Discriminative Analysis of Lip Motion Features for Speaker Identification and Speech-Reading

H. Ertan Çetingül, *Student Member, IEEE*, Yücel Yemez, *Member, IEEE*, Engin Erzin, *Member, IEEE*, and A. Murat Tekalp, *Fellow, IEEE*

***Abstract*—There have been several studies that jointly use audio, lip intensity, and lip geometry information for speaker identification and speech-reading applications. This paper proposes using explicit *lip motion* information, instead of or in addition to lip intensity and/or geometry information, for speaker identification and speech-reading within a unified feature selection and discrimination analysis framework, and addresses two important issues: 1) Is using explicit lip motion information useful, and, 2) if so, what are the best lip motion features for these two applications? The best lip motion features for speaker identification are considered to be those that result in the highest discrimination of individual speakers in a population, whereas for speech-reading, the best features are those providing the highest phoneme/word/phrase recognition rate. Several lip motion feature candidates have been considered including dense motion features within a bounding box about the lip, lip contour motion features, and combination of these with lip shape features. Furthermore, a novel two-stage, spatial, and temporal discrimination analysis is introduced to select the best lip motion features for speaker identification and speech-reading applications. Experimental results using an hidden-Markov-model-based recognition system indicate that using explicit lip motion information provides additional performance gains in both applications, and lip motion features prove more valuable in the case of speech-reading application.**

***Index Terms*—Bayesian discriminative feature selection, lip motion, speaker identification, speech recognition, temporal discriminative feature selection.**

I. Introduction

LIP information has been extensively employed in the state-of-the-art audio-visual speech and speaker recognition applications, since lip movements are highly correlated with the audio signal. Hence, it is natural to expect that speech content can be revealed through lip reading; and lip movement patterns also contain information about the identity of the speaker. In audio-visual recognition literature, there exist three alternative representations for lip information: 1) lip texture; 2) lip geometry (shape); and 3) lip motion features. The first alternative implicitly represents lip movements with texture. The texture information itself might sometimes carry useful discrimination information; but in some other cases it may degrade the recognition performance since it is sensitive to acquisition conditions. The second, lip geometry, usually requires tracking of the lip contour and fitting contour model parameters and/or computing geometric features such as horizontal/vertical openings, contour perimeter, lip area, etc. This option may seem as the most powerful one for modeling lip movement, especially for the speech-reading problem, since it is easier to match mouth openings-closings with the corresponding phonemes. However, lip tracking and contour fitting are very challenging tasks, since contour tracking algorithms are in general very sensitive to lighting conditions and image quality. The last option is the use of explicit lip motion features, which are potentially easy to compute and robust to lighting variations between the training and test data sets. Determination of the best lip motion features for speech-reading and speaker identification is the focus of this work. We also consider combination of lip motion and lip geometry in this study.

In audio-visual speech recognition (speech-reading), lip texture information is widely used. In [1] and [2], principal component analysis (PCA) has been applied to raw lip intensity image to reduce its dimension, and the reduced vector is used as the visual feature. Another possibility is to use discrete cosine transform (DCT) coefficients of the gray-scale lip image [3]. Potamianos *et al.* [3] apply linear discriminant analysis (LDA) to the final feature vector formed by concatenating a number of consecutive feature vectors centered at the current frame so as to capture dynamic speech information. However, lip texture features are sensitive to intensity variations between the training and test data sets.

Geometric features have been employed in speech-reading [4]–[9], since it is easier to match mouth openings-closings with the corresponding phonemes. *Deformable templates* [4], [5], *active shape models* (ASM) [6], [10], [11], and *snakes* [12] have been used to obtain different lip geometry features; however, they all suffer from complex feature extraction and training procedures. In [5], Gaussian mixture models (GMM) are used to model both the lip and the non-lip region, and lip tracking is performed by deformable templates. A number of horizontal and vertical Euclidean distances representing the lip openings are then selected as features. Kaynak *et al.* [8] also use horizontal/vertical distances along with the orientation angle to represent lip shape. In fact, most of the techniques in the speech-reading literature utilize a combination of lip texture and primitive geometric lip shape features. In [13], the lip feature vector is formed by concatenating the Karhunen Lòeve transform (KLT) coefficients of the inner-outer lip contour points with the texture information which is represented in a similar way as in the

Manuscript received July 19, 2005; revised February 9, 2006. This work was supported in part by TUBITAK under the project EEEAG-101E026 and in part by the European FP6 Network of Excellence SIMILAR (http://www.similar.cc). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tamas Sziranyi.

The authors are with the Multimedia, Vision and Graphics Laboratory, College of Engineering, Koç University, Sariyer, Istanbul 34450, Turkey (e-mail: ecetingul@ku.edu.tr; yyemez@ku.edu.tr; eerzin@ku.edu.tr; mtekalp@ku.edu.tr).

Digital Object Identifier 10.1109/TIP.2006.877528

so-called *eigenlip* technique [1]. In [11], the geometric information modeled by ASM is used along with the gray-level appearance features and then fused with audio for speech recognition. Perez *et al.* [10] utilize a set of lip shape features extracted by ASM together with DCT coefficients of the gray-level appearance information.

There is only a limited amount of work reported in which explicit lip motion information is used for speech-reading. Aleksic *et al.* [12] use gradient vector flow (GVF) snakes to extract outer lip contour and calculate the lip movement at ten predefined points by point-wise coordinate difference. They then reduce the feature dimension by PCA and use lip features together with other facial animation features. However, selection of the best lip motion features has not been addressed within a framework.

For speaker identification, unlike speech-reading, lip information has been employed in only a few works. In [14], [15], the DCT coefficients of gray-scale lip images are considered as lip features. It is relatively easy to obtain this feature, but it again suffers from illumination variation between the training and test data sets. Lip geometry is used in [16], where lip segmentation is carried out by forming an accumulated difference image, and considering moving parts of that image. Then, a number of predefined horizontal and vertical distances are taken as geometric lip features. Mok *et al.*, in [17], find the outer lip contour by active shape models, and form a feature vector using both the model parameters and some additional distances representing the lip shape. In the audio-visual fusion system presented in [18], the lip contour is first tracked and then each contour pixel is associated with chromatic features that constitute the initial feature vector. The dimension of the feature vector is then reduced via PCA followed by LDA. However, the initial step of PCA reduction filters out some useful discrimination information valuable to biometric speaker identification, and temporal correlations in lip motion are not taken into account in discrimination analysis. The lip feature vector proposed in [19] for speaker verification is composed of lip shape parameters concatenated with intensity values along the lip contour. The feature dimension is then reduced by PCA with no discrimination analysis at all.

In the speaker identification literature, there are only two reported works employing explicit lip motion as lip features. In [20], following the computation of the optical flow between two consecutive lip frames, the power spectrum from the three-dimensional motion field is calculated and used as lip motion features. In [21], the lip motion is represented by the full set of DCT coefficients of the dense optical flow vectors computed within rectangular lip frames and then fused with face texture and acoustic features for multimodal speaker identification. However no discrimination analysis is performed, and no specific attention is paid to optimize the unimodal performance of the lip motion modality. In our recent studies [22], [23], we observed that speaker identification systems can benefit from discriminative lip motion feature extraction.

Although numerous methods have been proposed for integration of lip information to speech and speaker recognition solutions, there is no framework proposed for selection of the most discriminative lip motion features optimally in the literature. This paper aims at providing quantitative answers to the following open questions.

1) Is explicit lip motion, instead of or in addition to lip intensity and/or geometry useful for speech/speaker recognition?
2) If so, what are the best lip motion features for speech-reading and speaker identification applications?

In order to answer these questions, several lip motion feature candidates have been considered including dense motion features within a bounding box about the lip, lip contour motion features, and combination of these with lip shape features. Furthermore, a novel framework for two-stage, spatial and temporal, discrimination analysis is introduced to select the best lip motion features for speaker identification and speech-reading applications. Hence, the main contribution of this paper is introduction of a framework for determination of the most discriminative lip motion and shape features for speech-reading and speaker identification.

A speaker/speech recognition system has three major components: feature extraction, probabilistic modeling of features, and classification. The standard modeling and classification aspects of our system are briefly presented in Section II for speaker identification and speech-reading problems. The main focus of this paper is on the feature extraction/analysis part. Section III describes the lip extraction/tracking procedures that are employed, and different alternatives that are considered for lip motion representation. The success of a recognition system eventually depends on how efficiently the extracted lip information is represented in a relatively low-dimensional feature vector. For speech-reading, the general approach (in the literature) has been to extract the principal components of the lip movement in order to establish a one-to-one correspondence between phonemes of speech and visemes of lip shape. For the speaker identification problem, however, the use of lip motion requires more sophisticated processing, which has not been addressed in the literature. The main reason for this is that the principal components of the lip movement are not usually sufficient to well discriminate the biometric properties of a speaker. High frequency or nonprincipal components of the motion should also be valuable especially when the objective is to model specific lip movements of an individual rather than what is uttered. In other words, discrimination among speakers should be emphasized and selected features should minimize the recognition error rather than the reconstruction error. The discrimination analysis framework proposed in Section IV addresses this dimension reduction problem optimally, taking into account the intra-class and inter-class distribution of individual single-frame lip feature vectors as well as the temporal discrimination information. The experimental results are provided in Section V both for speaker identification and speech-reading problems.

II. System Overview

A. Speaker/Speech Recognition

Speaker and speech recognition tasks can be formulated as open-set or closed-set identification problems. In the closed-set identification problem, a reject scenario is not defined, and an

unknown observation is classified as belonging to one of the R registered pattern classes. In the open-set problem, the objective is, given the observation from an unknown pattern, to find whether it belongs to a pattern class registered in the database or not; the system identifies the pattern if there is a match and rejects otherwise. Hence, the problem can be thought of as an $R+1$ class identification problem, including a reject class.

