodal face database and associated documents [35, 36]. We would also like to acknowledge the use of audio-visual data [33, 34] from the Advanced Multimedia Processing Lab at the Carnegie Mellon University.

REFERENCES

[1] T. Sullivan and R. Stern, "Multi-microphone correlation-based processing for robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 91–94, Minneapolis, Minn, USA, April 1993.

[2] R. M. Stern, A. Acero, F.-H. Liu, and Y. Ohshima, "Signal processing for robust speech recognition," in *Automatic Speech and Speaker Recognition: Advanced Topics*, C.-H. Lee, F. K. Soong, and K. K. Paliwal, Eds., pp. 357–384, Kluwer Academic Publishers, Boston, Mass, USA, 1996.

[3] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," *Computer Speech and Language*, vol. 1, no. 2, pp. 109–130, 1986.

[4] B.-H. Juang, "Speech recognition in adverse environments," *Computer Speech and Language*, vol. 5, no. 3, pp. 275–294, 1991.

[5] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, 1954.

[6] N. P. Erber, "Interaction of audition and vision in the recognition of oral speech stimuli," *Journal of Speech and Hearing Research*, vol. 12, pp. 423–425, 1969.

[7] E. D. Petajan, *Automatic lipreading to enhance speech recognition*, Ph.D. thesis, University of Illinois, Urbana-Champaign, Ill, USA, 1984.

[8] D. G. Stork and M. E. Hennecke, Eds., *Speechreading by Humans and Machines*, vol. 150 of *NATO ASI Series F: Computer and Systems Sciences*, Springer-Verlag, Berlin, Germany, 1996.

[9] C. Neti, G. Potamianos, J. Luettin, et al., "Audio-visual speech recognition," Tech. Rep., CLSP/Johns Hopkins University, Baltimore, Md, USA, 2000.

[10] R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proceedings of the IEEE*, vol. 83, no. 5, pp. 705–740, 1995.

[11] P. S. Penev and J. J. Atick, "Local feature analysis: A general statistical theory for object representation," *Network: Computation in Neural Systems*, vol. 7, no. 3, pp. 477–500, 1996.

[12] J. Luettin, *Visual speech and speaker recognition*, Ph.D. thesis, University of Sheffield, Sheffield, UK, 1997.

[13] T. J. Wark and S. Sridharan, "A syntactic approach to automatic lip feature extraction for speaker identification," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 6, pp. 3693–3696, Seattle, Wash, USA, May 1998.

[14] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski, "Integration of acoustic and visual speech signals using neural networks," *IEEE Communications Magazine*, vol. 27, pp. 65–71, November 1989.

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

[16] C. Bregler and Y. Konig, "Eigenlips for robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, pp. 669–672, Adelaide, Australia, 1994.

[17] P. L. Silsbee and A. C. Bovik, "Computer lipreading for improved accuracy in automatic speech recognition," *IEEE Trans. Speech, and Audio Processing*, vol. 4, no. 5, pp. 337–351, 1996.

[18] G. Potamianos, J. Luettin, and C. Neti, "Hierarchical discriminant features for audio-visual LVCSR," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 165–168, Salt Lake City, Utah, USA, May 2001.

[19] K. Mase and A. Pentland, "Automatic lipreading by optical flow analysis," *Systems and Computers in Japan*, vol. 22, no. 6, pp. 67–75, 1991.

[20] A. J. Goldschen, O. N. Garcia, and E. Petajan, "Continuous optical automatic speech recognition by lipreading," in *IEEE Proc. the 28th Asilomar Conference on Signals, Systems and Computers*, pp. 572–577, Pacific Grove, Calif, USA, October–November 1994.

[21] A. L. Yuille, P. Hallinan, and D. S. Cohen, "Feature extraction from faces using deformable templates," *International Journal of Computer Vision*, vol. 8, no. 2, pp. 99–112, 1992.

[22] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.

[23] J. Luettin, N. A. Thacker, and S. W. Beet, "Active shape models for visual speech feature extraction," in *Speechreading by Humans and Machines*, D. G. Stork and M. E. Hennecke, Eds., vol. 150 of *NATO ASI Series F: Computer and Systems Sciences*, pp. 383–390, Springer-Verlag, Berlin, 1996.

