DISCRIMINATION ANALYSIS OF LIP MOTION FEATURES FOR MULTIMODAL SPEAKER IDENTIFICATION AND SPEECH-READING

by

Hasan Ertan Çetingül

A Thesis Submitted to the
Graduate School of Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Electrical & Computer Engineering

Koç University

July, 2005

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Hasan Ertan Çetingül

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Prof. A. Murat Tekalp

_____

Assoc. Prof. Levent M. Arslan

_____

Assist. Prof. Engin Erzin

_____

Assist. Prof. Yücel Yemez

_____

Assist. Prof. Alper T. Erdoğan

Date: _____

To my parents and my fiancée

ABSTRACT

In this thesis a new multimodal speaker/speech recognition system that integrates audio, lip texture, lip geometry, and lip motion modalities is presented. There have been several studies that jointly use audio, lip intensity and/or lip geometry information for speaker identification and speech recognition applications. This work proposes using explicit lip motion information, instead of or in addition to audio, 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: i) Is using explicit lip motion information useful? and ii) 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. The audio modality is represented by the well-known mel-frequency cepstral coefficients (MFCC) along with the first and second derivatives, whereas lip texture modality is represented by the 2D-DCT coefficients of the luminance component within a bounding box about the lip region. Several lip motion feature candidates are considered including dense motion features within a bounding box around the lip, lip contour motion features, lip shape features, and combinations of them. Furthermore, a novel two-stage discriminant analysis is introduced to select the best lip motion features for speaker identification and speech-reading applications. The fusion of audio, lip texture and lip motion modalities is performed by the so-called Reliability Weighted Summation (RWS) decision rule. Experimental results show that the proposed discriminative analysis significantly improves the unimodal performance of the lip motion modality. Moreover, using explicit lip motion information in addition to audio and lip texture yields further performance gains in bimodal speaker/speech recognition systems.

ÖZETÇE

Bu tezde ses, dudak dokusu, dudak geometrisi ve dudak devinimlerini birleştiren yeni bir çok-kipli konuşmacı/konuşma tanıma sistemi sunulmaktadır. Konuşmacı ve konuşma tanıma uygulamalarında ses, dudak yeğinliği ve/veya dudak geometri bilgisini beraber kullanan birkaç çalışma mevcuttur. Bu çalışmada konuşmacı tanıma ve konuşma okuma için, ses, dudak yeğinlik ve/veya geometri bilgisi ile birlikte ya da bu bilgilerin yerine, açık dudak devinim bilgisinin kullanımı önerilmekte; konu öznitelik seçimi ile ayırım analizi çerçevesinde incelenmektedir. Çalışma iki önemli soruya cevap aramaktadır: i) Açık dudak devinim bilgisi yararlı mıdır? ve ii) Devinim bilgisi yararlı ise, sözü edilen uygulamalarda eniyi dudak devinim öznitelikleri nelerdir? Konuşmacılar arasında en yüksek ayrımı sağlayan öznitelikler, konuşmacı tanıma probleminde eniyi dudak devinim öznitelikleri olmakla beraber konuşma okumada eniyi öznitelikler, en yüksek fonem/kelime/deyiş tanıma oranına erişenlerdir. Ses doruğu, mel frekans kepstral katsayıları ile katsayıların birinci ve ikinci türevleriyle gösterilirken, dudak doku kipi, dudak bölgesinin yeğinlik değerlerinin 2B-AKD (Ayrık Kosinüs Dönüşümü) katsayıları ile ifade edilmektedir. Birden çok dudak devinim öznitelik adayı ele alınmaktadır: dudak bölgesi içinde ızgara-tabanlı yoğun devinim öznitelikleri, dudak çevriti üzerinde devinim öznitelikleri ve son olarak dudak şekil parametreleri ile bunların bileşimleri. Buna ek olarak, konuşmacı tanıma ve konuşma okumada eniyi dudak devinim öznitelik lerini belirlemek üzere iki basamaklı yeni bir ayrımsama analizi tanıtılmaktadır. Ses, dudak dokusu ve dudak devinim kiplerinin tümleştirilmesi *Güvenilirlik Ağırlıklı Toplama* karar kuralıyla gerçekleştirilmiştir. Deneysel sonuçlarda, önerilen ayırımsal analizin dudak deviniminin tek-kipli başarımını oldukça geliştirdiği görülmektedir. Bunun yanında, ses ve dudak doku bilgisi ile birlikte açık dudak devinim bilgisinin kullanımı, iki-kipli konuşmacı/konuşma tanıma sistemlerinin başarımlarında ilave kazanım sağlamaktadır.

ACKNOWLEDGMENTS

First I would like to thank my supervisor Prof. A. Murat Tekalp, my co-advisor Assist. Prof. Engin Erzin and Assist. Prof. Yücel Yemez who have been a great source of inspiration and provided the right balance of suggestions, criticism, and freedom.

I am grateful to members of my thesis committee for critical reading of this thesis and for their valuable comments.

I would like to thank to Dr. Alice Caplier, Dr. Nicolas Eveno and Prof. Pierre-Yves Coulon for their valuable help in implementation of the lip tracking module in our recognition system.

I also would like to thank those people who have shared time, given acquisitions for our audio-visual database and edited recordings.

Finally I thank my family and my fiancée for providing me a morale support that helps me in hard days of my research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

DCT	Discrete Cosine Transform
EER	Equal Error Rate
FAR	False Acceptance Rate
FRR	False Reject Rate
GMM	Gaussian Mixture Models
HMM	Hidden Markov Model
HTK	Hidden Markov Model Toolkit
LS	Least-Squares
LDA	Linear Discriminant Analysis
MFCC	Mel-Frequency Cepstral Coefficient
PCA	Principal Component Analysis

Chapter 1

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: i) lip texture, ii) lip geometry (shape), and iii) lip motion features. The first represents lip movements implicitly along with texture information that might sometimes carry useful discrimination information; but in some other cases the texture 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 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.

It is worth noting that there are relatively less number of technologies employing the explicit lip motion in audio-visual recognition as compared to the lip texture and lip geometry. Thus, investigating the best lip motion features for unimodal speech-reading and speaker identification will definitely be an asset for the literature. However, it is generally agreed that no single technology, i.e., modality, will meet the needs of all potential recognition applications. Hence, integration of multiple modalities should also be attacked so as to obtain

improved recognition systems.

The design of a multimodal recognition system requires addressing three basic issues: i) Which modalities to fuse, ii) How to represent each modality with a discriminative and low-dimensional set of features, and iii) How to fuse existing modalities.

For the first issue, modalities to fuse, it is a fact that recent speaker/speech recognition technologies have generally employed available visual data together with audio. Audio is probably the most natural source to recognize what is uttered and a valuable source to identify a speaker [1]. The speech content and voice can be interpreted as two different, though correlated, modalities existing in audio signals. Likewise, video signal can be split into different modalities, such as face/lip texture and lip motion information that is correlated with the audio. The lip motion modality, whether implicit or explicit, has been extensively utilized in speech recognition systems but not so common in speaker recognition. The first reason for this is that the lip motion, or the lip modality in general, is not considered as the primary modality to be used in speaker recognition. The second reason is the sophisticated feature processing, which will be briefly explained while discussing the second issue.

The second issue, representative feature selection, also includes modeling of classifiers through which each class is represented with a statistical model or a representative feature set. The lip motion is known to be the least investigated modality in feature representation as compared to face and audio, for which there are well-known representations. The main reason for the lack of feature-level investigation of lip motion is, as stated before, the sophisticated feature processing to reveal biometrics. More specifically, as far as speech is concerned, it is usually sufficient to extract the principal components of the lip movement. However, the principal components of the lip movement are not usually sufficient to well discriminate the biometric properties of a speaker. High frequency or non-principal components of the signal should also be valuable especially when the objective is to model the biometrics. For speaker recognition, the best features are those that result in the highest discrimination of individual speakers in a population whereas for speech recognition, the features providing the highest phoneme/word/phrase recognition rate are considered as the best ones. Other than discrimination capability, curse of dimensionality, computational efficiency, robustness and invariance are other important criteria in selection of the feature set and the recognition methodology for each modality.

Audio-only speaker/speech recognition systems are far from being perfect especially under noisy conditions. Performance problems are also observed in video-only speaker/speech recognition systems, where poor picture quality, changes in pose and lighting conditions, and varying facial expressions may have detrimental effects [2, 3]. Hence, robust solutions for both speaker and speech recognition should employ multiple sources, such as audio, lip texture, and lip motion in a unified scheme. For the final issue, fusion problem, different strategies are possible: In the so-called *early integration*, modalities are fused at data or feature level, whereas in *late integration* decisions or scores resulting from each unimodal recognition are combined to give the final conclusion. Multimodal decision fusion can also be viewed from a broader perspective as a way of combining classifiers, which is a well studied problem in pattern recognition. The main motivation here is to compensate possible misclassification errors of a certain classifier with other available classifiers and to end up with a more reliable overall decision. Misclassification errors are in general inevitable due to numerous factors such as environmental noise, measurement and modeling errors or time-varying characteristics of signals. A comprehensive survey and discussion on classifier combination techniques can be found in [4].

In this thesis, we will develop a multimodal speaker/speech recognition scheme that uses an improved lip motion information together with lip texture and audio. The lip motion modality is improved by means of a novel two-stage discrimination analysis that selects best lip motion features. In the remaining part of this chapter, we will give a brief summary of the relevant past research and our contribution. Then in Chapter 2, we will develop a theoretical framework that the whole thesis work will be based on. Chapter 3 describes the lip tracking procedure together with the lip motion feature representation. The proposed two-stage discriminative lip motion feature extraction technique is presented in Chapter 4. The question "how to fuse" is addressed in Chapter 5 together with the other audio-visual modalities available. Experimental results will be presented in 2 different chapters: Chapter 6 outlines the unimodal system performance and discussions on the discrimination analysis, whereas in Chapter 7 multimodal system performances are presented. Finally the conclusions are given in Chapter 8.

1.1 State-of-the-art

In audio-visual speech recognition (speech-reading), lip texture information is widely used. In [5, 6], 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 DCT coefficients of the gray-scale lip image [7]. They then 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 [8, 9, 10, 11, 12, 13], since it is easier to match mouth openings-closings with the corresponding phonemes. Deformable templates [8, 9], active shape models (ASM) [13, 14, 15], and snakes [16] have been used to obtain different lip geometry features; however, they all suffer from complex feature extraction and training procedures. In [9], 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. [11] use horizontal/vertical distances along with the orientation angle to represent the 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 [17], the lip feature vector is formed by concatenating the Karhunen-Loève transformed inner-outer lip contour points with the texture information which is represented in a similar way as in the so-called *eigenlips* technique [5]. In [15], the geometric information extracted by active shape models is used along with the gray-level appearance features and then fused with audio for speech recognition. Perez et al. [14] 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. [16] use gradient vector flow (GVF) snakes to extract outer lip contour and calculate the lip movement at 10 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 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 [18, 19], 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 [20], 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. [21] 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 [22], 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 [23] 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 [24], 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 [25], the lip motion is represented by the full set of frequency-domain 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 was performed, and no specific attention was paid to optimize the unimodal performance of the lip motion modality.

In modality fusion problem, the speaker recognition schemes proposed in [22, 23, 25, 26, 27] are basically opinion fusion techniques that combine multiple expert decisions through adaptive or non-adaptive weighted summation of scores, whereas in [28, 29], fusion is carried out at feature-level by concatenating individual feature vectors so as to exploit the temporal correlations that may exist between audio and video signals. In audio-visual speech recognition, audio and lip data can be concatenated [5], while unimodal recognition rates

are combined to obtain the fused result [30]. Furthermore, recent works show the success of multi-stream HMM structures in speech recognition [7, 12, 14, 17].

1.2 System Overview and Contribution

Although numerous methods have been proposed for integration of lip information to speech/speaker recognition solutions, there is no framework proposed for selection of the most discriminative lip motion features optimally in the literature. This work aims at providing quantitative answers to the following open questions:
i) Is explicit lip motion, instead of or in addition to audio, lip intensity and/or geometry useful for speech/speaker recognition, and
ii) If so, what are the best lip motion features for speech-reading and speaker identification applications?
In order to answer these questions, first the problem of finding the best lip motion representation is considered in our work. Several lip motion feature candidates have been evaluated including dense motion features within a bounding box about the lip, lip contour motion features, and combination of these with lip shape features. In the dense motion computation case, no explicit information about the lip shape is included in the feature vector. The main disadvantage of this strategy is that some irrelevant noisy motion vectors may show up especially inside the inner lip boundary as parts of this region are occluded or uncovered during the speaking act. In the contour motion computation case, the lip boundary is tracked over time and only the motion of lip boundary pixels are taken into account. In this way, noisy motion vectors are mostly eliminated at the cost of disregarding some useful motion information around the lip. One advantage of this strategy is that extracted lip shape information can explicitly be included and exploited in the feature set as additional information.

After finding the best lip representation, a novel two-stage discriminant analysis is introduced to select the best lip motion features from this representation. At the first stage, the most discriminative features are selected from the full set of DCT coefficients of a single lip motion frame by using a probabilistic measure that maximizes the ratio of intra-class and inter-class probabilities. At the second stage, the resulting discriminative feature vectors are interpolated and concatenated for each time instant within a neighborhood, and further

analyzed by LDA to reduce dimension, this time taking into account temporal discrimination information. Following the determination of the best lip motion features, a multimodal system for both speech-reading and speaker identification has been implemented by fusing the best lip motion features with lip texture and audio.

Hence, the main contribution of this work is the introduction of a framework for determination of the most discriminative lip motion and shape features for speech-reading and speaker identification. The other contribution is made by fusing the available audio-visual modalities.