The maximum *a posteriori* probability solution to the R-class closed-set classification problem requires, given a feature vector $\boldsymbol{f}$ representing the sample data of an unknown observation, computing $P(\lambda_r|\boldsymbol{f})$ for each class λ_r, $r=1,\ldots,R$. Alternatively, one can employ the maximum likelihood solution, which maximizes the class-conditional probability, $P(\boldsymbol{f}|\lambda_r)$, for $r=1,\ldots,R$. Hence, a decision in the closed-set identification is taken as

$$\lambda_* = \underset{\lambda_1,\ldots,\lambda_R}{\arg\max}\, P(\boldsymbol{f}|\lambda_r). \tag{1}$$

In the open-set identification problem, an imposter class λ_{R+1} can be introduced as the $R+1$th class. Since it is difficult to accurately model the imposter class, λ_{R+1}, we employ the following solution which includes a reject strategy through the definition of the log-likelihood ratio

$$\rho(\lambda_r) = \log\frac{P(\boldsymbol{f}|\lambda_r)}{P(\boldsymbol{f}|\lambda_{R+1})} = \log P(\boldsymbol{f}|\lambda_r) - \log P(\boldsymbol{f}|\lambda_{R+1}). \tag{2}$$

The decision strategy of the open-set identification can then be implemented in two steps. First, determine

$$\lambda_* = \underset{\lambda_1,\ldots,\lambda_R}{\arg\max}\ \rho(\lambda_r) \tag{3}$$

and then

$$\begin{array}{ll} \text{if } \rho(\lambda_*) \geq \tau, & \text{accept} \\ \text{otherwise}, & \text{reject} \end{array} \tag{4}$$

where τ is the optimal threshold which is usually determined experimentally to achieve the desired false accept or false reject rate.

Computation of class-conditional probabilities needs a prior modeling step, through which a feature probability density function is estimated for each class $r=1,\ldots,R$ by using available training data. A common and effective approach to model the impostor class is to use a universal background model, which is estimated by using all available training data regardless of which class they belong to.

B. Probabilistic Modeling of Features

Hidden Markov models (HMMs) are widely used for both audio-based speaker identification and speech recognition applications [24]. The speaker identification problem can further be classified as text-dependent or text-independent. In the text-independent case, speaker identification is performed over a content free utterance, whereas in the text-dependent case, each speaker is expected to utter a personalized secret phrase. State-of-the-art speaker identification systems use HMMs for the text-dependent case and GMMs for text-independent case [25]. HMM-based techniques are preferred in the text-dependent scenario since HMMs can successfully exploit the temporal correlations in a speech signal. Because lip motion is strongly coupled with the audio utterance, HMMs can also model temporal correlations of lip motion features effectively.

In this work, we address the speaker identification application under the text-dependent scenario as an open-set identification problem. We use word-level continuous-density HMM structures for temporal characterization of lip features. Each speaker in the database is modeled using a separate HMM that is trained over some repetitions of the lip motion sequences of the corresponding class. Given a test feature set, each HMM structure associated with a speaker produces a likelihood. A world HMM model representing the impostor class is also trained over the whole training data of the population. The log-ratios of the speaker likelihoods to the world class likelihood result in a stream of log-likelihood ratios that are used to identify or reject a speaker.

The performance of speaker identification systems are often measured using the equal error rate (EER) figure. The EER is calculated as the operating point where false accept rate (FAR) equals false reject rate (FRR). In the open-set identification case, false accept, and false reject rates can be defined as

$$\text{FAR} = 100\times\frac{F_a}{N_a+N_r} \text{ and } \text{FRR} = 100\times\frac{F_r}{N_a} \tag{5}$$

where F_a and F_r are the number of false accepts and rejects, and N_a and N_r are the total number of trials for the true and imposter clients in the testing, respectively.

In the speech-reading application, we employ the same HMM system described above. However, we address the speech-reading problem as a closed-set identification problem. Hence, an impostor class is not defined and the best match is given by the utterance class for which the corresponding HMM produces the highest likelihood as defined in (1). The performance of speech recognition systems is usually measured by the ratio of the number of true matches to the total number of trials.

III. Lip Motion Feature Extraction

The proposed lip motion feature extraction and analysis system is depicted in Fig. 1. It consists of a preprocessing module, a lip motion estimation module, a Bayesian discrimination module, and a temporal discrimination module. We consider two alternatives for lip motion estimation:

1) dense motion vectors within a rectangular grid;
2) motion vectors along the lip contour together with lip shape information.

Each of these modules are explained in detail as follows.

A. Preprocessing

The purpose of the preprocessing module is to register lip regions in successive frames by eliminating global head motion so that the extracted motion features within the lip region correspond to speaking act only. Hence, each frame of the sequence is aligned with the first frame using a two-dimensional (2-D) parametric motion estimator. For every two consecutive frames, global head motion parameters are calculated using hierarchical Gaussian image pyramids and the 12-parameter

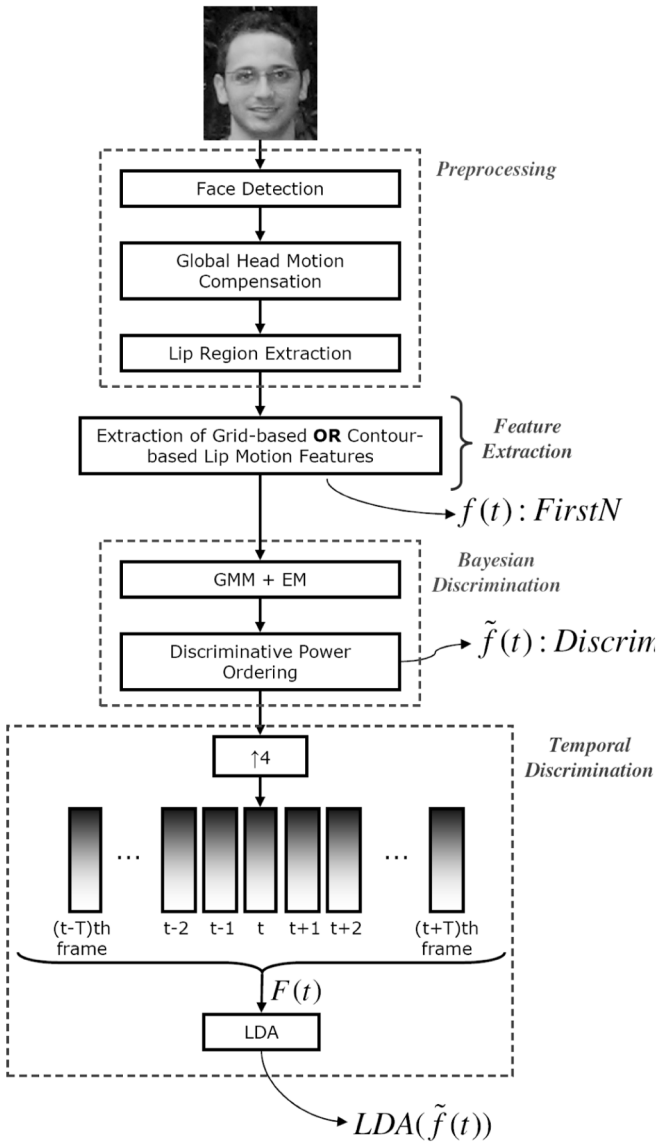


Fig. 1. Block diagram of the lip feature extraction system and the two-stage discrimination analysis.

quadratic motion model [26]. The frames are successively warped using the calculated parameters. Thus by only hand-labeling the mid-point of the lip region in the first frame, we can automatically extract the lip region for the whole sequence.

Using the quadratic transform to model head motion, the image intensity $I(x, y, t)$ of a pixel (x, y) at frame t is estimated from the intensity $I(u(x, y), v(x, y), t-1)$, by

$$\begin{aligned} u(x,y) &= a_1x^2 + a_2y^2 + a_3xy + a_4x + a_5y + a_6 \\ v(x,y) &= b_1x^2 + b_2y^2 + b_3xy + b_4x + b_5y + b_6. \end{aligned} \quad (6)$$

The quadratic transform provides an exact description of the three-dimensional (3-D) rotation, translation and scaling of an object with a parabolic surface under parallel projection [27]. Thus, it is effective in modeling rigid motion of the head between consecutive frames, where the movement is not very abrupt.

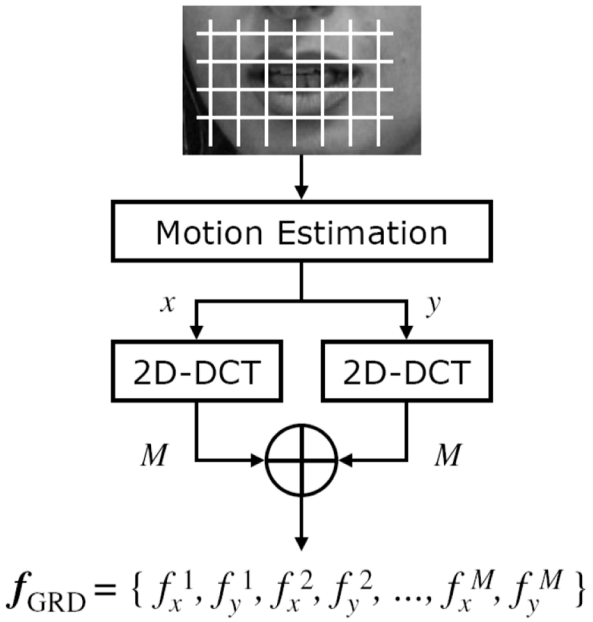


Fig. 2. Block diagram for extraction of grid-based lip motion features.