[24] M. U. Ramos Sanchez, J. Matas, and J. Kittler, "Statistical chromaticity models for lip tracking with B-splines," in *Proc. the 1st Int. Conf. on Audio- and Video-Based Biometric Person Authentication*, Lectures Notes in Computer Science, pp. 69–76, Springer-Verlag, Crans-Montana, Switzerland, 1997.

[25] M. Vogt, "Interpreted multi-state lip models for audio-visual speech recognition," in *Proc. of the ESCA Workshop on Audio-Visual Speech Processing, Cognitive and Computational Approaches*, pp. 125–128, Rhodes, Greece, September 1997.

[26] G. I. Chiou and J. Hwang, "Lipreading from color video," *IEEE Trans. Image Processing*, vol. 6, no. 8, pp. 1192–1195, 1997.

[27] M. Chan, "Automatic lip model extraction for constrained contour-based tracking," in *Proc. IEEE International Conference on Image Processing*, vol. 2, pp. 848–851, Kobe, Japan, October 1999.

[28] A. A. Montgomery and P. L. Jackson, "Physical characteristics of the lips underlying vowel lipreading performance," *Journal of the Acoustical Society of America*, vol. 73, no. 6, pp. 2134–2144, 1983.

[29] A. Q. Summerfield, "Lipreading and audio-visual speech perception," *Philosophical Transactions of the Royal Society of London, Series B*, vol. 335, no. 1273, pp. 71–78, 1992.

[30] A. Q. Summerfield, A. MacLeod, M. McGrath, and M. Brooke, "Lips, teeth and the benefits of lipreading," in *Handbook of Research on Face Processing*, A. W. Young and H. D. Ellis, Eds., pp. 223–233, Elsevier Science Publishers, Amsterdam, North Holland, 1989.

[31] M. McGrath, *An examination of cues for visual and audio-visual speech perception using natural and computer-generated face*, Ph.D. thesis, University of Nottingham, Nottingham, UK, 1985.

[32] D. Rosenblum and M. Saldaña, "An audiovisual test of kinematic primitives for visual speech perception," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 22, no. 2, pp. 318–331, 1996.

[33] http://amp.ece.cmu.edu/projects/audiovisualspeechprocessing .

[34] T. Chen, "Audiovisual speech processing: lip reading and lip synchronization," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 9–21, 2001.

[35] http://www.ee.surrey.ac.uk/Research/vssp/xm2vtsdb.

[36] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. 2nd Int. Conf. on Audio- and Video-Based Biometric Personal Verification*, pp. 72–77, Washington, DC, USA, March 1999.

[37] C. Garcia and G. Tziritas, "Face detection using quantized skin color regions merging and wavelet packet analysis," *IEEE Trans. Multimedia*, vol. 1, no. 3, pp. 264–277, 1999.

[38] J. Yang, R. Stiefelhagen, U. Meier, and A. Waibel, "A real-time face tracker," in *Proc. 3rd IEEE Workshop on Application of Computer Vision*, pp. 142–147, Sarasota, Fla, USA, 1996.

[39] K. Jack, *Video Demystified: A Handbook for the Digital Engineer*, LLH Technology Publishing, Eagle Rock, Va, USA, 1996.

[40] J. R. Kender, "Instabilities in color transformations," *IEEE Computer Society*, pp. 266–274, June 1977.

[41] N. Otsu, "A threshold selection method from gray level histograms," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[42] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," *International Journal of Computer Vision*, vol. 46, no. 1, pp. 81–96, 2002.

[43] A. W. Senior, "Face and feature finding for a face recognition system," in *Proc. 2nd International Conference on Audio- and Video-Based Biometric Person Authentication*, pp. 154–159, Washington, DC, USA, 1999.

[44] A. Hurlbert and T. Poggio, "Synthesizing a color algorithm from examples," *Science*, vol. 239, pp. 482–485, January 1988.

[45] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, 1984.

[46] C. F. Borges, "On the estimation of Markov random field parameters," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 3, pp. 216–224, 1999.

[47] S. Geman and C. Graffigne, "Markov random field image models and their applications to computer vision," in *Proc. Int. Congress of Mathematicians*, pp. 1496–1517, Berkeley, Calif, USA, 1986.

[48] P. B. Chou and C. M. Brown, "The theory and practice of Bayesian image labeling," *International Journal of Computer Vision*, vol. 4, no. 3, pp. 185–210, 1990.