Chapter 2

THEORETICAL FRAMEWORK

2.1 Speaker Recognition

Speaker recognition task can be formulated as either verification or identification problem. The latter can further be classified as open-set or closed-set identification. 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 also a reject class. Open-set identification has a variety of applications such as the authorized access control for computer and communication systems, where a registered user can log onto the system with her/his personalized profile and access rights. In this work, we formulate the speaker recognition problem in an open-set identification framework, which is a more challenging and realistic way of addressing the problem as compared to closed-set speaker identification and verification. Note that verification is a special case of the general open-set identification problem.

The speaker identification problem is often formalized by using a probabilistic approach: Given a feature vector $\boldsymbol{f}$ representing the sample data of an unknown individual, compute the a posteriori probability $P(\lambda_r|\boldsymbol{f})$ for each speaker's model λ_r. The sample feature vector is then assigned to the class λ^* that maximizes the a posteriori probability,

$$\lambda^* = \arg\max_{\lambda_r} P(\lambda_r|\boldsymbol{f}). \tag{2.1}$$

One can rewrite (2.1) in terms of class-conditional probabilities using Bayes Rule:

$$P(\lambda_r|\boldsymbol{f}) = \frac{P(\boldsymbol{f}|\lambda_r)P(\lambda_r)}{P(\boldsymbol{f})}. \tag{2.2}$$

Since $P(\boldsymbol{f})$ is class independent and assuming equally likely class distribution, (2.1) is

equivalent to

$$\lambda^* = \arg\max_{\lambda_r} P(\boldsymbol{f}|\lambda_r). \tag{2.3}$$

In the open-set identification problem, an imposter class λ_{R+1} is 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 likelihood ratio $\bar{\rho}(\lambda_r)$ [31]:

$$\bar{\rho}(\lambda_r) = \log \frac{P(\boldsymbol{f}|\lambda_r)}{P(\boldsymbol{f}|\lambda_{R+1})} = \log P(\boldsymbol{f}|\lambda_r) - \log P(\boldsymbol{f}|\lambda_{R+1}). \tag{2.4}$$

Computation of class-conditional probabilities $P(\boldsymbol{f}|\lambda_r)$ needs a prior modeling step, through which a probability density function of feature vectors is estimated for each class $r = 1, 2, \ldots, R$ 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.

Then (2.3), which is accurate for a closed-set identification problem, is modified as,

$$\lambda_* = \arg\max_{\lambda_1,\ldots,\lambda_R} \bar{\rho}(\lambda_r), \tag{2.5}$$

and then

$$\begin{array}{ll} \text{if } \ \bar{\rho}(\lambda_*) \geq \tau & \text{accept} \\ \text{otherwise} & \text{reject} \end{array} \tag{2.6}$$

where τ is the optimal threshold which is usually determined experimentally to achieve the desired false accept or false reject rate [32].

The performance of the 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,

$$\text{FAR} = 100 \times \frac{F_a}{N_a + N_r} \quad \text{and} \quad \text{FRR} = 100 \times \frac{F_r}{N_a}, \tag{2.7}$$

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.

2.2 Speech Recognition

Speech recognition task can be formulated to identify a specific utterance, such as in the isolated word recognition task. Therefore the closed-set identification framework can be used to address the speech recognition problem with an isolated word dictionary.

The identification problem is formalized within the maximum likelihood framework. We can employ the maximum likelihood solution, which maximizes the class-conditional probability, $P(\boldsymbol{f}|\lambda_r)$, for $r = 1, ..., R$. Hence a decision in the closed-set identification is taken as,

$$\lambda_* = \underset{\lambda_1,...,\lambda_R}{\arg\max} \log P(\boldsymbol{f}|\lambda_r) = \underset{\lambda_1,...,\lambda_R}{\arg\max} \rho(\lambda_r). \tag{2.8}$$

The performance of the speech recognition systems are generally measured by the phoneme/word/phrase recognition rate in percentage, that is the ratio of the true matches to the total number of trials.

2.3 Recognition using Hidden Markov Models (HMM)

The Hidden Markov Models (HMM), [33], is a special case of Markov chains. It can be described as a doubly stochastic process, where the sequence of one stochastic process is observed and the other is not (it is the hidden part which gives the name of $Hidden$). The identification task addresses the problem of finding the most probable path or sequence of the hidden stochastic process, given an observation sequence and HMM parameters. Since it is able to provide a mathematical framework for sequentially evolving pattern recognition tasks, HMM can fit to both speech recognition and speaker identification problems.

The temporal characterization of an audio-video stream can successfully be modeled using an HMM structure, where state transitions model temporal correlations and in each state Gaussian classifiers model signal characteristics. Considering a left-to-right continuous density HMM structure, an HMM can be defined by the following parameter set:

- N is the number of states, where states are denoted by $\mathbf{S} = \{s_1, s_2, \cdots, s_N\}$.
- $\boldsymbol{A} = \{a_{ij}\}$ is the matrix of state transition probabilities where a_{ij} is the probability of making a transition from state i to state j, such that $a_{ij} = P(q_{\tau+1} = s_j|q_\tau = s_i)$,

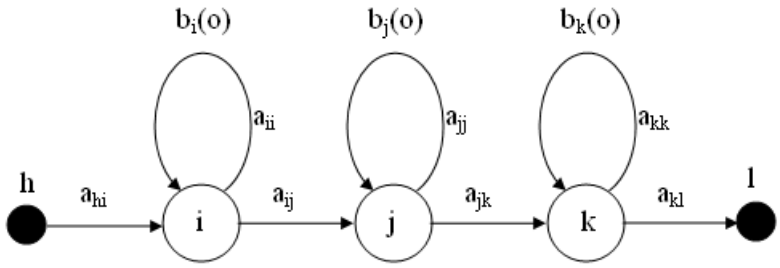


Figure 2.1: A Hidden Markov Model with three emitting states and continuous output distributions

where q_τ is the state at time τ. The state transition probabilities are assumed to be time independent.

- $\boldsymbol{B} = \{b_j(\boldsymbol{f})\}$ is the vector of observation probabilities associated with each emitting state j, with $b_j(\boldsymbol{f}) = P(\boldsymbol{f}|q_\tau = s_j)$.

- $\boldsymbol{\Pi} = \{\pi_i\}$ is the vector with the initial state probabilities of entering the model at state i such that $\pi_i = P(q_1 = s_i)$.

A HMM can now be represented by the compact parameter set $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\Pi})$. Since the speech signal evolves forward in time, the transition probability matrix A is normally constrained to only allow self-loops, by residing in the same state for several consecutive frames, or transitions from left to right.

The likelihood function for the temporal characterization, that is the probability of observing feature vector sequence $\mathbf{F} = (\boldsymbol{f}_1, \boldsymbol{f}_2, ..., \boldsymbol{f}_K)$, given the model λ is defined as,

$$P(\mathbf{F}|\lambda) = \sum_{\text{all } \boldsymbol{q}} P(\mathbf{F}, \boldsymbol{q}|\lambda), \tag{2.9}$$

where $\boldsymbol{q} = (q_1, q_2, ..., q_K)$ is a possible state transition sequence. Further we can write the joint probability of the observation sequence and the state transition sequence given the model as,

$$P(\mathbf{F}, \boldsymbol{q}|\lambda) = P(\mathbf{F}|\boldsymbol{q}, \lambda)P(\boldsymbol{q}|\lambda), \tag{2.10}$$

where

$$\begin{aligned} P(\mathbf{F}|\boldsymbol{q},\lambda) &= b_{q_1}(\boldsymbol{f}_1)b_{q_2}(\boldsymbol{f}_2)\cdots b_{q_K}(\boldsymbol{f}_K), \quad \textit{and} \\ P(\boldsymbol{q}|\lambda) &= \pi_{q_1}a_{q_1q_2}a_{q_2q_3}\cdots a_{q_{K-1}q_K}. \end{aligned}$$

The resulting likelihood function from (2.9) will be in the form of,

$$P(\mathbf{F}|\lambda) = \sum_{\text{all } \boldsymbol{q}} \pi_{q_1} b_{q_1}(\boldsymbol{f}_1)a_{q_1q_2}b_{q_2}(\boldsymbol{f}_2)a_{q_2q_3}\cdots b_{q_{K-1}}(\boldsymbol{f}_{K-1})a_{q_{K-1}q_K}b_{q_K}(\boldsymbol{f}_K), \tag{2.11}$$

in which observation symbol probabilities $b_j(\boldsymbol{f})$ are modeled using Gaussian mixture densities as,

$$b_j(\boldsymbol{f}_k) = \sum_{l=1}^{L} \omega_{jl}\mathcal{N}(\boldsymbol{f}_k, \mu_{jl}\Sigma_{jl}), \tag{2.12}$$

where for each state j feature vector probabilities are represented as the weighted sum of L Gaussian mixture densities with means μ_{jl}, covariance matrices Σ_{jl} and weights ω_{jl}, such that $\sum_l \omega_{jl} = 1$ and $0 < \omega_{jl} \leq 1$.

HMMs are known to be as effective structures to model the temporal behavior of the speech signal, and thus they are widely used both in audio-based speaker identification and speech recognition applications [1]. State-of-the-art systems use HMMs for text-dependent and Gaussian mixture models (GMM) for text-independent speaker identification [34]. HMM-based techniques are preferred in text-dependent scenarios since HMM structures can successfully exploit the temporal correlations of a speech signal. Since lip motion is strongly coupled with audio utterance, HMMs can also be employed for temporal characterization of lip features.

We use word-level continuous-density HMM structures for both speaker identification and speech recognition tasks. Each speaker or utterance in the database is modeled using a separate HMM that is trained over some repetitions of the lip motion streams of the corresponding speaker or utterance. In the recognition process, given a test feature set, each HMM structure associated with a speaker or an utterance produces a likelihood. In the speaker identification case, a world HMM model is also trained over the whole training data of the population. The log-ratio of the speaker likelihoods and the world class likelihood results in a stream of log-likelihood ratios that are used in the speaker identification process. The system identifies the person if there is a match and rejects otherwise. Alternatively in

speech recognition, the impostor or world class is not defined; thus the best match is given by the utterance class that maximizes the produced likelihood as described in Section 2.2.

Chapter 3

LIP MOTION FEATURE EXTRACTION

The proposed lip motion feature extraction and analysis system is depicted in Figure 3.1. It consists of a preprocessing module, a lip motion estimation module, and a two-stage discrimination module. We consider two alternatives for lip motion estimation: i) Dense motion vectors within a rectangular grid, and ii) Motion vectors along the lip contour together with lip shape information. Each of these modules are explained in detail below.

3.1 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 2D parametric motion estimator. For every two consecutive frames, global head motion parameters are calculated using hierarchical Gaussian image pyramids and the 12-parameter quadratic motion model [35]. 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 initial quadratic transform to model head motion, at each pixel (x_p, y_p) in the region of interest, the flow vector $[u(x_p, y_p), v(x_p, y_p)]$ is estimated from the image intensities $I(x_p, y_p, t)$ and $I(x_p + u(x_p, y_p), y_p + v(x_p, y_p), t+1)$ at time instants t and $t+1$ respectively as,

$$\begin{aligned} u(x_p, y_p) &= a_1 x_p^2 + a_2 y_p^2 + a_3 x_p y_p + a_4 x_p + a_5 y_p + a_6, \\ v(x_p, y_p) &= b_1 x_p^2 + b_2 y_p^2 + b_3 x_p y_p + b_4 x_p + b_5 y_p + b_6. \end{aligned} \tag{3.1}$$

The optimal motion parameters $\{a_1, \cdots, a_6, b_1, \cdots b_6\}$, which best describe the motion at each pixel in the region of interest, are found through an optimization task and used to back-warp the face frames.

The use of quadratic transform for motion modeling assumes certain physical conditions for the 3D geometric surface of the object, its 3D real motion, and the camera projection model. The quadratic transform gives an exact description of the 3D rotation, translation and scaling of an object with a parabolic surface under parallel projection [36]. Thus it serves only as a good approximation for the rigid motion of the human head and is quite effective in modeling the motion between consecutive frames for which the movement is mostly not very abrupt.

3.2 Motion Estimation Alternatives

In order to investigate the effect of different types of motion estimation approaches, two different methods to compute the lip motion vectors have been utilized: i) Optical flow, ii) Block matching.

3.2.1 Motion Estimation by Optical Flow

Optical flow is defined as an apparent motion of image brightness, $I(x, y, t)$, which changes in time to provide an image sequence. There exist two crucial assumptions which the optical flow rely on:

1. Brightness $I(x, y, t)$ depends on coordinates x, y in greater region of the image.

2. Brightness of every point of a moving or static object does not change in time.

Suppose that some object in the image, or some point of an object, moves and after time dt the object displacement is (dx, dy). Using Taylor series for brightness $I(x, y, t)$ gives the following:

$$I(x + dx, y + dy, t + dt) = I(x, y, t) + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt + \cdots , \tag{3.2}$$

where "$\cdots$" are higher order terms. According to the second assumption, (3.2) becomes

$$I(x + dx, y + dy, t + dt) = I(x, y, t), \tag{3.3}$$

and

$$\frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt + \cdots = 0 \tag{3.4}$$

By dividing (3.4) by dt and defining $\frac{dx}{dt} = u$ and $\frac{dy}{dt} = v$, the following equation called *optical flow constraint equation* is obtained:

$$\frac{\partial I}{\partial t} = \frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v \tag{3.5}$$

Here u and v are components of optical flow field in x and y coordinates respectively. Since (3.5) has more than one solution, more constraints are required.