B. Extraction of Grid-Based Motion Features

We first consider dense motion estimation over a uniform grid of size $G_x \times G_y$ on the extracted lip region image. We use hierarchical block matching to estimate the lip motion with quarter-pel accuracy by interpolating the original lip image using the 6-tap Wiener and bilinear filters specified in H.264/MPEG-4 AVC [28].

The motion estimation procedure yields two $G_x \times G_y$ 2-D matrices, $\boldsymbol{V}_x$ and $\boldsymbol{V}_y$, which contain the x- and y- components of the motion vectors at grid points, respectively. The motion matrices, $\boldsymbol{V}_x$ and $\boldsymbol{V}_y$, are separately transformed via 2-D-DCT. The first M DCT coefficients along the zig-zag scan order, both for x and y directions, are combined to form a feature vector $\boldsymbol{f}$ of dimension $2M$ as depicted in Fig. 2. This feature vector representing the dense grid motion will be denoted by $\boldsymbol{f}_{\mathrm{GRD}}$.

Transforming the motion data into DCT domain has two advantages. First, it serves as a tool to reduce the feature dimension by filtering out the high frequency components of the motion signal. These high frequency components are mostly due to noise and irrelevant to our analysis since it is unnatural to have very abrupt motion changes between neighboring pixels of the lip region, where the motion signal is expected to have some smoothness. Second, DCT de-correlates the feature vector so that the discriminative power of each feature component can independently be analyzed as will later be addressed in Section IV.

C. Extraction of Contour-Based Motion Features

1) Lip Contour Extraction: The accuracy and robustness of the lip contour extraction method are crucial for a recognition system that uses lip shape information. There exist many techniques in the literature that attempt to solve the lip segmentation/tracking problem [12], [29]–[35]. The performance of these techniques usually depend on acquisition specifics such as image quality, resolution, head pose and illumination conditions. In region-based lip segmentation techniques, color

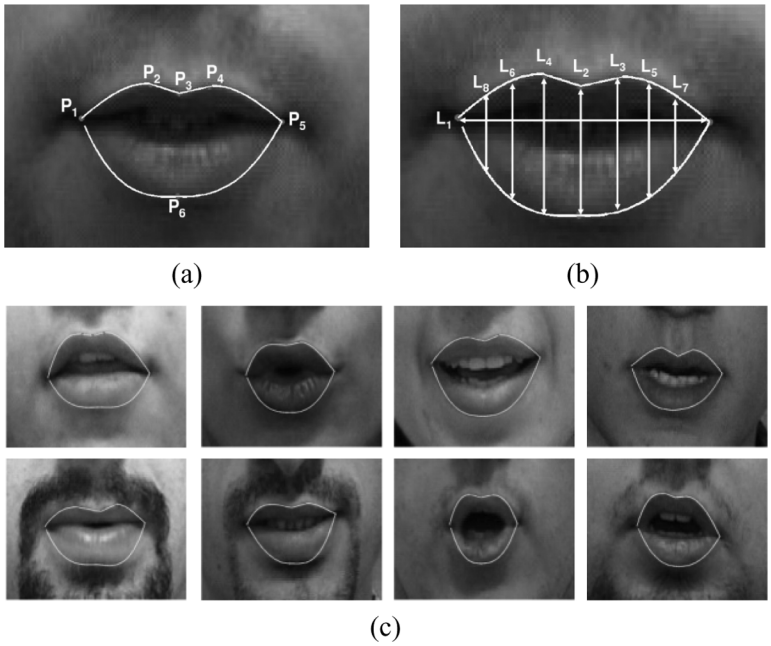


Fig. 3. Extraction of contour-based lip motion features: (a) The six key points and the parametric models fitted on the outer contour (P_2P_3, and P_3P_4 are line segments, whereas P_1P_2, P_4P_5, P_5P_6 and P_6P_1 are cubic polynomials). (b) The eight lip shape parameters. (c) Extracted lip contours.

information is often used as an important cue to differentiate lip pixels from those of the skin. In order to achieve this, the state-of-the-art techniques use, for instance, Markov random fields [34], LDA [35], adaptive Gaussian mixture models [29] or fuzzy clustering methods as in [30], [32]. There are also a number of boundary-based techniques to represent and to extract the lip contour, such as splines, active shape models, snakes, and parametric models, that use color gradient and/or edge information. Active shape models [6], [36] impose a priori information about possible lip movements so as to avoid unrealistic lip models, however they require a large training set of registered lip images acquired under predefined face orientation and lighting. Classical active contours [31] and their extensions such as GVF-snakes [12] suffer from complex parameter tuning, and they are unable to perfectly fit to certain characteristic lip parts such as Cupid's bow.

For lip contour extraction, we employ the quasi-automatic technique proposed in [33], where we fit polynomials on the outer lip contour. The technique is based on six designated key points detected on the lip contour (see Fig. 3). The algorithm starts with manual selection of a seed point above the mouth, that guides the so-called "jumping snake" onto the upper-lip boundary. The jumping snake along with the pseudo hue gradient information is then used to locate the upper and lower key points. The detected key points serves as the junction points of the four cubic polynomials and two line segments to be fitted onto the lip contour via least-squares optimization. The key points are tracked from one image to the other using a variant of the Lucas-Kanade algorithm [37] adapted to the particular geometry of the lip. Fig. 3(a) shows the six key points and the fitted parametric model on a sample lip image. When tested on our visual database, the technique proposed in [33] mostly yields very accurate lip tracking results. Nevertheless, the algorithm fails in about one-tenth of the sample video sequences. For some speakers, the lack of discriminative color information, especially on the lower lip boundary, becomes occasionally so severe that even a human eye can hardly make a distinction. Thus we have integrated a user interaction mechanism into the original algorithm described in [33]. In cases where it fails, the algorithm is assisted with some extra key points which are hand-labelled on the lip boundary. Fig. 3(c) displays examples of lip contours extracted from various images of our database.

2) Contour-Based Motion Features: In the contour-based lip motion representation, only motion vectors computed on the pixels along the extracted lip contour are taken into account and the rest is discarded. In this case, the two sequences of x and y motion components on the contour pixels are separately transformed using one-dimensional DCT. Note that the length of the resulting sequence of motion components on each direction may vary from one frame to another according to varying lip shape. In order to obtain a feature vector of fixed size in each frame, prior to 1-D DCT transformation, the length of the sequence is normalized to a fixed number by using linear interpolation. This number, $M_{\max}$, is the maximum number of contour points achieved in any lip frame of all available sequences. The DCT coefficients computed separately for x and y directions are concatenated to form the feature vector that is denoted by $\boldsymbol{f}_{\mathrm{CTR}}$. Fig. 4 depicts the procedure for extraction of contour-based lip motion features.

D. Lip Shape Features

The contour-based lip motion feature vector $\boldsymbol{f}_{\mathrm{CTR}}$ can further be fused with lip shape parameters to improve the representation. We will denote the lip shape feature vector by $\boldsymbol{f}_{\mathrm{SHP}}$. Recall that we parameterize the lip shape with four cubic polynomial and two line segments. Polynomial segments can be specified by sampling four points on each whereas a pair of endpoints is sufficient to represent a line segment. Since the lip contour is

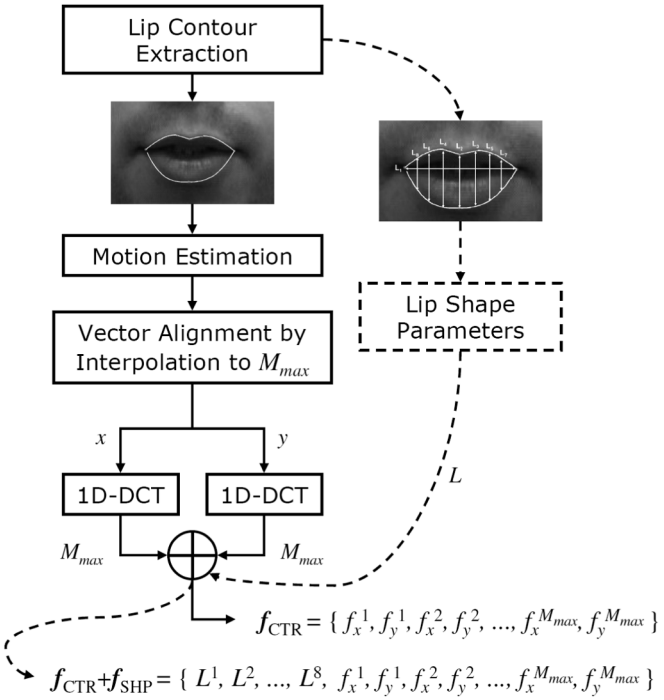


Fig. 4. Extraction of lip shape and contour-based motion features (the dashed lines show the optional path for the feature level fusion of lip shape and contour-based lip motion).

composed of these six segments articulated at their endpoints, a minimum number of 14 points is necessary to uniquely represent the parameterized lip shape, which corresponds to a feature vector of 28 point coordinates in x and y directions. These points should appropriately be sampled on the lip contour. In order to assure translation and rotation invariance, we represent the lip shape in terms of horizontal and vertical distances between the sampled points. A possible such feature vector is composed of eight simple parameters: the maximum horizontal distance, L_1, and the 7 vertical distances from the Cupid's bow and from the equidistant upper lip points to the lower lip boundary $(L_2, \ldots, L_8)$ as depicted in Fig. 3(b). The vertical lines are selected to be perpendicular to the line joining the two corners of the lip. The concatenation of lip shape parameters with contour-based motion information is illustrated in Fig. 4.