[49] J. Marroquin, S. Mitter, and T. Poggio, "Probabilistic solution of ill-posed problems in computational vision," in *Proc. Image Understanding Workshop*, pp. 293–309, San Diego, Calif, USA, 1985.

[50] J. Marroquin, S. Mitter, and T. Poggio, "Probabilistic solution of ill-posed problems in computational vision," *Journal of American Statistical Association*, vol. 82, no. 397, pp. 76–89, 1987.

[51] D. Geiger and F. Girosi, "Parallel and deterministic algorithms from MRFs: Surface reconstruction," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 13, no. 5, pp. 401–412, 1991.

[52] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society B*, vol. 48, no. 3, pp. 259–302, 1986.

[53] P. Chou, C. Brown, and R. Raman, "A confidence-based approach to the labeling problem," in *Proc. IEEE Workshop on Computer Vision*, pp. 51–56, Miami Beach, Fla, USA, 1987.

[54] X. Zhang, R. M. Mersereau, M. Clements, and C. C. Broun, "Visual speech feature extraction for improved speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 1993–1996, Orlando, Fla, USA, May 2002.

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

[56] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164–171, 1970.

[57] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, Entropic, Cambridge, UK, 1999.

[58] A. Adjoudani and C. Benoit, "On the integration of auditory and visual parameters in an HMM-based ASR," in *Speechreading by Humans and Machines*, D. G. Stork and M. E. Hennecke, Eds., vol. 150 of *NATO ASI Series F: Computer and Systems Sciences*, pp. 461–471, Springer-Verlag, Berlin, 1996.

[59] M. T. Chan, Y. Zhang, and T. S. Huang, "Real-time lip tracking and bimodal continuous speech recognition," in *Proc. IEEE Signal Processing Society 1998 Workshop on Multimedia Sig-*

nal Processing, pp. 65–70, Los Angeles, Calif, USA, December 1998.
[60] R. R. Rao, *Audio-visual interaction in multimedia*, Ph.D. thesis, Georgia Institute of Technology, Atlanta, Ga, USA, 1998.
[61] S. Dupont and J. Luettin, "Audio-visual speech modelling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.
[62] J. Luettin, G. Potamianos, and C. Neti, "Asynchronous stream modeling for large vocabulary audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 169–172, Salt Lake City, Utah, USA, May 2001.
[63] J. R. Movellan and G. Chadderdon, "Channel separability in the audio-visual integration of speech: A Bayesian approach," in *Speechreading by Humans and Machines*, D. G. Stork and M. E. Hennecke, Eds., vol. 150 of *NATO ASI Series F: Computer and Systems Sciences*, pp. 473–487, Springer-Verlag, Berlin, 1996.
[64] E. Lombard, "Le signe de l'elévation de la voix," *Ann. Maladiers Oreille, Larynx, Nez, Pharynx*, vol. 37, pp. 101–119, 1911.
[65] H. L. Lane and B. Tranel, "The Lombard sign and the role on hearing in speech," *Journal of Speech and Hearing Research*, vol. 14, no. 4, pp. 677–709, 1971.
[66] F. J. Huang and T. Chen, "Consideration of Lombard effect for speechreading," in *Proc. Workshop Multimedia Signal Processing*, pp. 613–618, Cannes, France, October 2001.
[67] A. V. Nefian, L. Liang, X. Pi, X. Liu, C. Mao, and K. Murphy, "A coupled HMM for audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 2013–2016, Orlando, Fla, USA, May 2002.
[68] C. C. Broun and X. Zhang, "Multimodal fusion of polynomial classifiers for automatic person recognition," in *SPIE 15th AeroSense Symposium*, pp. 166–174, Orlando, Fla, USA, April 2001.
[69] J. Schuermann, *Pattern Classification, a Unified View of Statistical and Neural Approaches*, John Wiley & Sons, New York, USA, 1996.
[70] W. M. Campbell and C. C. Broun, "Using polynomial networks for speech recognition," in *Neural Networks for Signal Processing X, Proceedings of the 2000 IEEE Workshop*, pp. 795–803, Sydney, Australia, 2000.
[71] W. M. Campbell, K. T. Assaleh, and C. C. Broun, "Speaker recognition with polynomial classifiers," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 4, pp. 205–212, 2002.
[72] G. H. Golub and C. F. Van Loan, *Matrix Computations*, John Hopkins University Press, Baltimore, Md, USA, 2nd edition, 1989.
[73] http://users.ece.gatech.edu/xzhang/note_on_defect_seq_in_xm2vts.txt.
[74] J. Luettin and G. Maitre, "Evaluation protocol for the XM2VTS database," IDIAP-COM 98-05, IDIAP, October 1998.
[75] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner, "Acoustic-labial speaker verification," *Pattern Recognition Letters*, vol. 18, no. 9, pp. 853–858, 1997.
[76] J. Matas, M. Hamouz, K. Jonsson, et al., "Comparison of face verification results on the XM2VTS database," in *Proc. the 15th International Conference on Pattern Recognition*, A. Sanfeliu, J. J. Villanueva, M. Vanrell, R. Alqueraz, J. Crowley, and Y. Shirai, Eds., vol. 4, pp. 858–863, Los Alamitos, Calif, USA, September 2000.