In the Lucas-Kanade method [37], by using the optical flow equation for group of adjacent pixels and assuming that all of them have the same velocity, a system of linear equations can be formed. In a non-singular system for two pixels a velocity vector can be computed to solve the system. However, combining equations for more than two pixels is more effective. It is possible to get a system that has no solution; yet it can be solved roughly using the least-squares (LS) method. The weighted combination of equations is used. This method involves the solution of 2×2 linear system:

$$\begin{aligned} \sum_{x,y} W(x,y)I_x I_y u + \sum_{x,y} W(x,y)I_y^2 v &= -\sum_{x,y} W(x,y)I_y I_t \\ \sum_{x,y} W(x,y)I_x^2 u + \sum_{x,y} W(x,y)I_x I_y v &= -\sum_{x,y} W(x,y)I_x I_t, \end{aligned} \tag{3.6}$$

where $W(x,y)$ is the Gaussian window. I_x, I_y and I_t are the partial derivatives of I with respect to x, y and t respectively. The Gaussian window may be represented as a composition of two separable kernels with binomial coefficients. Iterating through the system can yield even better results. That is, retrieved offset is used to determine a new window in the second image from which the window in the first image is subtracted while I_t is calculated.

In our optical flow approach, a three-level hierarchical structure using Lucas-Kanade technique in an image neighborhood of 13×13 is employed. Hierarchical motion estimation speeds the motion search up by repeatedly down-converting both the current and the reference frame by a factor of two in both dimensions, and doing motion estimation on smaller pictures. At each stage of the hierarchy, vectors from lower levels, i.e., smaller versions of the picture, are used as a guide for searching at higher levels. This dramatically reduces the size of search for large motions. Intel's open-source computer vision library OpenCV© possesses a build-in implementation of the algorithm with sub-pixel accuracy. Details about the pyramidal implementation of the algorithm can be found in [38].

3.2.2 Motion Estimation by Block Matching

Block matching (BM) is a technique that searches corresponding image point in two successive images by comparing a block of surrounding image points in both images. There exist different search ways depending on the motion such as full-search BM, neighborhood-search BM, etc. It is a known fact that the full-search approach is computationally more expensive than the neighborhood-search. Thus when there is *a priori* information that the motion is not too large, as in lip movement, one can utilize the neighborhood-search BM by specifying the maximum allowable lengths of the motion in both x and y directions. Another issue in block-matching is the matching criterion: there are well-known criteria such as sum-of-absolute differences (SAD) and sum-of-squared differences (SSD).

Mathematically speaking, block matching aims to find the displacement vector $\boldsymbol{d} = (d_x, d_y)$ by minimizing the residual function $\Delta(\boldsymbol{d})$ for a block of size $(2b_x + 1) \times (2b_y + 1)$ surrounding the pixel (p_x, p_y) of the image I in the next image J:

$$\Delta(\boldsymbol{d_{SSD}}) = \sum_{x=p_x-b_x}^{p_x+b_x} \sum_{y=p_y-b_y}^{p_y+b_y} (I(x,y) - J(x+dx, y+dy))^2, \tag{3.7}$$

$$\Delta(\boldsymbol{d_{SAD}}) = \sum_{x=p_x-b_x}^{p_x+b_x} \sum_{y=p_y-b_y}^{p_y+b_y} |I(x,y) - J(x+dx, y+dy)| \,. \tag{3.8}$$

In this work, for the sake of better comparison between motion estimation methods, a three-level hierarchical block-matching algorithm with SSD criterion and blocks of size 13×13 has been implemented. Quarter-pel accuracy is reached by interpolating the original lip image and using appropriate 6-tap Wiener and bilinear filters used in H.264/MPEG-4 AVC [39]. The maximum allowable displacement vector is set to $\boldsymbol{d_{max}} = [\pm 7.75, \pm 7.75]$ considering all pyramid levels. This maximum allowable setting is found to be successful to model all motion vectors accurately [40].

3.3 Extraction of Grid-based Motion Features

The first alternative that we consider is the use of a dense uniform grid of size $G_x \times G_y$ on the intensity lip image. This grid definition allows to analyze the whole motion information contained within the rectangular mouth region. The motion analysis is conducted using both

the optical flow and the block matching methods described in Section 3.2.1 and Section 3.2.2, respectively.

The motion estimation procedure yields two $G_x \times G_y$ 2D 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 2D-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 Figure 3.2. 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. 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 4.1.

3.4 Extraction of Contour-based Motion Features

3.4.1 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 [16, 41, 42, 43, 44, 45, 46, 47]. 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 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 [46], linear discriminant analysis [47], adaptive Gaussian mixture models [41] or fuzzy clustering methods as in [42, 44]. 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 (ASM) [13, 48] 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 [43] and their extensions such as GVF-snakes [16] suffer from complex pa-

rameter 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 [45], where we fit polynomials on the outer lip contour. The technique is based on 6 designated key points detected on the lip contour. The algorithm can briefly be outlined as follows:

- The algorithm starts by manually putting one single seed above the mouth and near its vertical symmetry axis to initialize the *jumping snake* [49].

- The upper lip boundary is found by after the convergence of the snake and the three points forming the Cupid's bow on the upper lip, i.e., P_2, P_3, and P_4, are detected by a simple local maxima-minima function.

- Another key point P_6 on the lower lip boundary near the vertical axis of the mouth is located by analyzing the one-dimensional gradient of the pseudo-hue along the vertical axis passing by P_3.

- Mouth corners P_1 and P_5 are detected using both the minima of luminance computed along each vertical pixel group and an edge criterion. Figure 3.3.a shows a lip image during the mouth corner detection.

- The modeling stage is basically an optimization task that uses the color information to draw two lines fitting to the Cupid's bow and four cubic polynomials for the rest. Figure 3.3.b illustrates the parametric model candidates for the upper right contour and the optimal one.

- The key points are tracked from one image to the other using a variant of the Lucas-Kanade algorithm adapted to the particular geometry of the mouth and the modeling stage is repeated for the other images in the sequence.

Details about the algorithm can be found in [45]. Figure 3.4.a shows the 6 key points and the fitted parametric model on a lip image.

When tested on our visual database, the technique proposed in [45] mostly yields very accurate lip tracking results, but only under some assumptions on the acquisition environment and illumination conditions. 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 [45]. In cases where it fails, the algorithm is assisted with some extra key points which are hand-labeled on the lip boundary. Figure 3.4.c displays examples of lip contours extracted from various images of our database.

3.4.2 Contour-based Motion Features

The extracted parametric lip contour is in fact a rough sketch of the real lip and does not contain sufficiently detailed information for especially the speaker identification problem, i.e., to characterize discriminative biometrics of different speakers. The discriminative information can be captured by incorporating the motion vectors computed along the parametric lip contour. Thus 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 $\boldsymbol{f}_{\mathrm{CTR}}$. Figure 3.5 depicts the procedure for extraction of contour-based lip motion features.

3.4.3 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 composed

of these 6 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. One possible such feature vector representation is composed of 8 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 ($L_2, ..., L_8$) as depicted in Figure 3.4.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 Figure 3.5.

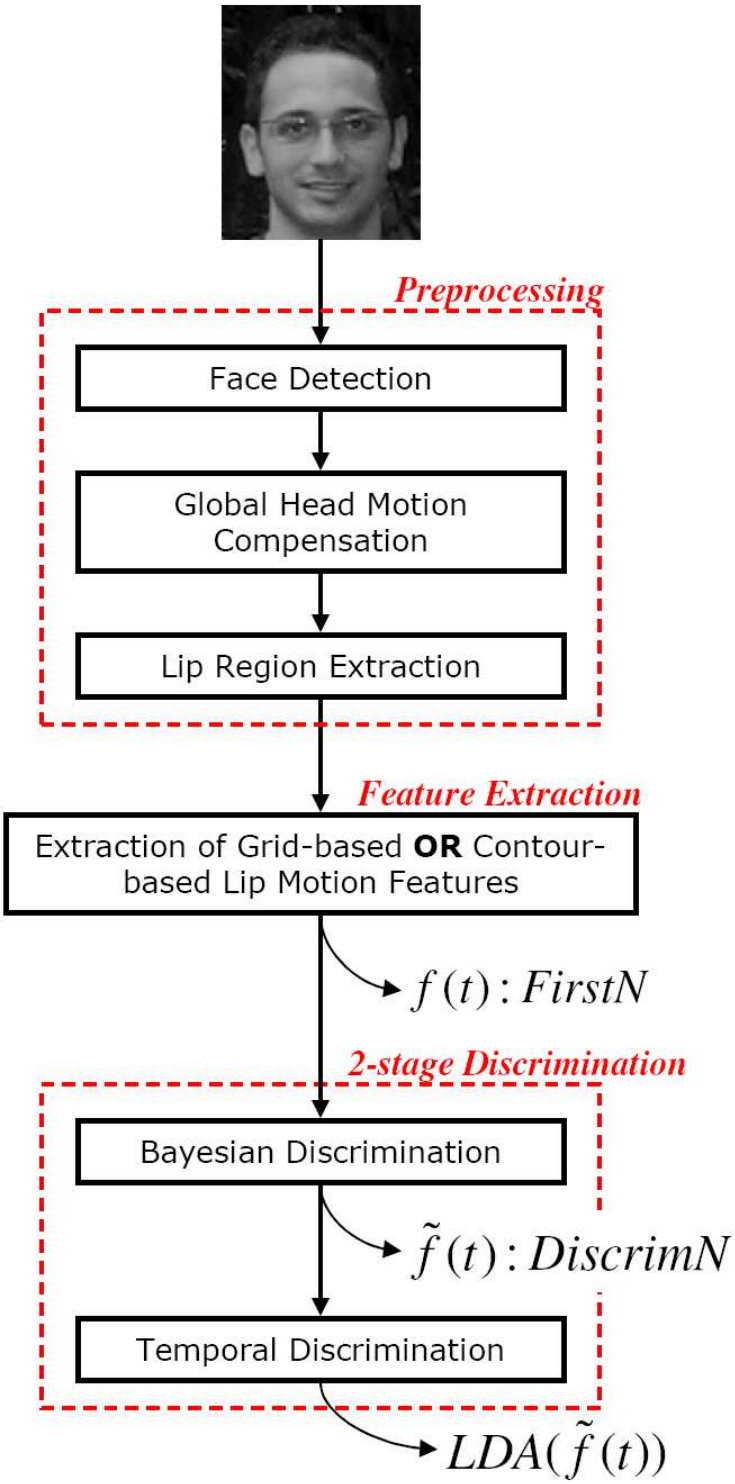


Figure 3.1: Block diagram of the overall lip feature extraction system

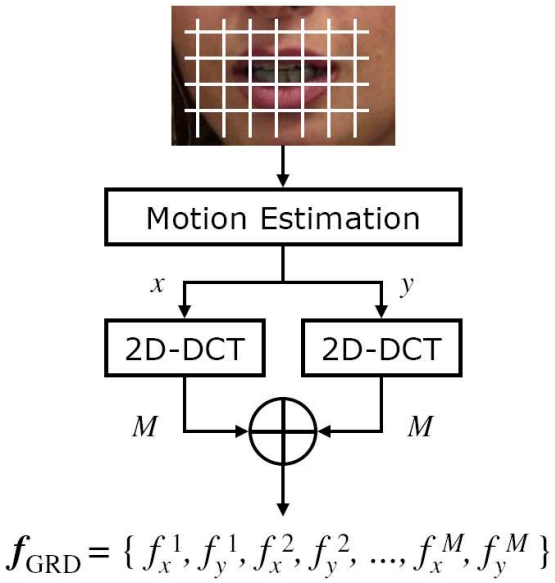
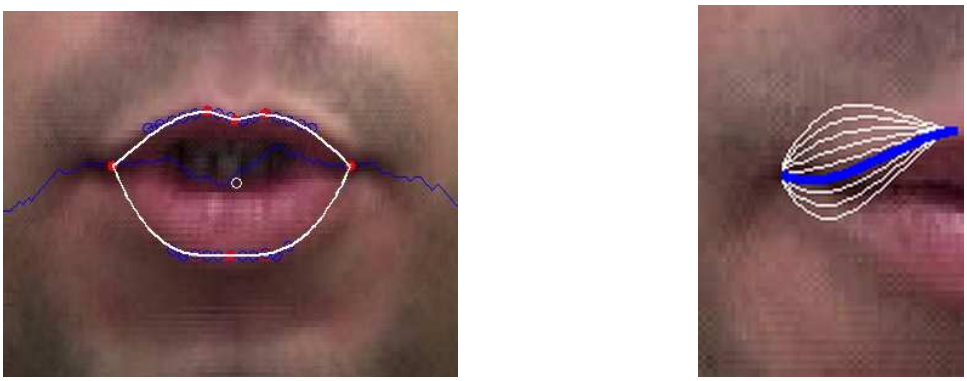
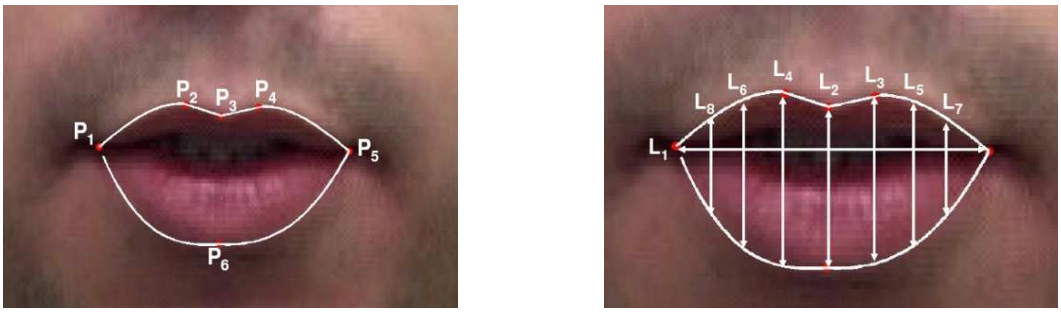


Figure 3.2: Grid-based lip motion feature extraction



(a) (b)

Figure 3.3: Crucial lip contour extraction stages (a) Mouth corner detection (b) Parametric model fitting.