IV. Discrimination Analysis

There are a number of subspace representation techniques that can be used for reduction of dimensionality of feature vectors in recognition systems. Linear discriminant analysis (LDA) is a well-known dimension reduction and feature extraction method to achieve discrimination among multiple classes [3], [38], [39]. In this paper, we propose a novel approach for feature reduction, where we select the most discriminative lip motion features in two successive stages, the so-called Bayesian and temporal discrimination stages. In the Bayesian discrimination analysis stage, we use a probabilistic measure that maximizes the ratio of intra-class and inter-class probabilities. The temporal stage uses LDA analysis. The details of these two stages are discussed in the following.

A. Bayesian Discriminative Feature Selection

Let f_k denote the kth component of a feature vector $\boldsymbol{f}$. Given an observation f_k, the maximum *a posteriori* (MAP) estimator selects the class λ_i with the MAP probability $P(\lambda_i|f_k)$ which can be written in terms of class conditional probability distributions

$$\begin{aligned} P(\lambda_i|f_k) &= \frac{P(f_k|\lambda_i)P(\lambda_i)}{P(f_k)} \\ &= \frac{P(f_k|\lambda_i)P(\lambda_i)}{P(f_k|\lambda_i)P(\lambda_i) + \sum_{j\neq i} P(f_k|\lambda_j)P(\lambda_j)} \\ &= \left[1 + \frac{\sum_{j\neq i} P(f_k|\lambda_j)P(\lambda_j)}{P(f_k|\lambda_i)P(\lambda_i)}\right]^{-1}. \end{aligned} \tag{7}$$

Then the MAP estimator becomes the maximum mutual information estimator (MMIE) [40] by maximizing the ratio $l(\lambda_i|f_k)$

$$l(\lambda_i|f_k) = \log \frac{P(f_k|\lambda_i)P(\lambda_i)}{\sum_{j\neq i} P(f_k|\lambda_j)P(\lambda_j)}. \tag{8}$$

This ratio can be interpreted as the ratio of intra-class and inter-class probabilities, and when maximized, it can serve as a measure of discrimination between the class λ_i and all other classes for the corresponding feature component f_k.

When the class conditional probability distributions are available for a K dimensional feature vector $(f_1, f_2, \ldots, f_K)$, where the components are statistically independent, one can compute the discriminative power of the independent feature f_k^i that belongs to class λ_i using $l(\lambda_i|f_k^i)$. The larger the ratio $l(\lambda_i|f_k^i)$, the more discriminative is the feature; that is, the class conditional probability for its own class is high and the average of the class conditional probabilities over all other classes is low. In most cases, the class probabilities, $P(\lambda_i)$, can be assumed to be equally likely. The class conditional probability distributions are generally computed over some training data using expectation-maximization type algorithms, assuming an underlying probability distribution. Let us refer to this training data as f_k^i, that is a collection of observations of the kth feature component from the ith class, which is available for all feature components and for all classes. We propose the following discrimination measure, $d(f_k)$, to estimate the discriminative power of each feature f_k

$$d(f_k) = \sum_i \frac{1}{L} \sum_{l=0}^{L-1} l\left(\lambda_i|f_k^i(l)\right) \tag{9}$$

where L is the number of observations in each class λ_i.

1) Discriminative Feature Ranking: The proposed discrimination measure, when computed for each independent feature, creates an ordering $\{f_{k_i}\}$ among the components of the feature vector such that

$$d(f_{k_1}) \geq d(f_{k_2}) \geq \cdots \geq d(f_{k_K}). \tag{10}$$

This ordering can be used to select the most N discriminative features, or similarly to eliminate the least $K-N$ discriminative features from the full set of features. Then the reduced discriminative feature vector can be written as

$$\tilde{\boldsymbol{f}}^N = (f_{k_1}, f_{k_2}, \ldots, f_{k_N}). \tag{11}$$

This selection strategy makes sense whenever the joint discrimination measure of any two features is less than the sum of their individual discriminative powers. A sufficient condition for this is to have statistically independent features. In this case, the proposed ordering is a valid ordering with respect to feature discriminative power.

We considered two alternative feature vectors $\boldsymbol{f}_{\mathrm{GRD}}$ and $\boldsymbol{f}_{\mathrm{CTR}}$ to represent the lip motion in Section III. Both involve the DCT coefficients of the motion vectors computed either on a 2-D rectangular grid covering the lip region or along the one–dimensional (1-D) lip boundary pixels. Under the Gaussian distribution assumption, the DCT transformation de-correlates observation vectors so that each feature approximately becomes independent from the rest of the features. After applying the DCT transformation, traditionally, the low indexed N coefficients, that we refer as *FirstN*, are used as the representative features since they yield the best reconstruction for the original observations. Following the notation introduced in this section, this feature vector can be expressed as $\boldsymbol{f}^N = (f_1, f_2, \ldots, f_N)$. The discriminative set of features, $\tilde{\boldsymbol{f}}^N$, that are introduced in (11), will be referred to as *DiscrimN*. Note that they are selected according to the discriminative power ordering specified in (10). The class conditional probability distribution of each transform domain coefficient is estimated so that the discrimination measure for each coefficient can be calculated using (9). The Gaussian mixture models (GMMs) are used to represent the class-conditional probability density functions. For GMM estimation, the EM (expectation-maximization) algorithm is employed using diagonal covariance matrices, since feature components are assumed to be independent of each other.

2) Total Discrimination Measure: The proposed discrimination analysis also offers a means to assess and compare the expected identification performances of the different lip feature sets. Note that the measure $d(f_k)$ in (9) is an estimate of the discrimination power of each component in the feature vector. The discriminative power of the N selected features (the reduced feature vector) can then be estimated by the total discrimination measure, $D_N(\boldsymbol{f})$, which is defined as follows:

$$D_N(\boldsymbol{f}) = \sum_{n=1}^{N} d(f_{k_n}). \tag{12}$$

The numerical estimates for $D_N(\boldsymbol{f}_{\mathrm{GRD}})$ and $D_N(\boldsymbol{f}_{\mathrm{CTR}})$ will later be provided in the experimental results section along with the corresponding recognition results. Note that the Bayesian discrimination analysis can not be applied to the lip shape feature vector $\boldsymbol{f}_{\mathrm{SHP}}$ since the lip shape parameters, which are few in number, are not in general statistically independent of each other.

B. Temporal Discriminative Feature Selection Using LDA

The Bayesian discriminative feature selection technique described in Section IV-A does not model and exploit the temporal correlations existing between successive lip frames. Following the work of Potamianos *et al.* [3], we use the LDA for temporal discrimination analysis, where we successively concatenate the Bayesian-reduced lip feature vectors through a window of fixed duration so as to capture dynamic visual speech information, and obtain a new sequence of higher dimensional feature vectors. Then, each of these feature vectors is projected to a lower dimensional discriminative feature space using the LDA analysis.

The LDA maps a given high dimensional feature vector to a subspace of reduced dimension that best describes the discrimination among classes. This is achieved using two statistical measures, the within-class scatter matrix $(\boldsymbol{S}_w)$ and the between-class scatter matrix $(\boldsymbol{S}_b)$ [41]. The goal is to maximize the between-class scattering while minimizing the within-class variations. Hence, LDA seeks for a projection matrix $\boldsymbol{W}$ that maximizes the function:

$$\boldsymbol{\epsilon}(W) = \frac{\det\left(\boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W}\right)}{\det\left(\boldsymbol{W}^T \boldsymbol{S}_w \boldsymbol{W}\right)} \tag{13}$$

provided that $\boldsymbol{S}_w$ is a nonsingular matrix. The $\boldsymbol{\epsilon}(W)$ function is maximized when the column vectors of the projection matrix $\boldsymbol{W}$ are the eigenvectors of $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$. The LDA has two important limitations.

- The matrix $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$ has nonzero eigenvalues at most one less than the total number of classes $(R-1)$, that puts an upper bound on the reduced dimension.
- At least $K+R$ training samples are needed to guarantee the existence of the inverse matrix $\boldsymbol{S}_w^{-1}$, where K denotes the initial feature vector dimension.

Thus, the common practice is, prior to LDA, to use an intermediate dimension reduction technique such as PCA that does not involve a discrimination analysis. This intermediate reduction is also preferable to reduce the computational complexity of the LDA analysis. In this regard, the Bayesian analysis that we propose in Section IV-A, can also serve as an intermediate dimension reduction method that selects a discriminative set of features from a larger set of DCT coefficients including some non-principle (or minor) feature components at each time instant.

As shown in Fig. 1, the Bayesian discrimination analysis results in a feature vector $\tilde{\boldsymbol{f}}(t)$ for each time instant t. Prior to concatenation within a window, the feature vector $\tilde{\boldsymbol{f}}(t)$ is linearly interpolated in time by some factor whose value depends on the frame rate. In the interpolated temporal domain, each feature vector at time instant t is concatenated with the previous and the next T feature vectors, so as to form a new higher dimensional feature vector that we denote by $\boldsymbol{F}(t)$

$$\boldsymbol{F}(t) = [\tilde{\boldsymbol{f}}(t-T), \tilde{\boldsymbol{f}}(t-T+1), \ldots, \tilde{\boldsymbol{f}}(t), \ldots, \tilde{\boldsymbol{f}}(t+T-1), \tilde{\boldsymbol{f}}(t+T)]. \tag{14}$$

The LDA analysis is then performed on this concatenated vector of dimension $(2T+1)N$. The dimension of the resulting dis-

criminative feature space is bounded above by $R-1$, that is one less than the total number of classes. Fig. 1 illustrates the formation of the final feature vector, that we will denote by $LDA(\tilde{\boldsymbol{f}}(t))$, via temporal and spatial discrimination analysis.