Xiaozheng Zhang received the Diploma in electrical engineering in 1997 from the University of Erlangen-Nuremberg, Germany. In the same year, she joined the Telecommunications Institute at the same university, where she worked on motion compensation related video compression techniques. Starting from winter 1998, she continued her education at the Georgia Institute of Technology, Atlanta, GA, where she conducted research on joint audio-visual speech processing. She received the Ph.D. degree in electrical and computer engineering in May 2002. She is the recipient of the 2002 Outstanding Research Award from the Center of Signal and Image Processing, Georgia Institute of Technology. Her research interests are in the areas of computer vision, image analysis, statistical modeling, data fusion, and multimodal interaction.

Charles C. Broun received his B.S. degree in 1990, and the M.S. degree in 1995, both in electrical engineering from the Georgia Institute of Technology. He specialized in digital signal processing, communications, and optics. From 1996 to 1998, he conducted research in speech and speaker recognition at Motorola SSG in the Speech and Signal Processing Lab located in Scottsdale, Arizona. There he worked both the speaker verification product CipherVOX, and the System Voice Control component of the U.S. Army's Force XXI/Land Warrior program. In 1999 he joined the newly formed Motorola Labs—Human Interface Lab in Tempe, Arizona, where he expanded his work to *multimodal* speech and speaker recognition. Currently, he is the project manager for several realization efforts within the Human Interface Lab—Intelligent Systems Lab. The most significant is the Driver Advocate, a platform supporting research in driver distraction mitigation and workload management.

Russell M. Mersereau received his S.B. and S.M. degrees in 1969 and the Sc.D. degree in 1973 from the Massachusetts Institute of Technology. He joined the School of Electrical and Computer Engineering at the Georgia Institute of Technology in 1975. His current research interests are in the development of algorithms for the enhancement, modeling, and coding of computerized images, and computer vision. He is the coauthor of the book *Multidimensional Digital Signal Processing*. Dr. Mersereau has served on the Editorial Board of the *Proceedings of the IEEE* and as Associate Editor for signal processing of the IEEE Transactions on Acoustics, Speech, and Signal Processing and Signal Processing Letters. He is the corecipient of the 1976 Bowder J. Thompson Memorial Prize of the IEEE for the best technical paper by an author under the age of 30, a recipient of the 1977 Research Unit Award of the Southeastern Section of the ASEE, and three teaching awards. He was awarded the 1990 Society Award of the Signal Processing Society.

Mark A. Clements received the S.B. degree in 1976, the S.M. degree in 1978, the Electrical Engineering degree in 1979, and the Sc.D. degree in 1982, all in electrical engineering and computer science from MIT. During his graduate work, he was supported by a National Institutes of Health fellowship for research in hearing prostheses, and corporate sponsorship for the development of real-time automatic speech recognition systems. He has been Professor of Electrical and Computer Engineering at the Georgia Institute of Technology since 1982. He has numerous publications and patents in the area of speech processing. His present interests concern automatic speech recognition and digital speech coding. He is a member of the Acoustical Society of America, and is a Senior Member of IEEE. He has been a member of the IEEE Speech Technical Committee, and has served as an Editor for IEEE Transactions on Acoustics, Speech, and Signal Processing. He is currently the Director of the Interactive Media Technology Center (IMTC) and a Founder and Director of Fast-Talk Communications.