(a)

(b)

(c)

Figure 3.4: Lip contour modeling (a) The 6 key points and parametric models fitted on the outer contour, (b) The 8 lip shape parameters, (c) Extracted lip contours.

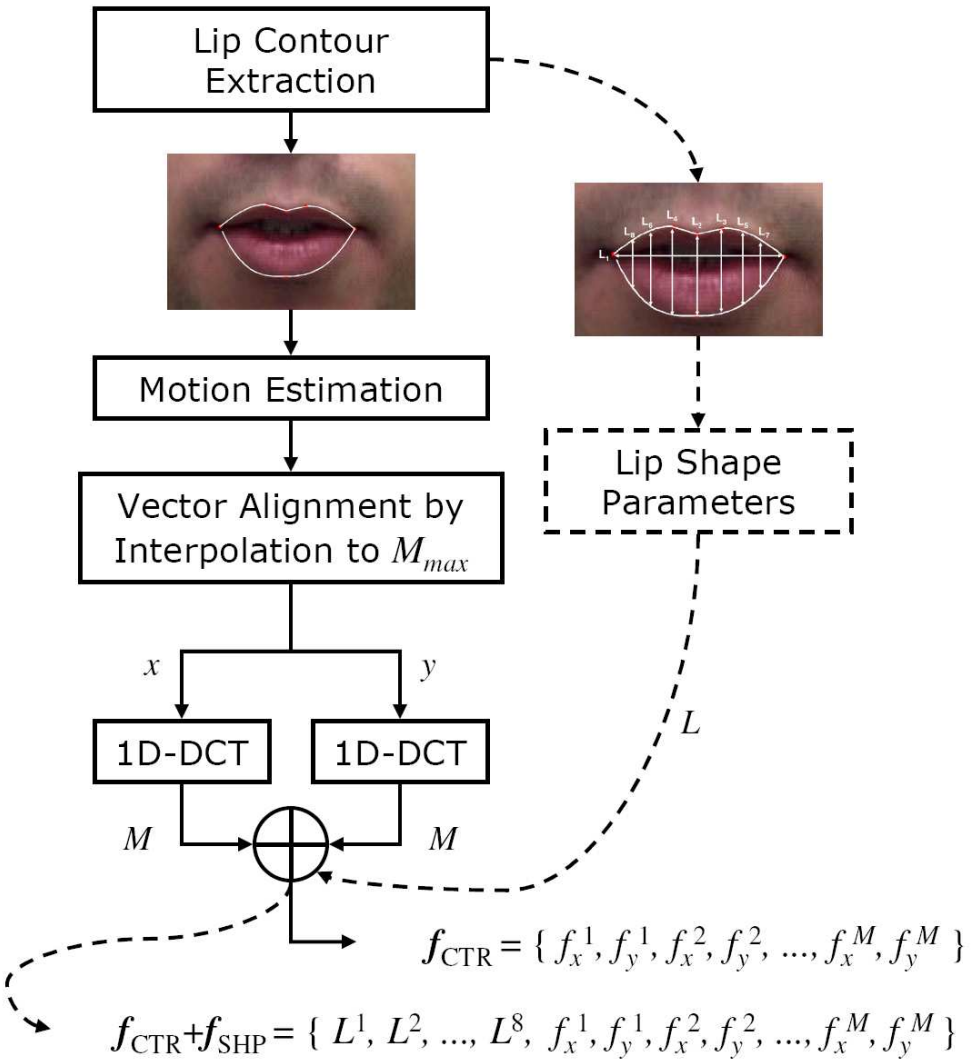


Figure 3.5: Contour-based lip motion feature extraction and concatenation of motion and shape features (the dashed lines show the optional path to include lip shape parameters).

Chapter 4

DISCRIMINATION ANALYSIS

In this work, 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 discrimination stage uses the linear discriminant analysis (LDA). The LDA is a well-known dimension reduction and feature extraction method to achieve discrimination among multiple classes [7, 50, 51]. The details of these two stages are discussed in the following.

However, it will be beneficial to first give a brief information on Gaussian mixture models since it serves as a tool to estimate unknown probability densities in Bayesian feature selection. The theory of the LDA will be also discussed within the theoretical background.

4.1 Bayesian Discriminative Feature Selection

Let f_k denote the k-th 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 maximum posterior probability $P(\lambda_i|f_k)$ which can be written in terms of class-conditional probability distributions:

$$\begin{aligned} P(\lambda_i|f_k) &= \frac{P(f_k|\lambda_i)P(\lambda_i)}{P(f_k)} \\ &= \frac{P(f_k|\lambda_i)P(\lambda_i)}{P(f_k|\lambda_i)P(\lambda_i) + \sum_{j\neq i} P(f_k|\lambda_j)P(\lambda_j)} \\ &= \left[1 + \frac{\sum_{j\neq i} P(f_k|\lambda_j)P(\lambda_j)}{P(f_k|\lambda_i)P(\lambda_i)}\right]^{-1} \end{aligned} \tag{4.1}$$

Then the MAP estimator becomes the maximum mutual information estimator (MMIE) [52] by maximizing the ratio $l(\lambda_i|f_k)$,

$$l(\lambda_i|f_k) = \log \frac{P(f_k|\lambda_i)P(\lambda_i)}{\sum_{j\neq i} P(f_k|\lambda_j)P(\lambda_j)}. \tag{4.2}$$

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.

In most cases, the class probabilities, $P(\lambda_i)$, can be assumed to be equally likely. Thus the only unknown in the $l(\lambda_i|f_k)$ ratio is the class-conditional probabilities $P(f_k|\lambda_i)$. The class-conditional probability distributions are generally computed over some training data using expectation-maximization type algorithms, assuming an underlying probability distribution. Gaussian mixture density modeling has been widely used in various disciplines that require signal characterization for classification and recognition, as well as estimation of unknown probability densities. In this work, the Gaussian mixture models are used for the class-conditional probability density function, $p(f_k|\lambda_i)$, estimation, where f_k and λ_i are respectively the k-th feature coefficient and the i-th class model. The maximum-likelihood method has been one of the most commonly used techniques to estimate the parameters of the mixture densities. Mathematically speaking, if we have an input sequence of N samples $\{x_1, \ldots, x_N\}$, the underlying density function for K mixtures is given as,

$$f(x) = \sum_{k=1}^{K} \omega_k \mathcal{N}_k(\mu_k, \Sigma_k), \tag{4.3}$$

where ω_k values are the mixture weights, μ_k and Σ_k are respectively mean vector and covariance matrix of the k-th Gaussian mixture density $\mathcal{N}_k$. The *EM* (Expectation-Maximization) algorithm to train the mixture densities over our sample data is given as:

- *Initialize* mixture weights, means and covariance matrix. Repeat the following E-step and M-step until achieving convergence.

- *E-step:* Calculate the responsibility $p(k|x_n)$ of each Gaussian mixture $\mathcal{N}_k$ for each training data point x_n as,

$$p_{kn} = p(k|x_n) = \frac{p(x_n|k)\omega_k}{p(x_n)}, \tag{4.4}$$

  where $p(x_n)$ can be calculated as, $p(x_n) = \sum_{k=1}^{K} p(x_n|k)\omega_k$.

- *M-step:* Re-estimate mixture weights, means and covariance matrix,

$$\hat{\omega}_k = \frac{\sum_n p_{kn}}{\sum_k \sum_n p_{kn}}, \quad \hat{\mu}_k = \frac{\sum_n p_{kn} x_n}{\sum_n p_{kn}}, \quad \hat{\sigma}^2_{ik} = \frac{\sum_n p_{kn}(x_n - \hat{\mu}_k)(x_n - \hat{\mu}_k)'}{\sum_n p_{kn}}. \tag{4.5}$$

The *EM* algorithm is employed using diagonal covariance matrices, since feature components are assumed to be independent of each other.

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.

Let us refer to the training data, which is used to compute the class-conditional probability densities, as f_k^i, that is a collection of observations of the k-th feature component from the i-th 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:

$$d(f_k) = \sum_i \frac{1}{L} \sum_{l=0}^{L-1} l(\lambda_i|f_k^i(l)), \tag{4.6}$$

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

4.1.1 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

$$d(f_{k_1}) > d(f_{k_2}) > \cdots > d(f_{k_K}). \tag{4.7}$$

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,

$$\tilde{\boldsymbol{f}}^N = (f_{k_1}, f_{k_2}, ..., f_{k_N}). \tag{4.8}$$

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 Chapter 3. Both involve the DCT coefficients of the motion vectors computed either on a 2D rectangular grid covering the lip region or along the 1D 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 to 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 $\boldsymbol{f}^N = (f_1, f_2, ..., f_N)$. The discriminative set of features, $\tilde{\boldsymbol{f}}^N$, that are introduced in (4.8), will be referred to as *DiscrimN*. Note that they are selected according to the discriminative power ordering specified in (4.7).

4.1.2 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 $d(f)$ in (4.6) 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:

$$D_N(\boldsymbol{f}) = \sum_{n=1}^{N} d(f_{k_n}). \tag{4.9}$$

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.

4.2 Temporal Discriminative Feature Selection using LDA

The Bayesian MMIE-based discriminative feature selection technique described in Section 4.1 does not model and exploit the temporal correlations existing between successive lip frames. One could include the first and second derivatives of the feature vectors to

better model temporal variations, but this would result in higher dimensional feature representations. Alternatively, following the work of Potamianos et al. [7], 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 ($\boldsymbol{S}_b$),

$$\boldsymbol{S}_w = \sum_{j=1}^{R} \sum_{i=1}^{Q_j} (\mathbf{x}_i^j - \mu_j)(\mathbf{x}_i^j - \mu_j)^T, \tag{4.10}$$

$$\boldsymbol{S}_b = \sum_{j=1}^{R} (\mu_j - \mu)(\mu_j - \mu)^T, \tag{4.11}$$

where $\mathbf{x}_i^j$ is the i-th sample of class j, μ_j is the mean of class j, μ is the mean of all classes, R is the number of classes, and Q_j the number of samples in class j [53].

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:

$$\boldsymbol{\epsilon(W)} = \frac{\det\left(\boldsymbol{W}^T \boldsymbol{S}_b \boldsymbol{W}\right)}{\det\left(\boldsymbol{W}^T \boldsymbol{S}_w \boldsymbol{W}\right)}, \tag{4.12}$$

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: i) 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, and ii) 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.

Due to the second limitation of the LDA, an intermediate step to reduce the feature vector dimension is needed. This intermediate reduction is also preferable to reduce the

computational complexity of the LDA analysis. In this regard, the Bayesian MMIE-based analysis that we propose in Section 4.1, 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.

The MMIE-based 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 $\boldsymbol{F}(t)$:

$$\boldsymbol{F}(t) = [\tilde{\boldsymbol{f}}(t-T), \tilde{\boldsymbol{f}}(t-T+1), \cdots, \tilde{\boldsymbol{f}}(t), \cdots, \tilde{\boldsymbol{f}}(t+T-1), \tilde{\boldsymbol{f}}(t+T)]. \tag{4.13}$$

The LDA analysis is then performed on this concatenated vector of dimension $(2T+1)N$. The dimension of the resulting discriminative feature space is bounded above by $R-1$, that is one less than the total number of classes. Figure 4.1 illustrates the formation of the final feature vector, that we will denote by $LDA(\tilde{\boldsymbol{f}}(t))$, via temporal and spatial discrimination analysis.

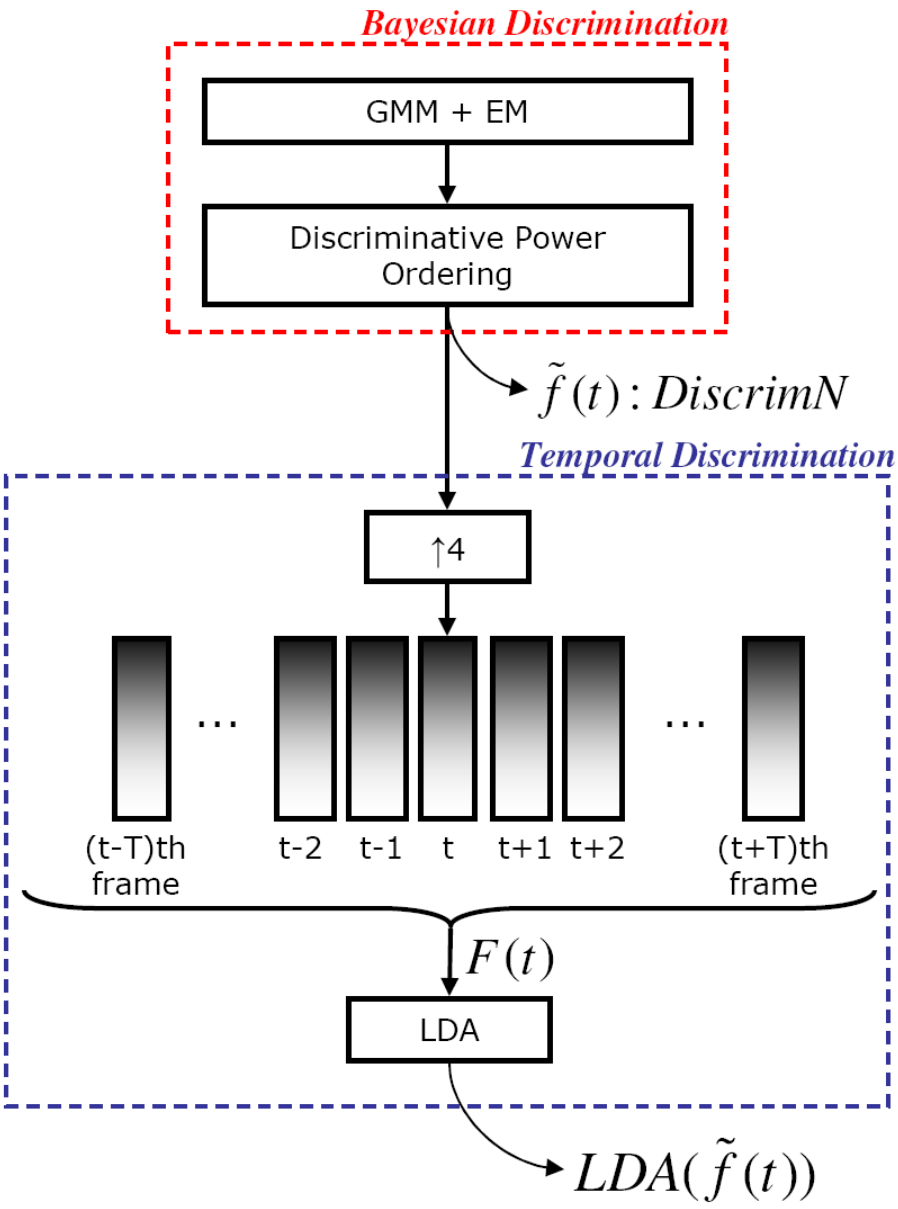


Figure 4.1: Two-stage discrimination analysis for lip motion features.