V. Experimental Results

Speaker identification and speech-reading experiments have been conducted using the MVGL-AVD audio-visual databases [42]. We have two distinct databases, the name ($\mathcal{D}_n$) and the digit ($\mathcal{D}_d$) datasets, each containing frontal views of 50 speakers ($R = 50$). The video frames are 720×576 pixels at a rate of 15 fps with 24 bit/pixel color. In the name dataset, each subject utters ten repetitions of her/his name as the secret phrase. A set of impostor data is also collected with each subject in the population uttering five different names from the population. In the digit dataset, each subject utters ten repetitions of a fixed digit password 348–572. All experiments have been conducted for three cases: 1) Speaker identification under name scenario using $\mathcal{D}_n$, 2) Speaker identification under digit scenario using $\mathcal{D}_d$, and 3) Speech-reading. The speech-reading dataset $\mathcal{D}_s$ is a subset of $\mathcal{D}_n$, that includes at least 12 repetitions of each name utterance. Hence this experiment addresses a limited vocabulary speech-reading application.

In order to extract lip motion features, first an initial lip region of size 128×80 is segmented from each video frame, following registration of successive face regions by global motion compensation. For grid-based motion analysis, a rectangular grid of size $G_x \times G_y = 64 \times 40$ is used for each lip segment. Following motion estimation and 2-D-DCT, a feature vector of size $2M$, is obtained by interlacing M features from x direction and M features from y direction, where $M = 50$ is used in the experiments. Then, the *FirstN* features, $\boldsymbol{f}^N_{\mathrm{GRD}}$, are extracted by eliminating some high-indexed DCT coefficients to obtain a vector of size N, where $N \leq 2M$. For contour-based motion analysis, we follow a similar procedure. First, the lip contour is extracted in each frame with the method described in Section III-C1. Following motion estimation and 1-D-DCT on the lip contour pixel locations, a feature vector of size $2M$, is obtained, where M is set to 50, i.e., the same number as in grid-based motion analysis. The low-indexed DCT coefficients then provide us with the contour-based *FirstN* features, $\boldsymbol{f}^N_{\mathrm{CTR}}$. The third lip feature representation is obtained by concatenating the contour-based motion features with the eight lip shape parameters, that is $\boldsymbol{f}^N_{\mathrm{CTR}} + \boldsymbol{f}_{\mathrm{SHP}}$.

The experimental results are presented with the following organization. In Section V-A, we present an evaluation of various lip motion features, $\boldsymbol{f}^N_{\mathrm{GRD}}$, $\boldsymbol{f}^N_{\mathrm{CTR}}$, and $\boldsymbol{f}^N_{\mathrm{CTR}} + \boldsymbol{f}_{\mathrm{SHP}}$ with $5 \leq N \leq 50$ for the cases of speaker identification under name scenario, speaker identification under digit scenario, and speech-reading. In Section V-B, we compare the results of the proposed Bayesian discrimination, DiscrimN, for feature reduction with those of the FirstN (PCA) analysis with or without LDA. We also demonstrate that the proposed discrimination measure (12) correlates well with the equal error rate (EER). Finally, Section V-C shows that the fusion of lip-motion and lip-intensity features provides improved performance compared to intensity-only features.

A. Evaluation of Various Lip Motion Features

1) Speaker Identification: Name Scenario: In the name scenario implementation, the $\mathcal{D}_n$ database is partitioned into two disjoint sets, $\{\mathcal{D}_{n_1} \text{ and } \mathcal{D}_{n_2}\}$, each having five repetitions from each subject in the database. The subset $\mathcal{D}_{n_1}$ is used for training, and $\mathcal{D}_{n_2}$ is used for testing. Since there are 50 subjects and five repetitions for each true and imposter client tests, the total number of trials for the true accepts and true rejects is respectively $N_a = 250$ and $N_r = 250$.

The three lip motion feature representations, $\boldsymbol{f}^N_{\mathrm{GRD}}$, $\boldsymbol{f}^N_{\mathrm{CTR}}$ and $\boldsymbol{f}^N_{\mathrm{CTR}} + \boldsymbol{f}_{\mathrm{SHP}}$, are tested on the database. Fig. 5(a) displays the EER performances with varying feature dimension N. We observe that the grid-based motion features, $\boldsymbol{f}^N_{\mathrm{GRD}}$, achieve 6.8% EER, and outperform the contour-based features. We also observe that addition of lip shape features, $\boldsymbol{f}_{\mathrm{SHP}}$, to the contour-based motion features, $\boldsymbol{f}^N_{\mathrm{CTR}}$, results in additional performance gain.

2) Speaker Identification: Digit Scenario: In the digit scenario, the $\mathcal{D}_d$ database is partitioned into two disjoint sets, $\{\mathcal{D}_{d_1} \text{ and } \mathcal{D}_{d_2}\}$, each having five repetitions of the same 6-digit number from each subject in the database. Again, $\mathcal{D}_{d_1}$ is used for training and $\mathcal{D}_{d_2}$ is used for testing. Note that, in the digit scenario, no imposter recordings are performed since every subject utters the same 6-digit number. Hence, the imposter clients are generated by the *leave-one-out* scheme, where each subject becomes the imposter of the remaining $R-1$ subjects in the population. Having $R = 50$ subjects and five testing repetitions, the resulting total number of trials for the true accepts and true rejects (imposters) becomes respectively $N_a = 250$ and $N_r = 250$.

Fig. 5(b) shows the EER performances for different lip motion representations with varying feature dimension N. We observe that the grid-based motion features, $\boldsymbol{f}^N_{\mathrm{GRD}}$, and the contour-based motion with shape features, $\boldsymbol{f}^N_{\mathrm{CTR}} + \boldsymbol{f}_{\mathrm{SHP}}$, achieve the same minimum 12.8% EER, and outperform the contour-based only features $\boldsymbol{f}^N_{\mathrm{CTR}}$. Note that the EER performance of speaker identification under the name scenario is significantly better than that of the digit scenario. This is as expected since in the name scenario each speaker in the database utters a different person-specific phrase, making the identification task easier.

3) Speech-reading Scenario: The speech reading database $\mathcal{D}_s$ is constructed as a subset of speaker identification database for the name scenario. It includes 35 different phrases, i.e., $R = 35$. Each phrase is a name from the name database with twelve repetitions. The number of source speakers for each phrase varies, we have at least four and at most seven speakers. The $\mathcal{D}_s$ database is partitioned into two disjoint sets $\mathcal{D}_{s_1}$ and $\mathcal{D}_{s_2}$, one for training and the other for testing, each having six utterance repetitions. Fig. 5(c) displays the recognition rates for different lip motion representations with varying feature dimension N. We observe that the contour-based motion combined with shape features, $\boldsymbol{f}^N_{\mathrm{CTR}} + \boldsymbol{f}_{\mathrm{SHP}}$, achieve the best recognition rate, 70.48%. However, the contour-based only, $\boldsymbol{f}^N_{\mathrm{CTR}}$, and the grid-based motion features, $\boldsymbol{f}^N_{\mathrm{GRD}}$, perform quite close to this best recognition rate, yielding 69.52% and 67.62%, respectively.

Furthermore, under all scenarios, as observed in Fig. 5, the performance of contour-based motion features starts to degrade

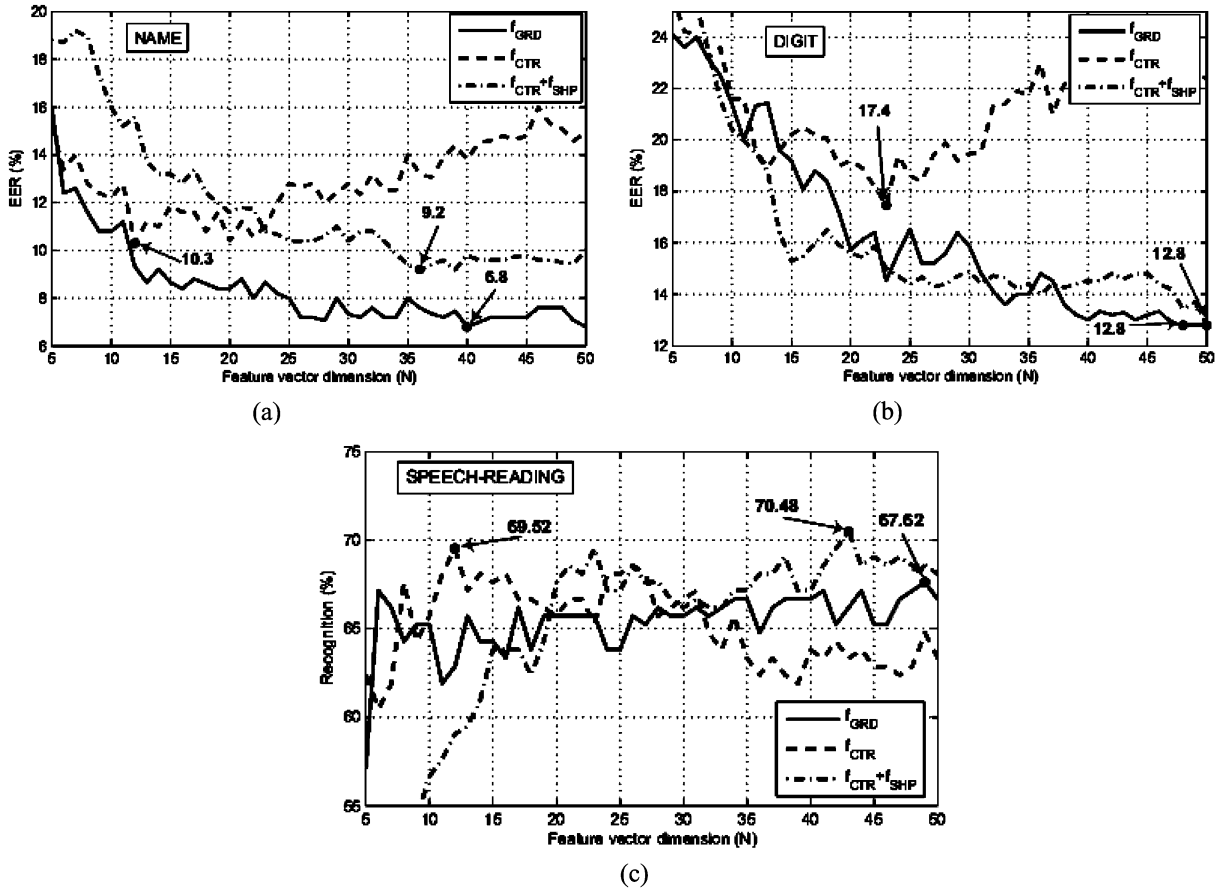


Fig. 5. Speaker identification and speech-reading results for grid-based motion, $\boldsymbol{f}^N_{\mathrm{GRD}}$, contour-based motion, $\boldsymbol{f}^N_{\mathrm{CTR}}$, and contour-based motion combined with shape, $\boldsymbol{f}^N_{\mathrm{CTR}} + \boldsymbol{f}_{\mathrm{SHP}}$.

after a certain value of the feature dimension N. This is because the length of the lip contour can adequately be sampled by this value of N, and there is no further gain from increasing the dimension N.

B. Evaluation of Bayesian Discrimination Analysis

This subsection compares the results of the proposed Bayesian discrimination analysis, DiscrimN, for feature reduction with those of the FirstN (PCA) analysis with or without temporal LDA, considering all lip motion features discussed in the previous subsection. The EER or recognition rates obtained by selecting the feature dimension N as the one that maximizes the performance for each case are provided in Table I. In Table I, $\boldsymbol{f}^N$ and $\tilde{\boldsymbol{f}}^N$ stand for the *FirstN* and *DiscrimN* features, whereas $LDA(\boldsymbol{f}^N)$ and $LDA(\tilde{\boldsymbol{f}}^N)$ denote the features obtained by applying the temporal LDA using $T = 6$ as the temporal window parameter. The best performance rate for each scenario is indicated in bold in the table. The best EER rate attained for speaker identification is 5.2% under both name and digit scenarios after two-stage discrimination, whereas the best recognition rate for speech-reading, 72.86%, is achieved using Bayesian discrimination alone. Note that the temporal LDA brings significant performance gain in speaker identification especially under the digit scenario. On the other hand, the Bayesian discriminative feature selection method, when used alone, yields performance gain in all scenarios. Also note that the use of lip shape parameters in addition to contour-based motion features improves the performance to 9.2% and 8.8% EER in name and digit scenarios, respectively, and to 70.48% recognition rate in speech-reading.

TABLE I
EVALUATION OF THE PROPOSED TWO-STAGE DISCRIMINATION ANALYSIS FOR SPEAKER IDENTIFICATION AND SPEECH-READING

Feature Set	EER (%)		Recog. Rate (%)
	Name	Digit	Speech
$\boldsymbol{f}^N_{GRD}$	6.8	12.8	67.62
$\tilde{\boldsymbol{f}}^N_{GRD}$	6.5	12.2	**72.86**
$LDA(\boldsymbol{f}^N_{GRD})$	5.6	5.8	67.14
$LDA(\tilde{\boldsymbol{f}}^N_{GRD})$	**5.2**	**5.2**	67.62
$\boldsymbol{f}^N_{CTR}$	10.3	17.4	69.52
$\tilde{\boldsymbol{f}}^N_{CTR}$	9.8	17.6	70.00
$LDA(\tilde{\boldsymbol{f}}^N_{CTR})$	12.0	18.88	60.95
$\boldsymbol{f}_{SHP}$	18.9	23.5	51.43
$\boldsymbol{f}^N_{CTR} + \boldsymbol{f}_{SHP}$	9.2	12.8	70.48
$\tilde{\boldsymbol{f}}^N_{CTR} + \boldsymbol{f}_{SHP}$	9.4	13.8	69.52
$LDA(\tilde{\boldsymbol{f}}^N_{CTR} + \boldsymbol{f}_{SHP})$	10.4	8.8	61.90

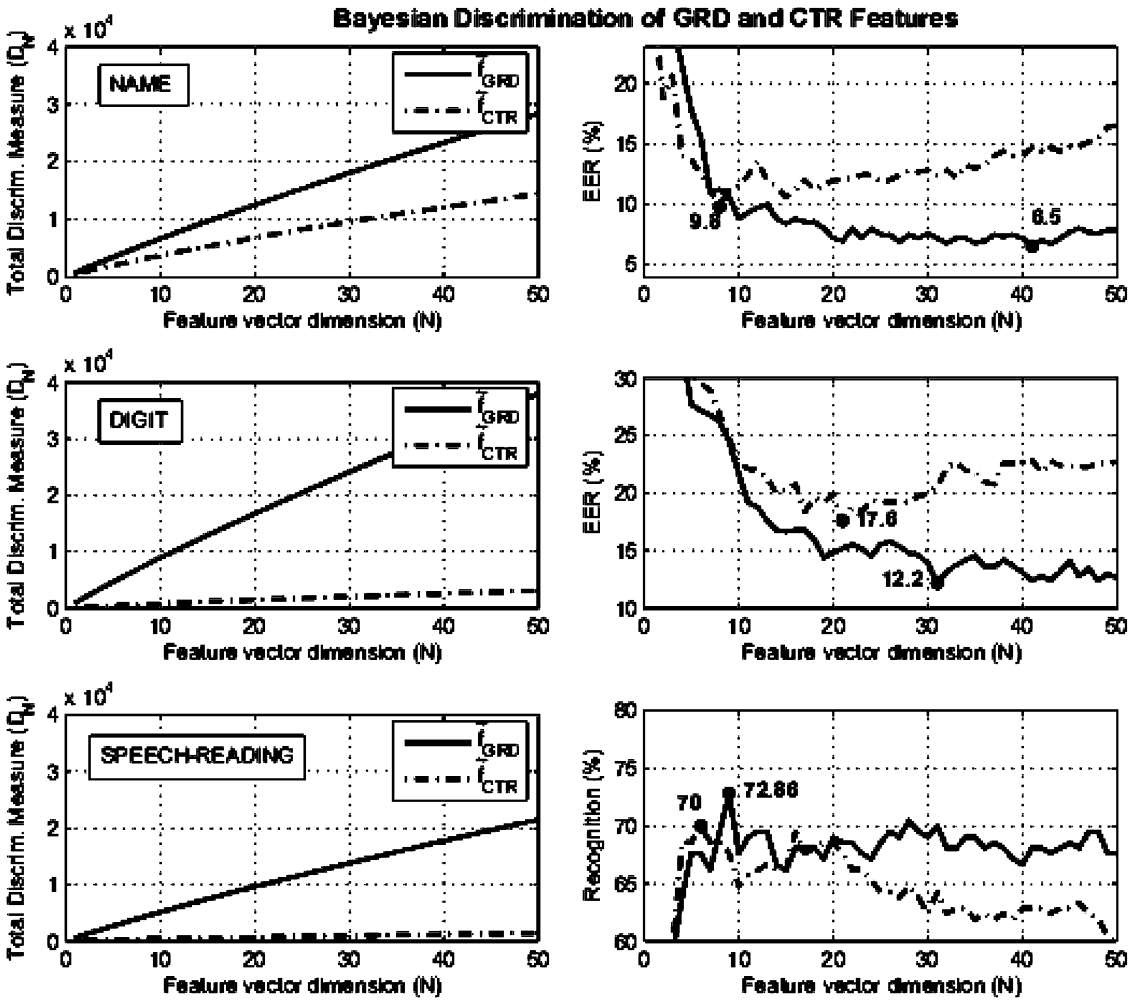


Fig. 6. Total discrimination measures (D_N) and the corresponding experimental performances of grid-based *DiscrimN* ($\bar{\boldsymbol{f}}_{\mathrm{GRD}}$) and contour-based *DiscrimN* ($\boldsymbol{f}_{\mathrm{CTR}}$) motion features with varying dimension N for name, digit, and speech-reading scenarios.

In Table I, we observe that the best performances are obtained using the grid-based motion features for both speaker identification and speech-reading. The key observations of these experiments are the following.

- DiscrimN achieves the same or better performance at relatively lower dimensions by selecting a discriminative subset of coefficients, which are not necessarily the principle components.
- The use of temporal LDA in addition to Bayesian discrimination, brings additional EER gain in speaker identification with grid-based features.

In Table I, we also observe that the discrimination analysis with or without LDA, when applied to contour-based motion features, does not always perform well and improve the performance as expected. There seem to be two main reasons for this poor performance. First, the dimension of contour-based motion vectors (computed only along the lip contour) is much smaller than that of the grid-based motion vectors, and second, these motion vectors are noisier due to possible tracking failures, which makes it more difficult to capture temporal correlations. This observation leads us to the conclusion that grid-based motion features are more robust and effective than contour-based motion features for use in recognition applications. Another observation is that the temporal LDA may degrade the recognition rate in speech reading even with grid-based motion features, which is likely to be due to the fact that our application addresses a word-level speech reading problem.

In Fig. 6, we observe that the total discrimination measure $D_N(\boldsymbol{f})$, defined by (12), correlates well with the relative recognition performances of different lip feature representations. The left column of Fig. 6 presents the total discrimination measures of the *DiscrimN* features of these two representations, and the right column presents the corresponding experimental EER and recognition rates for speaker identification and speech-reading. We observe that the numerical discrimination measures and the corresponding experimental performances match each other; that is, the higher the discriminative power for a given feature representation, the higher is the corresponding recognition performance.