Chapter 5

MULTIMODAL DECISION FUSION WITH AUDIO AND LIP TEXTURE

In this chapter, the decision fusion strategy employed and other modalities, namely audio and lip texture will be presented.

5.1 Other Audio-Visual Features

5.1.1 Audio Features

Audio stream is represented with the mel-frequency cepstral coefficients (MFCC), as they yield good discrimination of speech signals. The audio stream is processed over 10 msec frames centered on 25 msec Hamming window for 16 kHz sampled audio signal. Each analysis frame is first multiplied with a Hamming window and transformed to frequency domain using Fast Fourier Transform (FFT). Mel-scaled triangular filter-bank energies are calculated over the square magnitude of the spectrum and represented in logarithmic scale [54]. The resulting MFCC features, c_j, are derived using discrete cosine transform (DCT) over log-scaled filter-bank energies e_i:

$$c_j = \frac{1}{N_M} \sum_{i=1}^{N_M} e_i \cos\left((i - 0.5)\frac{j\pi}{N_M}\right), \qquad j = 1, 2, ..., N. \tag{5.1}$$

where N_M is the number of mel-scaled filter banks and N is the number of MFCC features that are extracted. The MFCC feature vector is defined as, $\boldsymbol{C} = [c_1 c_2 \cdots c_N]^T$. The audio feature vector $\boldsymbol{f}_{\mathrm{A}}$ is formed as a collection of MFCC vector $\boldsymbol{C}$ along with the first and second delta MFCCs, $\boldsymbol{f}_{\mathrm{A}} = [\boldsymbol{C}\ \Delta\boldsymbol{C}\ \Delta\Delta\boldsymbol{C}]$. Audio feature extraction is briefly illustrated in Fig. 5.1.a.

5.1.2 Lip Texture Features

It has been a common practice to use intensity-based features for the representation of lip texture [7, 18]. There are certain advantages and draw-backs of the intensity-based lip

features, such as representing texture information as well as shape but being sensitive to illumination changes. Fig. 5.1.b shows the intensity-based DCT feature extraction. The intensity-based lip features, which are denoted by $\boldsymbol{f}_{\mathrm{L_t}}$, are extracted by the Bayesian discrimination [55] from the zig-zag scan of 2D-DCT coefficients.

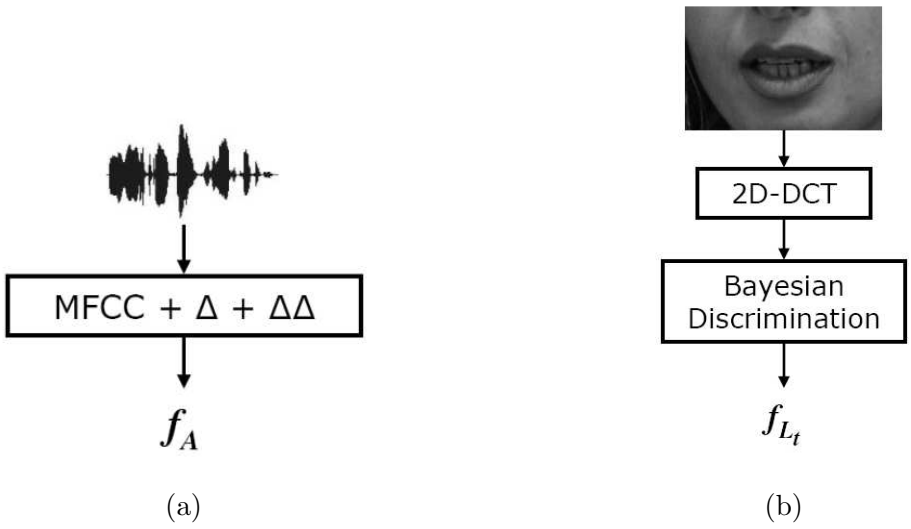


Figure 5.1: Block diagrams of the feature extraction for modalities: (a) Audio, (b) Lip texture.

5.2 Multimodal Decision Fusion At a Glance

When more than one information source is available, the fusion of information from different sources can reduce overall uncertainty and increase the robustness of a classification system. Suppose that P different classifiers, one for each of the P modalities $\boldsymbol{f}_1, \boldsymbol{f}_2, ..., \boldsymbol{f}_P$, are available. Each classifier, say the p-th classifier, produces a set of N-class log-likelihood ratios $\rho_p(\lambda_n)$, $n = 1, \ldots, N$. The problem then reduces to computing a single set of joint log-likelihood ratios $\rho(\lambda_1), \rho(\lambda_2), ..., \rho(\lambda_N)$ for these P modalities. In the Bayesian framework, assuming that $\boldsymbol{f}_1, \boldsymbol{f}_2, ..., \boldsymbol{f}_P$ are statistically independent, the joint log likelihood ratio is given by the sum of the individual ratios:

$$\rho(\lambda_n) = \log \frac{P(\boldsymbol{f}_1|\lambda_n) \cdots P(\boldsymbol{f}_P|\lambda_n)}{P(\boldsymbol{f}_1|\lambda_{N+1}) \cdots P(\boldsymbol{f}_P|\lambda_{N+1})} = \sum_p \rho_p(\lambda_n), \tag{5.2}$$

which is equivalent to the so-called product rule [4]. In practice, there are three main problems with the optimality of this rule. First, partial decisions coming from different classifiers may be correlated. Second, due to modeling errors and/or measurement noise, the estimated distribution model of training features, i.e., $P(\boldsymbol{f}_p|\lambda_n)$, may not always comply with the actual distribution of test features. Third, the impostor model, i.e., $P(\boldsymbol{f}_p|\lambda_{N+1})$, is a mere approximation of the reality. As a result, the log likelihood ratios coming from separate classifiers should each be considered as an opinion or a likelihood score rather than a probabilistic value. The statistics and the numerical range of these likelihood scores mostly vary from one classifier to another, and thus they need to be normalized into the interval $(0, 1)$ before the fusion process, using methods such as sigmoid and variance normalization. Unfortunately there is no formally "correct" or optimal way of normalization, which is investigated in detail in [56]. In this work a sigmoid normalization is used as in [26], which maps likelihood ratios to the $(0, 1)$ interval by normalizing the likelihood ratio ρ using the function

$$g(\rho) = \left[1 + e^{-(\frac{\rho-\mu}{2\sigma}+1)}\right]^{-1}, \tag{5.3}$$

where μ and σ are the mean and the standard deviation of the likelihood ratio ρ over the accept subjects, respectively.

In order to cope with the above problems, various approximation approaches have been proposed in the literature as alternatives to the product rule (i.e., the sum rule in log domain) such as max rule, min rule and reliability-based weighted summation. In fact, the most generic way of computing joint ratios (or scores) can be expressed as a weighted summation:

$$\rho(\lambda_n) = \sum_{p=1}^{P} \omega_p \rho_p(\lambda_n) \quad \text{for } n = 1, 2, ..., N, \tag{5.4}$$

where ω_p denotes the weighting coefficient for modality p, such that $\sum_p \omega_p = 1$. Then, the fusion problem becomes finding the optimal weight coefficients. Note that when $\omega_p = \frac{1}{P} \ \forall p$, (5.4) is equivalent to the product rule. On one side, there are hard-level combination techniques such as max rule, min rule and median rule [4], that use binary values for assignment of the weighting coefficients. These techniques combine decisions rather than likelihood scores and in this way try to filter out some of the erroneous likelihoods. The max rule and the min rule for example rely only on the classifier with the highest and the lowest best likelihood scores, respectively, and disregard the decisions of the other classifiers.

In this sense, the max rule tends to have a high false accept rate, whereas the min rule is suited to high security applications. Both methods rely solely on likelihood scores and do not employ an additional reliability measure. Soft-level combination techniques, on the other hand, regard each coefficient as a measure of the relative reliability R_p of each classifier so that each w_p becomes directly equal to R_p. We refer to this combination method as *Reliability Weighted Summation* (RWS) rule. Reliability values R_p can be set to some fixed values using some a priori knowledge about the performance of each modality classifier or can be estimated adaptively for each decision instant. The problem of the reliability-based weighting approach is that the numerical estimation of reliability values itself, which is ideally feature and class dependent, is not in general very accurate; thus erroneous likelihood scores contribute to the joint score, corrupting correct partial decisions.

5.3 Reliability Weighted Summation (RWS)

Among various reliability estimation techniques existing in the literature, we favor the one proposed in [18], since it is better suited to the open-set speaker identification problem by assessing both accept and reject decisions of a classifier, and it can easily be defined for the closed-set identification problem.

The RWS rule combines likelihood ratio values of the N modalities weighted by their reliability values ω_n as in (5.4). The reliability value ω_n is estimated based on the difference of likelihood ratios of the best two candidate classes λ_* and λ_{**}, that is, $\Delta_n = \rho_n(\lambda_*) - \rho_n(\lambda_{**})$, for modality n. In the absence of reject class, that is for closed-set identification, the likelihood difference of the best two candidates, Δ_n, can be used as the reliability value. However, in the presence of a reject class, one would expect a high likelihood ratio $\rho_n(\lambda_*)$ and a high Δ_n value for true accept decisions, and a low likelihood ratio $\rho_n(\lambda_*)$ and a low Δ_n value for true reject decisions. Hence, a normalized reliability measure ω_n can be estimated by,

$$\omega_n = \frac{1}{\sum_i \gamma_i}\gamma_n, \tag{5.5}$$

where

$$\gamma_n = \begin{cases} \Delta_n & \text{closed} - \text{set} \\ (e^{(\rho_n(\lambda_*)+\Delta_n)} - 1) + (e^{(\kappa-\rho_n(\lambda_*)-\Delta_n)} - 1) & \text{open} - \text{set} \end{cases} \tag{5.6}$$

The first and second terms for open-set identification in γ_n are associated with the true

accept and true reject, respectively. The symbol κ stands for an experimentally determined factor to reach the best compromise between accept and reject scenarios. The κ value is set to 0.65 as it is found to be optimal for open-set speaker identification task in [18].

Chapter 6

EXPERIMENTAL RESULTS I: UNIMODAL PERFORMANCE

6.1 Database and Test Environment

The audio-visual database have been acquired using a Sony DSR-PD150P video camera at Multimedia, Vision and Graphics Laboratory (MVGL) of Koç University. The data acquisition system built at MVGL can be seen in Figure 6.1. Speaker identification and speech-reading experiments have been conducted using the MVGL-AVD database [57], which contains audio-visual data collected from a population of 50 speakers ($R = 50$). A view of the variation in our database is presented in Figure 6.2. The visual dataset has color video frames of size 720×576 pixels at a rate of 15 fps, each containing the frontal view of a speaker's head, and the audio stream has 16 kHz sampling rate. The database includes two distinct scenarios, that are the name ($\mathcal{D}_n$) and the digit ($\mathcal{D}_d$) scenarios. In the name scenario, 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 scenario, each subject utters ten repetitions of a fixed digit password 348 572. Both scenarios are used in the speaker identification experiments. As for speech-reading, each name uttered in the name scenario dataset is regarded as an isolated phrase, and a subset from this dataset, $\mathcal{D}_s \subset \mathcal{D}_n$, which includes more than 12 repetitions of each name utterance, is used as the testbed for our speech-reading experiments.

In the experimental studies, first an initial lip region of size 128×80 is segmented from each video frame, following the 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 considered for each lip segment. Hence each grid covers a pixel block of size 2×2. Following motion estimation and 2D-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-

Figure 6.1: Data acquisition system at Koç University.

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 3.4.1. Following motion estimation and 1D-DCT on the lip contour pixel locations, a feature vector of size $2M$, is obtained, where $M = 50$. The low-indexed DCT coefficients then provide us with the contour-based *FirstN* features, $\boldsymbol{f}^N_{\text{CTR}}$. The third candidate for the best lip feature representation is obtained by concatenating contour-based motion features with the 8 lip shape parameters, that is $\boldsymbol{f}^N_{\text{CTR}} + \boldsymbol{f}_{\text{SHP}}$. In addition to the lip motion features, recall that the lip texture features are the 2D-DCT coefficients of the lip intensity image whereas the audio features are the MFCC coefficients along with the first and the second derivatives.

The temporal characterization of the audio and the audio-visual modalities are performed by HMM structures. The HMM structures are implemented using the HTK tool version 3.3, where each class is represented by a word-level, 6-state left-to-right HMM structure. While the performance of the speaker identification system is measured using the equal error rate figure, that of the speech recognition system is presented with the recognition rate, that is the ratio of the true matches to the total number of trials.

In the following sections, we will present the unimodal recognition results. In Section 6.2, we will first compare optical flow and block matching techniques for motion estimation in the recognition sense. We will then consider each scenario one by one in more detail, and for each scenario we will provide the performances of the three lip motion feature representations, $\boldsymbol{f}^N_{\text{GRD}}$, $\boldsymbol{f}^N_{\text{CTR}}$ and $\boldsymbol{f}^N_{\text{CTR}} + \boldsymbol{f}_{\text{SHP}}$ in Section 6.3, Section 6.4 and Section 6.5. Later,

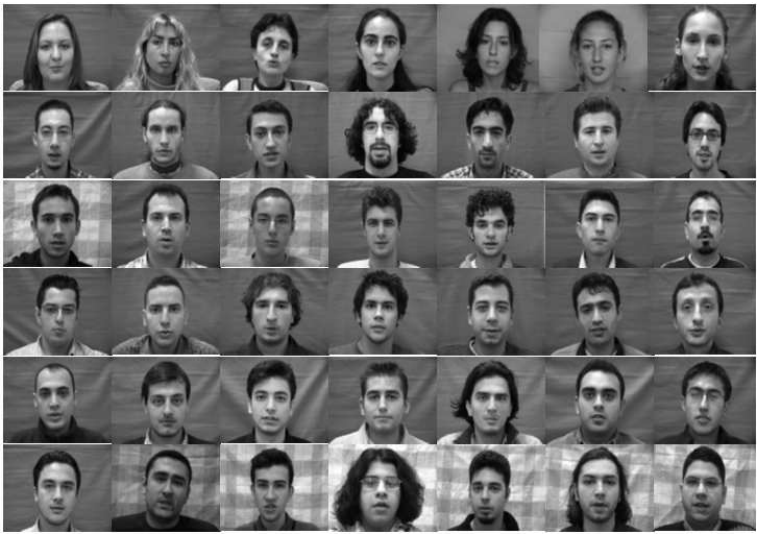


Figure 6.2: Sample subjects from the MVGL-AVD database.

experimental results based on the two-stage discrimination analysis of these lip motion features for each scenario are discussed in Section 6.6. The decision fusion results of the best lip motion and lip intensity features, and Bayesian feature selection are further discussed in Section 6.7 and Section 6.8 respectively.

6.2 Comparison of Optical Flow and Block Matching Techniques

Table 6.1 presents the unimodal EER performances of two different speaker identification scenarios, namely *Name* and *Digit* under different motion estimation techniques. It is observed that the lip motion features calculated by the block-matching method provide better EER performance as compared to the ones computed by the optical flow. The most reasonable explanation for this fact is the erroneous measurements introduced due to the theory and assumptions behind the optical flow computation. In addition, it is worth repeating that the optical flow estimation is carried out by the OpenCV©'s build-in function employing the Lucas-Kanade method described in Section 3.2.1. Hence the classical block matching method is computationally more expensive than the optical flow. So for these two motion estimation techniques there exist a trade-off between accuracy and computational load.

Table 6.1: Unimodal performance comparison of the optical flow and the block matching techniques under speaker identification scenarios.

Feature Type		Optical Flow	Block Matching
Name EER (%)	$\boldsymbol{f_{GRD}}$	8.4	6.8
	$\tilde{\boldsymbol{f}}_{\boldsymbol{GRD}}$	8.4	6.5
	LDA($\tilde{\boldsymbol{f}}_{\boldsymbol{GRD}}$)	7.6	5.2
Digit EER (%)	$\boldsymbol{f_{GRD}}$	14.8	12.8
	$\tilde{\boldsymbol{f}}_{\boldsymbol{GRD}}$	14.3	12.2
	LDA($\tilde{\boldsymbol{f}}_{\boldsymbol{GRD}}$)	7.6	5.2

6.3 Speaker Identification: Name Scenario

In the name scenario implementation, the $\mathcal{D}_n$ database is partitioned into two disjoint sets, $\{\mathcal{D}_{n_1}$ and $\mathcal{D}_{n_2}\}$, each having five repetitions from each subject in the database. The subsets $\mathcal{D}_{n_1}$ and $\mathcal{D}_{n_2}$ are then used for training and testing, respectively. 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 candidates, $\boldsymbol{f}^N_{\text{GRD}}$, $\boldsymbol{f}^N_{\text{CTR}}$ and $\boldsymbol{f}^N_{\text{CTR}} + \boldsymbol{f}_{\text{SHP}}$, are tested on the database. Figure 6.3 displays the EER performances with varying feature dimension N for speaker identification. We observe that the grid-based motion feature $\boldsymbol{f}^N_{\text{GRD}}$ achieves 6.8% EER, and outperforms the contour-based features. We also observe that the addition of lip shape information, $\boldsymbol{f}_{\text{SHP}}$, to the contour-based motion features, $\boldsymbol{f}_{\text{CTR}}$, improves the performance of contour-based features.

6.4 Speaker Identification: Digit Scenario

In the digit scenario, the $\mathcal{D}_d$ database is partitioned into two disjoint sets, $\{\mathcal{D}_{d_1}$ and $\mathcal{D}_{d_2}\}$, each having five repetitions of the same 6-digit number from each subject in the database. The subsets $\mathcal{D}_{d_1}$ and $\mathcal{D}_{d_2}$ are then used for training and testing, respectively. 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,

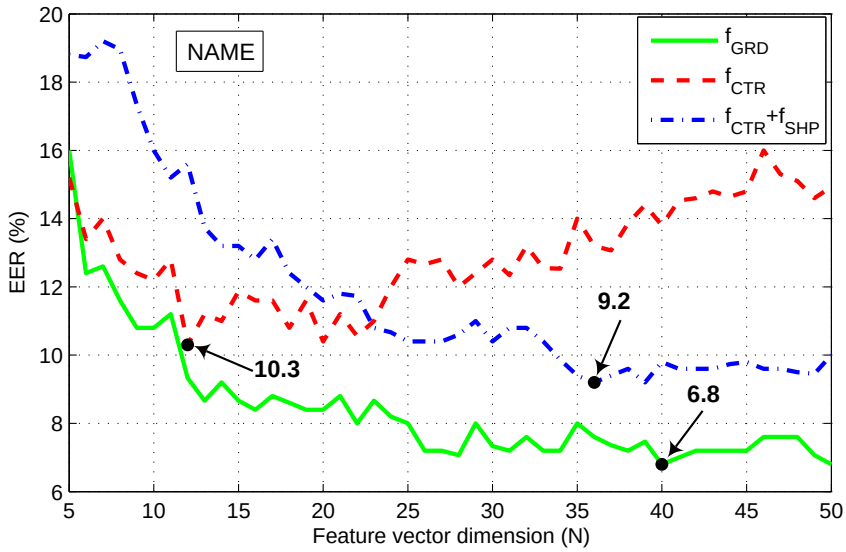


Figure 6.3: *Name* Scenario: The EER results for grid-based, contour-based and shape information-added motion features.

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 $N_r = 250$.

The three lip motion feature candidates, $\boldsymbol{f}^N_{\mathrm{GRD}}$, $\boldsymbol{f}^N_{\mathrm{CTR}}$ and $\boldsymbol{f}^N_{\mathrm{CTR}} + \boldsymbol{f}_{\mathrm{SHP}}$, are tested on the database. Figure 6.4 displays the EER performances with varying feature dimension N. We observe that the grid-based motion feature $\boldsymbol{f}^N_{\mathrm{GRD}}$ and the lip contour and shape based feature $\boldsymbol{f}^N_{\mathrm{CTR}} + \boldsymbol{f}_{\mathrm{SHP}}$ achieve the same minimum 12.8% EER, and outperforms the contour only feature $\boldsymbol{f}^N_{\mathrm{CTR}}$. Note that, speaker identification through digit scenario is harder, and the resulting EER performances are poorer than the name scenario performances.

6.5 Speech-Reading Scenario

In this scenario, the database $\mathcal{D}_s$ includes 35 different phrases, i.e., $R = 35$. Each phrase, which is actually the name of a speaker from the name database population, is repeated at least twelve times. 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 the same number of utterance repetitions.

Figure 6.5 displays performances of the three lip motion feature candidates, $\boldsymbol{f}^N_{\mathrm{GRD}}$, $\boldsymbol{f}^N_{\mathrm{CTR}}$

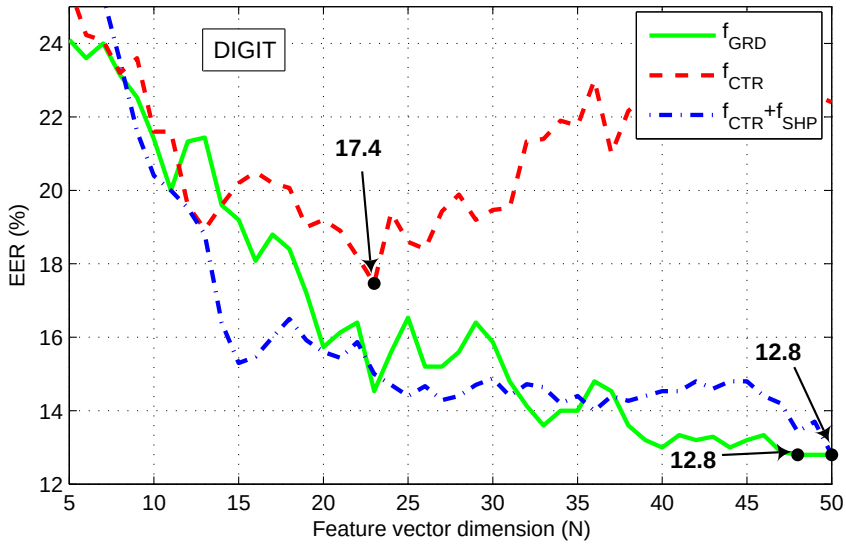


Figure 6.4: *Digit* Scenario: The EER results for grid-based, contour-based and shape information-added motion features.

and $\boldsymbol{f}^N_{\mathrm{CTR}} + \boldsymbol{f}_{\mathrm{SHP}}$. We observe that the lip contour and shape based feature $\boldsymbol{f}^N_{\mathrm{CTR}} + \boldsymbol{f}_{\mathrm{SHP}}$ achieves the best recognition rate, 70.48%. However, the lip contour only $\boldsymbol{f}^N_{\mathrm{CTR}}$ and lip motion based $\boldsymbol{f}^N_{\mathrm{GRD}}$ features perform quite close to this best recognition rate, which are respectively 69.52% and 67.62%.

6.6 Evaluation of Discrimination Analysis

The Bayesian discriminative feature selection method and the temporal LDA analysis have been applied to different lip motion feature representations in various combinations. The best EER and recognition rates attained are provided in Table 6.2. These values are obtained by choosing the feature size N as the one that maximizes the performance for each case. In Table 6.2, $\boldsymbol{f}^N$ 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

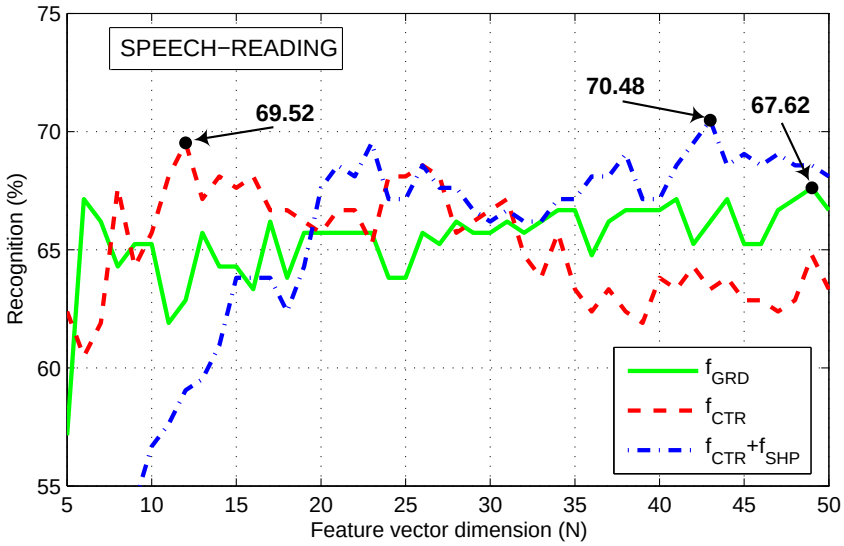


Figure 6.5: Speech-Reading: The recognition rates for grid-based, contour-based and shape information-added motion features.

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.

In Table 6.2, we observe that the best performances are obtained using the grid-based motion features for both speaker identification and speech-reading. Figure 6.6 plots the performances of the grid-based *FirstN* features, *DiscrimN* features and *DiscrimN* features with LDA at varying dimensions (10 to 50) for speaker identification and speech-reading scenarios. The key observations of these experiments are: i) *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 ii) As the feature vector dimension N increases the performance saturates, iii) The use of temporal LDA in addition to Bayesian discrimination, brings additional EER gain in speaker identification. However, this is not the case in speech-reading, where the temporal LDA may even degrade the recognition rate.

Table 6.2: Evaluation of two-stage discrimination analysis for lip motion and shape features under speaker identification and speech-reading scenarios.