Finally, we note that the Bayesian discrimination analysis specifies a reordering (ranking) of the transform domain coefficients and the first N of these coefficients are selected as the most discriminative features. Fig. 7 shows the discrimination values of the selected 50 coefficients out of 100 (50 from x and 50 from y directions, respectively) for different recognition scenarios in the case of grid-based motion feature vector

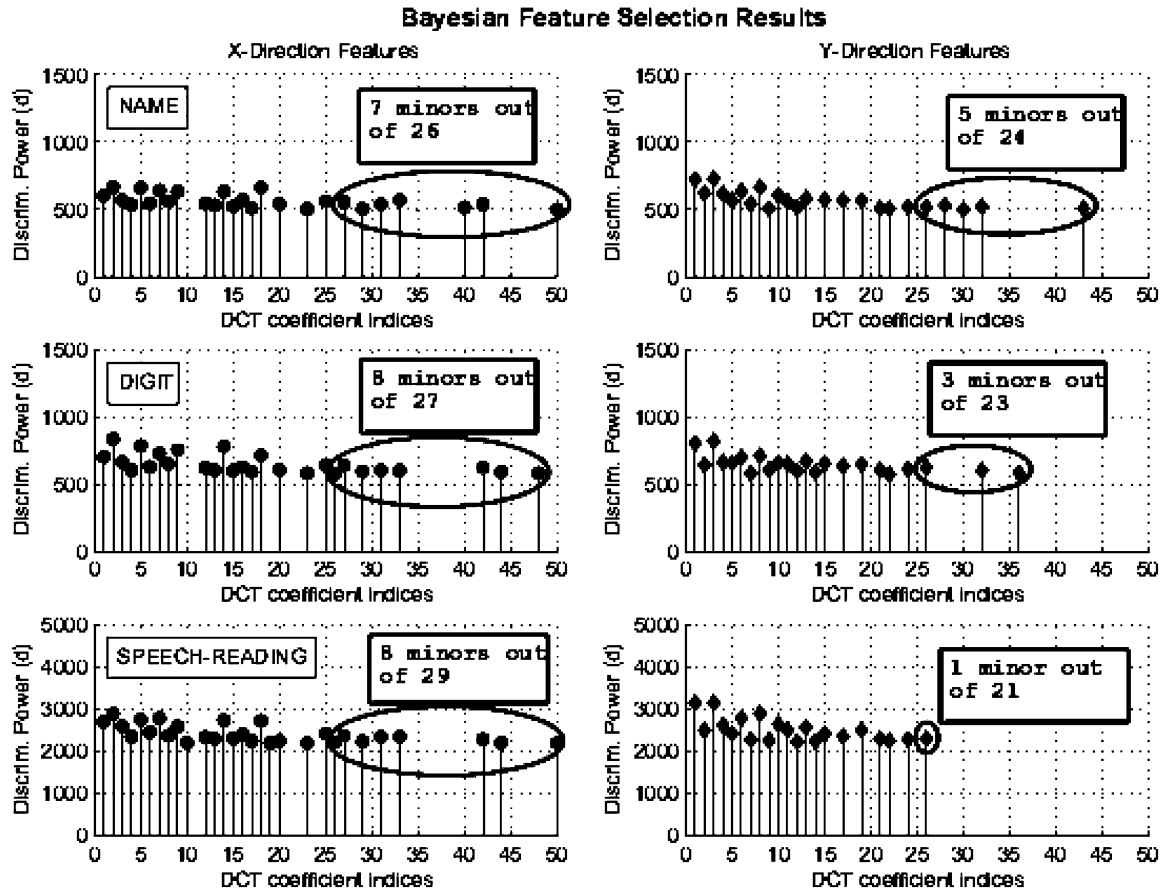


Fig. 7. DCT coefficients selected by Bayesian discrimination analysis in the case of grid-based lip motion for $N = 50$. Minor coefficients are defined as those with index greater than 25.

$\boldsymbol{f}_{\mathrm{GRD}}$. We note that the FirstN (PCA) feature selection procedure (with $N = 50$) would have selected the first 25 from x and y directions, respectively. We observe that the Bayesian feature selection procedure selects some feature coefficients whose index is greater than 25 in all cases. Moreover, the number of selected coefficients whose index is greater than 25 is higher in the speaker identification application, which indicates that higher frequency coefficients are more valuable for speaker identification than for speech-reading. One can also note that the number of coefficients with index greater than 25 is higher for the x component of the lip motion vectors than for the y component.

C. Combining Motion and Intensity Information

We have performed experiments to determine whether using explicit lip motion features, instead of or in addition to lip intensity information, provides further performance gain. Following the common practice of other lip-based recognition systems such as [3] and [14], we form the intensity-based lip feature vector by scanning the 2-D-DCT coefficients in the zig-zag order, that are computed from the raw intensity values within the rectangular lip region. The best performance rates achieved with intensity-only and motion-only features are presented in Table II for speaker identification (name and digit) and speech-reading. The last row of Table II displays the corresponding performance rates when lip motion is combined with lip intensity by using the decision fusion scheme, the reliability weighted summation, proposed in [14]. We use the best grid-based lip motion features for each scenario and the *DiscrimN* features to represent intensity information without any further temporal discrimination as they yield the best performance in all scenarios. We observe that the addition of intensity information yields a significantly higher performance gain in the case of speaker identification under the digit scenario as compared to the other lip-based scenarios and representations. This is mostly due to the texture information conveyed in the intensity-based lip features. The texture serves as an important discriminative information especially under the digit scenario since the imposters of this scenario are

TABLE II
SPEAKER IDENTIFICATION AND SPEECH-READING PERFORMANCE RESULTS FOR INTENSITY-ONLY FEATURES, MOTION-ONLY FEATURES, AND THEIR DECISION FUSION

Feature Type	EER (%)		Recog. Rate (%)
	Name	Digit	Speech
Intensity	5.6	1.74	62.86
Motion	5.2	5.2	72.86
Intensity⊕Motion	3.6	1.6	70.95

generated by the *leave-one-out* scheme and thus not registered in the population. It is also as expected to observe that, for speech-reading, the lip motion features perform better than the intensity-based lip features since the speech information is strongly coupled with the lip movement and thus better represented with motion-based features. The use of lip intensity information in addition to lip motion does not neither improve the performance in the case of speech-reading.

VI. Conclusion

In this paper, we have investigated different lip motion representations and proposed a two-stage discriminative lip feature selection method for speaker identification and speech-reading. We have shown by experiments that, for speaker/speech recognition:

- explicit lip motion is useful in addition to lip intensity and/or geometry;
- grid-based dense lip motion features are superior and more robust compared to contour-based lip motion features.

The best equal error rate and recognition rate performances of the grid-based lip motion features are reported as 5.2% and 72.86%, respectively, for speaker identification and speech-reading, whereas its fusion with lip intensity provides additional performance gain only in speaker identification, the EER rate being improved to 3.6% and 1.6% under the name and digit scenarios, respectively. The lip motion is found to be more valuable than the lip intensity for speech-reading.

We have proposed a two-stage discrimination analysis technique that involves the spatial Bayesian feature selection and the temporal LDA. The experimental results reveal that the Bayesian discrimination analysis improves the performance in both speaker identification and speech-reading. The Bayesian discriminative feature selection serves also as an intermediate dimension reduction step prior to the temporal LDA, by successfully selecting the lip features that are tailored for the specific recognition problem. The temporal LDA seems beneficial for speaker identification, especially under the digit scenario.

Acknowledgment

The authors would like to thank to N. Eveno, A. Caplier, and P.-Y. Coulon for their valuable help in implementation of the lip tracking module in our recognition system.

References

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

[2] M. J. Tomlinson, M. J. Russell, and N. M. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 1996, vol. II, pp. 821–824.

[3] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audio-visual speech," *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1326, Sep. 2003.

[4] S. W. Foo, Y. Lian, and L. Dong, "Recognition of visual speech elements using adaptively boosted hidden Markov models," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 5, pp. 693–705, May 2004.

[5] T. Chen, "Audiovisual speech processing," *IEEE Signal Process. Mag.*, vol. 18, pp. 9–21, Jan. 2001.

[6] S. L. Wang, W. H. Lau, S. H. Leung, and H. Yan, "A real-time automatic lipreading system," in *Proc. 2004 Int. Symp. Circuits and Systems*, 2004, vol. 2, pp. 101–104.

[7] L. G. D. Silveira, J. Facon, and D. L. Borges, "Visual speech recognition: a solution from feature extraction to words classification," in *Proc. XVI Brazilian Symp. Computer Graphics and Image Processing*, 2003, pp. 399–405.

[8] M. N. Kaynak, Q. Zhi, A. D. Cheok, K. Sengupta, Z. Jian, and K. C. Chung, "Analysis of lip geometric features for audio-visual speech recognition," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 34, no. 4, pp. 564–570, Jul. 2004.

[9] X. Zhang, C. C. Broun, R. M. Mersereau, and M. A. Clements, "Automatic speechreading with applications to human-computer interfaces," *EURASIP J. Appl. Signal Process.*, pp. 1228–1247, 2002.

[10] J. F. G. Perez, A. F. Frangi, E. L. Solano, and K. Lukas, "Lip reading for robust speech recognition on embedded devices," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 2005, vol. I, pp. 473–476.

[11] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 198–213, Feb. 2002.

[12] P. S. Aleksic, J. J. Williams, Z. Wu, and A. K. Katsaggelos, "Audio-visual speech recognition using MPEG-4 compliant visual features," *EURASIP J. Appl. Signal Process.*, pp. 1213–1227, 2002.

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

[14] E. Erzin, Y. Yemez, and A. M. Tekalp, "Multimodal speaker identification using an adaptive classifier cascade based on modality reliability," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 840–852, Oct. 2005.