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

In the *Name* scenario, the best EER performances for each feature set, *FirstN*, *DiscrimN* and *DiscrimN+LDA*, are 6.8%, 6.5% and 5.2% whereas in the *Digit* scenario, they are 12.8%, 12.2% and 5.2%. For speech-reading, the best recognition rate is measured for *DiscrimN* as 72.86%.

6.7 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 [7, 18], we form the intensity-based lip feature vector by scanning the 2D-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 6.3 for speaker identification (name and digit) and speech-reading.

The last row of Table 6.3 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 [18]. 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 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.

Table 6.3: 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-Reading
Intensity	5.60	1.74	62.86
Motion	5.20	5.20	72.86
Intensity⊕Motion	3.60	1.60	70.95

6.8 Discussions on the Bayesian Feature Selection

In this part of the experimental results, we will demonstrate how the discrimination power $D_N(\boldsymbol{f})$, as defined in (4.9), can be used to pre-estimate the relative recognition performances of different lip feature types. We will also provide some experimental details of the Bayesian

discriminative feature selection procedure applied to the grid-based ($\boldsymbol{f}_{\mathrm{GRD}}$) and to the contour-based ($\boldsymbol{f}_{\mathrm{CTR}}$) lip motion features. The left column of Figure 6.7 presents the discrimination power of *DiscrimN* ($\tilde{\boldsymbol{f}}^{N}_{GRD}$) 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 power estimates and the corresponding experimental performances match with each other, that is, the higher the discriminative power for a given feature type, the higher is the corresponding recognition performance.

The Bayesian discrimination analysis results in an ordering of the transform domain coefficients, where the first N of these coefficients are picked as the representing features. The indices of these discriminative coefficients are worth examining in more detail. Recall that the x- and y-direction grid-motion vectors are separately processed by 2D-DCT transformation and then the first 50 coefficients from x- and y-direction are concatenated into a single vector to be further processed by the Bayesian discrimination analysis. Figure 6.8 plots the indices of the first 50 discriminative coefficients from x- and y-direction DCT coefficients for different recognition scenarios. The non-principle coefficients are considered to be those having coefficient indices higher than 25. We observe that there are more number of valuable non-principle coefficients in the case of speaker identification as compared to speech-reading. From Figure 6.8 one can also observe that the number of non-principle coefficients computed from the x components of the lip motion vectors is more than those resulting from the y components.

A similar observation can be performed for the Bayesian discrimination of the grid-contour fusion based $\boldsymbol{f}_{\mathrm{GRD+CTR}}$ feature. Figure 6.9 plots the indices of the discriminative coefficients for the grid and contour based features, respectively on the left and right columns. On the right column of Figure 6.9, the first half of the coefficients are from x-direction, and the second half from y-direction. Note that, the number of discriminative coefficients from the grid-based features is significantly more than the contour-based features. In Figure 6.9, the contour-based features exhibit a poor discrimination, which is also validated with the poor experimental recognition results as presented in Figure 6.3, Figure 6.4 and Figure 6.5.

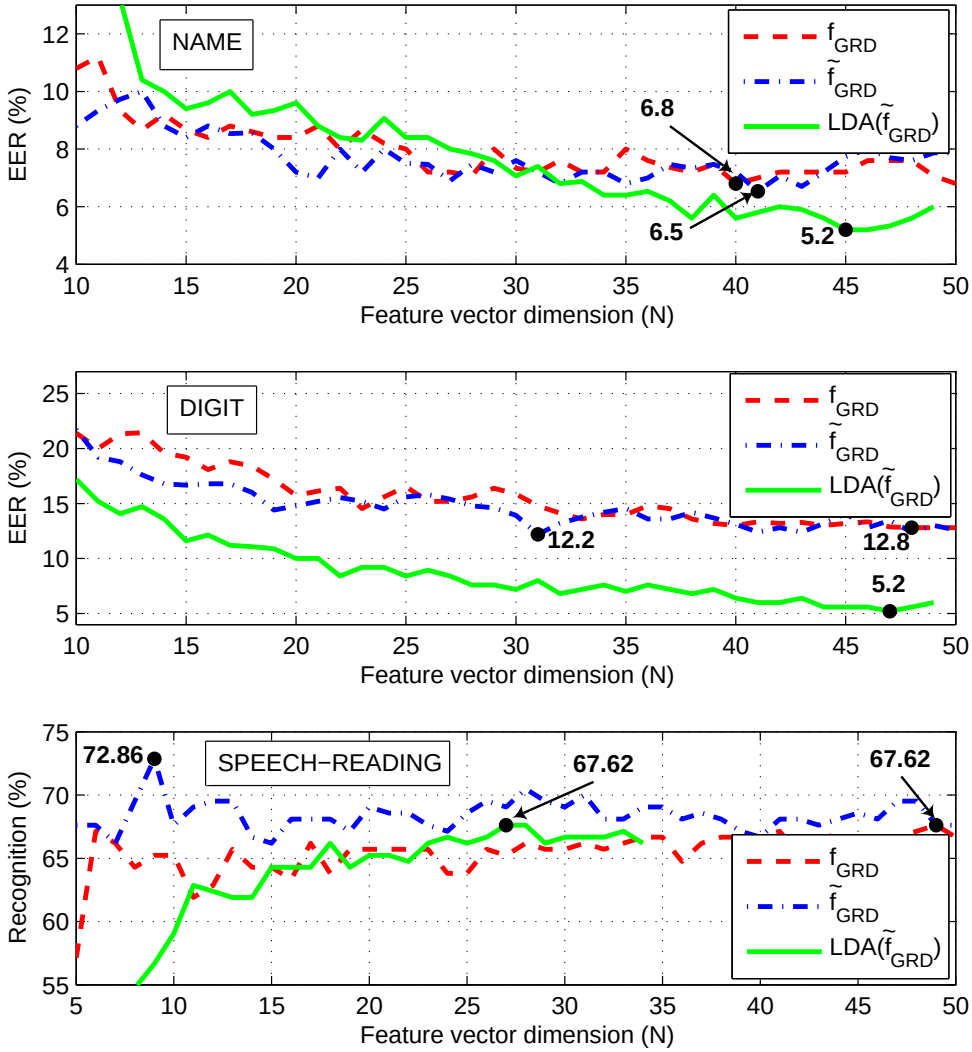


Figure 6.6: Speaker identification and speech-reading performance results for $\boldsymbol{f}_{GRD}^{N}$ (*FirstN*), $\tilde{\boldsymbol{f}}_{GRD}^{N}$ (*DiscrimN*) and $LDA(\tilde{\boldsymbol{f}}_{GRD}^{N})$ features with varying feature vector dimension N.

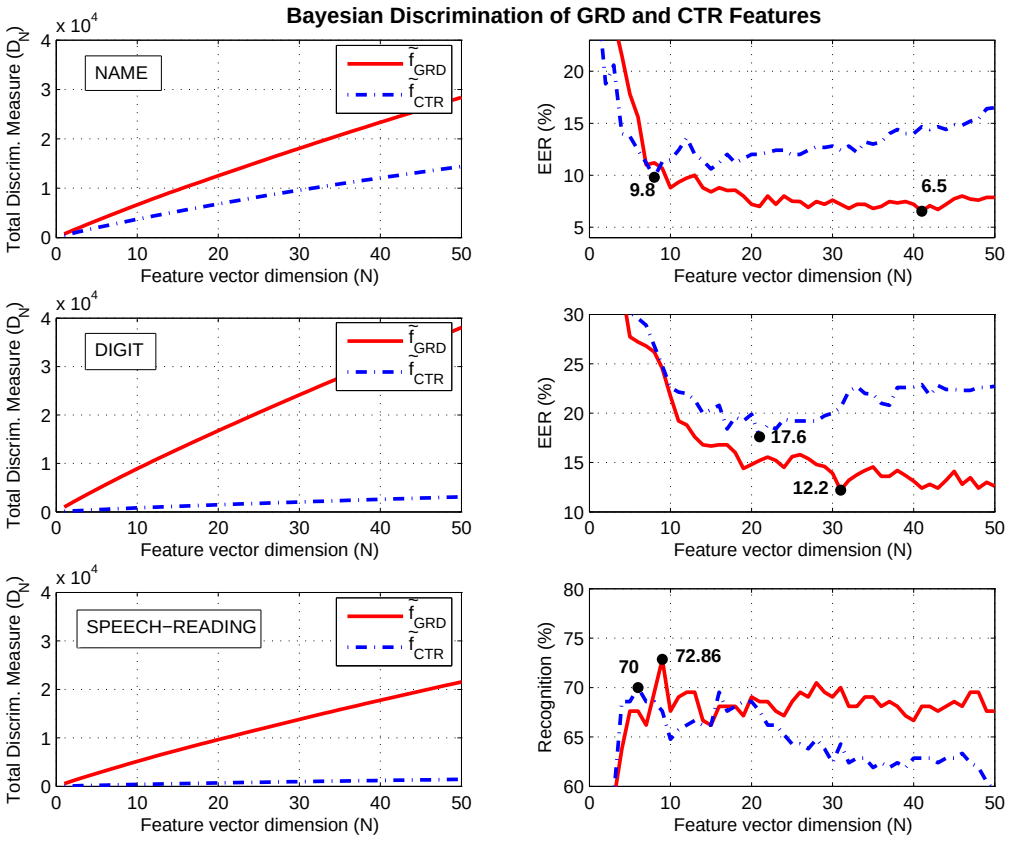


Figure 6.7: The discrimination powers ($D_N(\boldsymbol{f})$) and corresponding experimental performances of the grid-based and contour-based *DiscrimN* features at varying dimensions for name, digit and speech-reading scenarios.

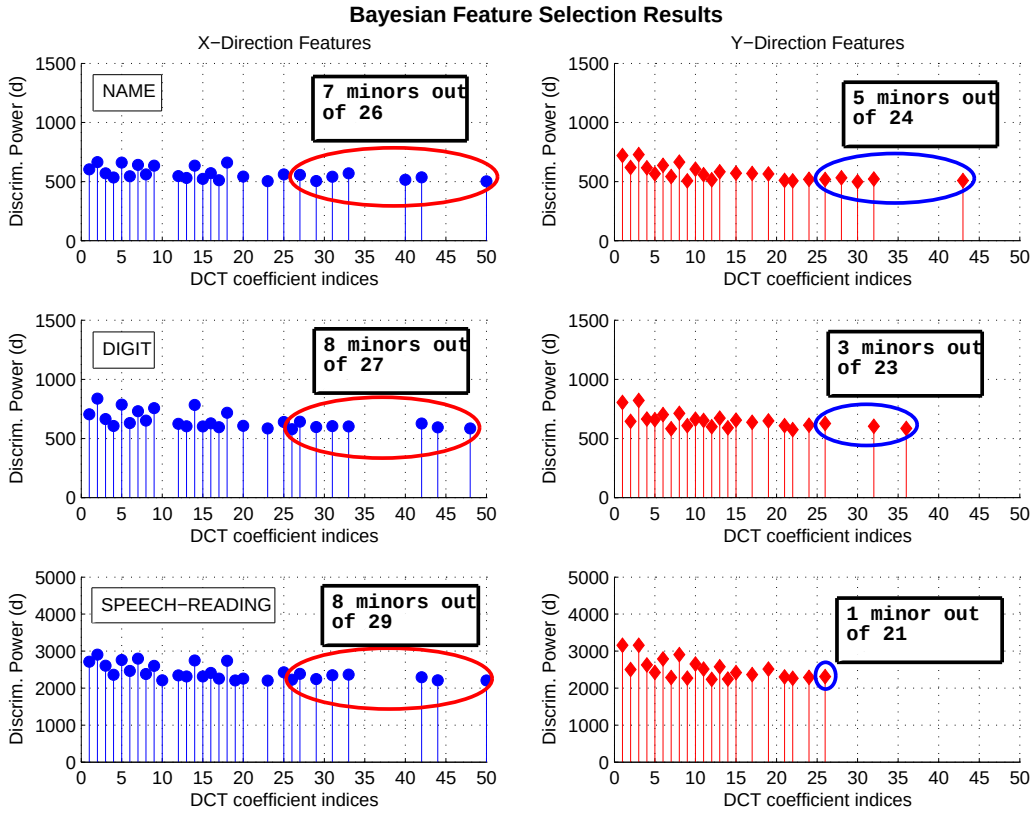


Figure 6.8: The DCT coefficients selected after the Bayesian discriminative feature selection for grid-based lip motion features.