[15] N. A. Fox and R. B. Reilly, "Robust multi-modal person identification with tolerance of facial expression," in *IEEE Int. Conf. Systems, Man and Cybernetics*, 2004, vol. 1, pp. 580–585.

[16] C. C. Broun, X. Zhang, R. M. Mersereau, and M. Clements, "Automatic speechreading with application to speaker verification," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 2002, vol. I, pp. 685–688.

[17] L. L. Mok, W. H. Lau, S. H. Leung, S. L. Wang, and H. Yan, "Lip features selection with application to person authentication," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 2004, vol. III, pp. 397–400.

[18] T. Wark and S. Sridharan, "Adaptive fusion of speech and lip information for robust speaker identification," *Digital Signal Process.*, vol. 11, no. 3, pp. 169–186, Jul. 2001.

[19] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner, "Acoustic-labial speaker verification," *Pattern Recognit. Lett.*, vol. 18, no. 9, pp. 853–858, 1997.

[20] B. Fröba, C. Rothe, and C. Küblbeck, "Evaluation of sensor calibration in a biometric person recognition framework based on sensor fusion," in *Proc. 4th IEEE Int. Conf. Automatic Face and Gesture Recognition*, Mar. 2000, pp. 512–517.

[21] R. W. Frischholz and U. Dieckmann, "Bioid: A multimodal biometric identification system," *IEEE Computer*, vol. 33, no. 2, pp. 64–68, Feb. 2000.

[22] H. E.Çetingül, Y. Yemez, E. Erzin, and A. M. Tekalp, "Discriminative lip-motion features for biometric speaker identification," in *Proc. Int. Conf. on Image Processing*, Oct. 2004, pp. 2023–2026.

[23] ——, "Robust lip-motion features for speaker identification," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Mar. 2005, vol. I, pp. 509–512.

[24] J. P. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep. 1997.

[25] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Commun.*, vol. 17, pp. 91–108, 1995.

[26] J.-M. Odobez and P. Bouthemy, "Robust multiresolution estimation of parametric motion models," *J. Vis. Commun. Image Represent.*, vol. 6, no. 4, pp. 348–365, Dec. 1995.

[27] Y. Yemez, B. Sankur, and E. Anarım, "A quadratic motion-based object-oriented video codec," *Signal Process.: Image Commun.*, vol. 15, pp. 729–766, 2000.

[28] A. Puri, X. Chen, and A. Luthra, "Video coding using the H.264/MPEG-4 AVC compression standard," *Signal Process.: Image Commun.*, vol. 19, pp. 793–849, 2004.

[29] M. Sadeghi, J. Kittler, and K. Messer, "Modelling and segmentation of lip area in face images," *IEE Proc. Vis. Image Signal Process.*, vol. 149, no. 3, pp. 179–184, Jun. 2002.

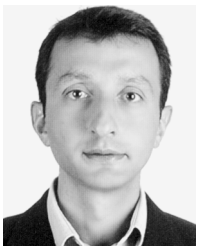
[30] S.-H. Leung, S.-L. Wang, and W.-H. Lau, "Lip Iimage segmentation using fuzzy clustering incorporating an elliptic shape function," *IEEE Trans. Image Process.*, vol. 13, no. 1, pp. 51–62, Jan. 2004.

[31] T. Wakasugi, M. Nishiura, and K. Fukui, "Robust lip contour extraction using separability of multi-dimensional distributions," in *Proc. 6th IEEE Int. Conf. Automatic Face and Gesture Recognition*, May 2004, pp. 415–420.

[32] S. L. Wang, W. H. Lau, S. H. Leung, and A. W. C. Liew, "Lip segmentation with the presence of beards," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 2004, vol. III, pp. 529–532.
[33] N. Eveno, A. Caplier, and P.-Y. Coulon, "Accurate and quasi-automatic Lip tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 5, pp. 706–715, May 2004.
[34] X. Zhang, R. M. Mersereau, M. A. Clements, and C. C. Broun, "Visual speech feature extraction for improved speech recognition," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 2002, pp. 1993–1996.
[35] A. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, "A couple hmm for audio-visual speech recognition," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 2002, pp. 2013–2016.
[36] J. Luettin, N. Thacker, and S. Beet, "Statistical lip modelling for visual speech recognition," in *Proc. 8th Eur. Signal Processing Conf.*, 1996, pp. 10–13.
[37] B. Lucas and T. Kanade, "An iterative image restoration technique with an application to stereo vision," in *Proc. DARPA IU Workshop*, 1981, pp. 121–130.
[38] C. C. Chibelushi, J. S. Mason, and F. Deravi, "Integration of acoustic and visual speech for speaker recognition," in *Proc. 3rd Eur. Conf. Speech Communication and Technology*, 1993, vol. 1, pp. 157–160.
[39] T. J. Wark, S. Sridharan, and V. Chandran, "An approach to statistical lip modeling for speaker identification via chromatic feature extraction," in *Proc. Int. Conf. Pattern Recognition*, 1998, vol. 1, pp. 123–125.
[40] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Englewood Cliffs: Prentice-Hall, 2001.
[41] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 228–233, Feb. 2001.
[42] E. Erzin, Y. Yemez, and A. M. Tekalp, *DSP in Mobile and Vehicular Systems*. New York: Springer Verlag, 2005, ch. Joint Audio-Video Processing for Robust Biometric Speaker Identification in Car.

H. Ertan Çetingül (S'03) received the B.Sc. degree in electrical and electronics engineering (with a minor degree in general management) from Middle East Technical University, Ankara, Turkey, in 2003, and the M.Sc. degree in electrical and computer engineering from Koç University, Istanbul, Turkey, in 2005. He is currently pursuing the Ph.D. degree in biomedical engineering at The Johns Hopkins University, Baltimore, MD.

His research interests include computer vision, pattern recognition, medical image processing, and machine learning.



Yücel Yemez (M'03) received the B.S. degree from Middle East Technical University, Ankara, Turkey, in 1989, and the M.S. and Ph.D. degrees from Bogaziçi University, Istanbul, Turkey, respectively, in 1992 and 1997, all in electrical engineering.

From 1997 to 2000, he was a Postdoctoral Researcher in the Image and Signal Processing Department, Télécom Paris (Ecole Nationale Supérieure des Télécommunications). Currently, he is an Assistant Professor in the Computer Engineering Department at Koç University, Istanbul. His current research is focused on various fields of computer vision and graphics.

Engin Erzin (S'88–M'96) received the B.Sc., M.Sc., and Ph.D. degrees from Bilkent University, Ankara, Turkey, in 1990, 1992, and 1995, respectively, all in electrical engineering.

From 1995 to 1996, he was a Postdoctoral Fellow in the Signal Compression Laboratory, University of California, Santa Barbara. He joined Lucent Technologies in September 1996, and was with Consumer Products for one year as a Member of Technical Staff of the Global Wireless Products Group. From 1997 to 2001, he was with the Speech and Audio Technology Group of the Network Wireless Systems. Since January 2001, he has been with the Electrical and Electronics Engineering and Computer Engineering Departments, Koç University, Istanbul, Turkey. His research interests include speech signal processing, pattern recognition, and adaptive signal processing.

A. Murat Tekalp (S'80–M'84–SM'91–F'03) received the M.S. and Ph.D. degrees in electrical, computer, and systems engineering from Rensselaer Polytechnic Institute (RPI), Troy, NY, in 1982 and 1984, respectively.

He has been with Eastman Kodak Company, Rochester, NY, from December 1984 to June 1987, and with the University of Rochester from July 1987 to June 2005, where he was promoted to Distinguished University Professor. Since June 2001, he has been a Professor at Koç University, Istanbul, Turkey. His research interests are in the area of digital image and video processing, including video compression and streaming, motion-compensated video filtering for high-resolution, video segmentation, content-based video analysis and summarization, 3DTV/video processing and compression, multi-camera surveillance video processing, and protection of digital content. He authored the book *Digital Video Processing* (Englewood Cliffs, NJ: Prentice-Hall, 1995) and holds seven U.S. patents. His group contributed technology to the ISO/IEC MPEG-4 and MPEG-7 standards.

Dr. Tekalp was named Distinguished Lecturer by the IEEE Signal Processing Society in 1998, and awarded a Fulbright Senior Scholarship in 1999. He received the TUBITAK Science Award (highest scientific award in Turkey) in 2004. He chaired the IEEE Signal Processing Society Technical Committee on Image and Multidimensional Signal Processing (January 1996–December 1997). He served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING (1990 to 1992) and the IEEE TRANSAACTIONS ON IMAGE PROCESSING (1994 to 1996), and the Kluwer journal *Multidimensional Systems and Signal Processing* (1994 to 2002). He was an Area Editor for the Academic Press Journal *Graphical Models and Image Processing* (1995 to 1998). He was also on the Editorial Board of the Academic Press journal *Visual Communication and Image Representation* (1995 to 2002). He was appointed as the Special Sessions Chair for the 1995 IEEE International Conference on Image Processing, the Technical Program Co-Chair for IEEE ICASSP 2000 in Istanbul, the General Chair of IEEE International Conference on Image Processing (ICIP) in Rochester in 2002, and Technical Program Co-Chair of EUSIPCO 2005 in Antalya, Turkey. He is the Founder and First Chairman of the Rochester Chapter of the IEEE Signal Processing Society. He was elected as the Chair of the Rochester Section of IEEE for 1994 to 1995. At present, he is the Editor-in-Chief of the EURASIP journal *Signal Processing: Image Communication* (Elsevier). He is serving as the Chairman of the Electronics and Informatics Group of the Turkish Science and Technology Foundation (TUBITAK) and as an independent expert to review projects for the European Commission.