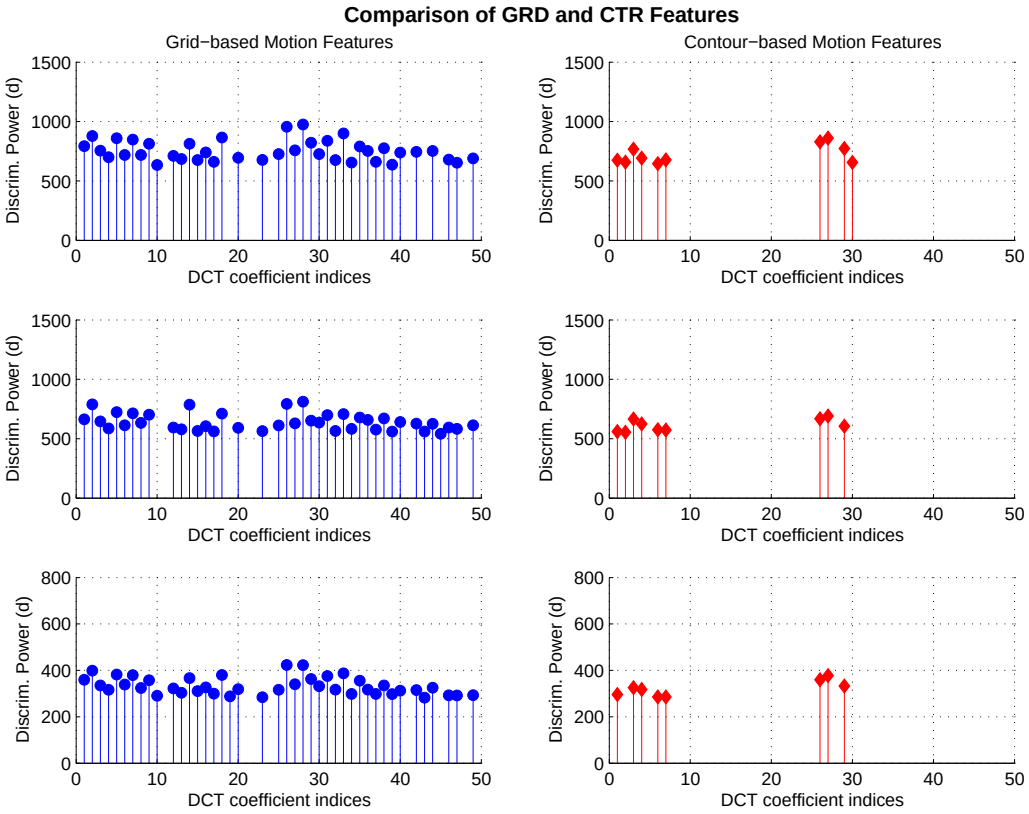


Figure 6.9: The DCT coefficients selected after the Bayesian discriminative feature selection for the combined grid and contour based lip motion features.

Chapter 7

EXPERIMENTAL RESULTS II: MULTIMODAL PERFORMANCE

In this chapter, the performance gain is presented by fusing 3 different modalities: i) the improved lip motion modality, ii) lip texture, and iii) audio, by RWS rule. The audio recordings are perturbed with varying levels of additive noise during the testing sessions to simulate adverse environmental conditions. The additive acoustic noise is picked to be either a mixture of office and babble noise or the car noise. Abbreviations and descriptions for the modalities and fusion techniques are given in Table 7.1.

Table 7.1: Abbreviations and descriptions for modalities and fusion techniques

A	Audio modality
L_t	Lip texture modality
L_m	Lip motion modality
$+$	Product rule
$\oplus$	RWS rule

7.1 Speaker Identification: Name Scenario

Table 7.2 presents the EER performance of the unimodal and multimodal speaker identification system (*Name*) for audio, lip texture and lip motion modalities. The audio modality is perturbed with the office&bubble noise. The EER performances of the lip texture and lip motion modalities are 5.6% and 5.2%, which are close to each other and better than the audio modality at 15 dB SNR and below. When the product rule and the RWS rule are applied to fuse pair of modalities or all the three modalities, the EER performance increases significantly. The RWS rule is observed to perform better than product rule, especially un-

der noisy conditions. The best EER performance is achieved with the fusion of all three modalities at 15 dB SNR and below. Above 15 dB SNR, the best performance is achieved with the fusion of lip texture and audio modalities.

Table 7.2: Speaker identification results for *Name* scenario: Equal error rates at varying office&bubble noise levels for different modalities and multimodal fusion structures.

EER (%)							
Source	Noise Level (dB SNR)						
Modality	clean	25	20	15	10	7	5
A	1.0	1.6	2.4	5.3	14.8	25.4	31.5
L_t	5.6						
L_m	5.2						
$L_m + A$	2.6	3.2	3.6	4.4	7.2	17.5	22.8
$L_m \oplus A$	0.8	1.2	1.8	3.2	5.6	13.6	19.2
$L_t + A$	0.4	0.4	0.8	2.0	4.4	11.2	15.9
$L_t \oplus A$	1.0	0.8	1.0	1.8	3.0	6.8	9.6
$L_m + L_t + A$	1.6	1.4	1.4	1.4	1.7	3.6	4.4
$L_m \oplus L_t \oplus A$	1.2	1.2	1.2	1.2	1.4	3.2	3.2

Table 7.3 presents the EER performance of the unimodal and multimodal speaker identification system (*Name*) for audio, lip texture and lip motion modalities, where the audio modality is perturbed with the car noise. The EER performances of the lip texture and lip motion modalities are again better than the audio modality at 10 dB SNR and below. Besides the fact that both decision fusion techniques yield performance improvement, the RWS rule is better as compared to the product rule, especially under noisy conditions. The best EER performance is achieved with the fusion of all three modalities at all noise levels.

Table 7.3: Speaker identification results for *Name* scenario: Equal error rates at varying car noise levels for different modalities and multimodal fusion structures.

EER (%)									
Source Modality	Noise Level (dB SNR)								
	clean	25	20	15	10	7	5	0	-5
A	1.0	1.2	1.6	4.8	13.2	18.8	22.4	30.8	39.7
L_t	5.6								
L_m	5.2								
$L_m + A$	2.6	2.2	2.8	3.4	6.0	10.8	12.0	15.1	22.4
$L_m \oplus A$	0.8	0.8	1.2	2.0	4.8	8.0	9.6	13.6	19.8
$L_t + A$	0.4	0.8	0.8	1.2	3.6	7.0	7.0	12.0	18.4
$L_t \oplus A$	1.0	0.8	1.2	2.0	2.8	4.8	5.0	6.8	10.8
$L_m + L_t + A$	1.6	1.2	1.2	2.0	1.6	2.0	2.8	4.8	4.6
$L_m \oplus L_t \oplus A$	1.2	0.8	1.0	1.6	1.2	1.6	2.4	4.0	4.2

7.2 Speaker Identification: Digit Scenario

Table 7.4 presents the EER performance of the unimodal and multimodal speaker identification system (*Digit*) for audio, lip texture and lip motion modalities. The audio modality is perturbed with the office&bubble noise. The EER performances of the lip texture and lip motion modalities are 1.74% and 5.2%. Since, in the digit scenario every subject utters the same six digit password, the audio modality suffers and the lip texture modality benefits with respect to the name scenario. When the product rule and the RWS rule are applied to fuse pair of modalities or all the three modalities, the EER performance increases significantly. The RWS rule is observed to perform better than product rule at all SNR conditions. The best EER performance is achieved with the fusion of all three modalities at all SNR levels.

Table 7.5 presents the EER performance of the unimodal and multimodal speaker identification system (*Digit*) for audio, lip texture and lip motion modalities, where the audio

Table 7.4: Speaker identification results for *Digit* scenario: Equal error rates at varying office&bubble noise levels for different modalities and multimodal fusion structures.

EER (%)							
Source	Noise Level (dB SNR)						
Modality	clean	25	20	15	10	7	5
A	2.4	3.4	6.9	12.2	24.9	33.1	37.1
L_t	1.74						
L_m	5.2						
$L_m + A$	2.4	2.4	2.4	4.0	10.4	18.0	23.2
$L_m \oplus A$	2.4	2.4	2.4	4.0	10.0	16.8	22.0
$L_t + A$	0.4	0.4	0.4	1.4	6.8	14.0	18.4
$L_t \oplus A$	0.4	0.4	0.4	0.8	4.0	10.0	13.8
$L_m + L_t + A$	0.8	0.8	1.2	1.2	2.6	4.2	5.2
$L_m \oplus L_t \oplus A$	0.4	0.4	0.6	0.8	2.4	3.8	5.2

modality is perturbed with the car noise. The RWS rule is again observed to perform better than product rule at all SNR conditions. The best EER performance is achieved with the fusion of all three modalities at all SNR levels. It is worth noting that at some noise levels, the EER performance of the multimodal system is 0%, which is mainly due to insufficient training-testing repetitions. The experiments can be repeated a number of times using different subsets from $\mathcal{D}_d$.

7.3 Speech-Reading Scenario

Table 7.6 presents the recognition performance of the unimodal and multimodal speech-reading system for audio, lip texture and lip motion modalities. The audio modality is perturbed with the office&bubble noise. The recognition performances of the lip texture and lip motion modalities are 62.86% and 72.86%. Since, the lip texture modality suffers to capture lip reading related information, the recognition rate of this modality is relatively

Table 7.5: Speaker identification results for *Digit* scenario: Equal error rates at varying car noise levels for different modalities and multimodal fusion structures.

EER (%)									
Source Modality	Noise Level (dB SNR)								
	clean	25	20	15	10	7	5	0	-5
A	2.4	2.6	2.8	5.6	11.0	18.6	24.2	37.2	45.4
L_t	1.74								
L_m	5.2								
$L_m + A$	2.4	2.4	2.8	2.8	5.8	6.6	8.6	17.1	25.9
$L_m \oplus A$	2.4	2.2	2.0	2.0	3.8	6.0	7.9	18.8	27.6
$L_t + A$	0.4	0.4	0.4	0.4	1.4	3.0	5.2	13.8	23.2
$L_t \oplus A$	0.4	0.4	0.4	0.4	0.8	1.6	2.8	10.8	18.8
$L_m + L_t + A$	0.8	0.8	0.4	0.4	0.4	0.8	1.2	1.6	1.8
$L_m \oplus L_t \oplus A$	0.4	0.4	0.0	0.0	0.0	0.4	0.6	1.6	4.2

poorer than the lip motion and audio modalities. When the product rule and the RWS rule are applied to fuse pair of modalities or all the three modalities, the recognition performance increases if the lip texture modality is not in the fusion. The best recognition performance is achieved with the RWS fusion of audio and lip motion modalities at all SNR levels.

Table 7.7 presents the recognition performance of the unimodal and multimodal speech-reading system for audio, lip texture and lip motion modalities, where the audio modality is perturbed with the car noise. Again when the product rule and the RWS rule are applied to fuse pair of modalities or all the three modalities, the recognition performance increases if the lip texture modality is not in the fusion. The best recognition performance is achieved with the RWS fusion of audio and lip motion modalities at all SNR levels.

Table 7.6: Speech-Reading results: Recognition rates at varying office&bubble noise levels for different modalities and multimodal fusion structures.

Recognition (%)							
Source	Noise Level (dB SNR)						
Modality	clean	25	20	15	10	7	5
A	90.00	88.57	87.62	86.67	80.00	62.86	39.05
L_t	62.86						
L_m	72.86						
$L_m + A$	86.19	84.28	84.28	84.28	80.95	72.38	63.33
$L_m \oplus A$	91.43	90.95	88.57	88.10	84.76	75.71	69.05
$L_t + A$	76.67	77.14	78.57	76.67	76.19	69.04	54.76
$L_t \oplus A$	76.67	77.14	75.24	74.76	73.33	68.57	61.69
$L_m + L_t + A$	80.95	81.42	80.95	81.90	79.04	75.71	69.52
$L_m \oplus L_t \oplus A$	78.57	78.57	76.19	77.14	74.28	72.38	68.10

Table 7.7: Speech-Reading results: Recognition rates at varying car noise levels for different modalities and multimodal fusion structures.

Recognition (%)									
Source	Noise Level (dB SNR)								
Modality	clean	25	20	15	10	7	5	0	-5
A	88.10	87.14	85.71	85.24	78.57	73.81	68.57	51.43	35.71
L_t	62.86								
L_m	72.86								
$L_m + A$	84.76	83.81	82.38	81.43	80.47	75.71	73.33	67.14	53.81
$L_m \oplus A$	90.00	88.57	86.67	85.71	82.85	79.05	75.24	73.33	60.47
$L_t + A$	77.62	78.10	77.14	76.66	74.76	71.90	68.57	61.90	45.71
$L_t \oplus A$	77.14	76.66	76.66	76.19	75.23	74.76	74.28	69.05	55.24
$L_m + L_t + A$	79.52	80.00	79.05	77.62	76.66	75.24	76.19	73.81	68.10
$L_m \oplus L_t \oplus A$	79.52	77.14	77.14	75.71	76.66	75.24	74.76	74.28	64.76

Chapter 8

CONCLUSIONS

Biometric person identification technologies focus on voice, face, iris and retina scans, signature strokes, fingerprint and gait as distinguishing source of personal information. However, state-of-the-art audio-visual speech recognition systems usually employ two critical source: speech signal and lip information. The lip motion information, which is highly correlated with the speech signal, has been extensively utilized in speech recognition. Despite the general belief that the lip motion possesses valuable biometric information, there have been few studies investigating this modality in speaker identification. More specifically, almost all of the existing systems employ the lip texture and/or geometry to model the lip motion. The use of the explicit lip motion, which is in fact what is meant by the lip information, is relatively rare. This has been the first issue in the speech/speaker recognition literature, that motivates us to investigate the lip motion modality. The second point open to debate is the optimal feature representation for the lip motion information. Determination of the best lip motion features has been the primary objective of this work. By obtaining the best lip motion features, i.e., the most discriminative features among classes, it is possible to maximize the unimodal recognition performance. However, no matter how successful the modality is, robustness has always been an issue for unimodal systems. More reliable and robust recognition systems should be build by fusing individual modalities. The audio information of speech and the temporal and visual characterization of lip constitute more information about the speech content and the identity of the speaker, making them good modalities to fuse. The necessity to build robust recognition systems directs us towards integrating visual features with audio.

Taking the outlined issues into account, we propose a new multimodal speaker/speech recognition system that integrates audio, with several lip modalities. We mainly focus on the explicit lip motion features in addition to or instead of the texture- and geometry-based lip features. For this purpose, we have investigated two kinds of lip motion representations:

firstly we compute the grid-based motion features within a bounding box around the lip region and thus, take the motion of the non-lip (skin) region into account. Secondly we calculate the contour-based motion features on the outer lip contour and discard the effect to the surrounding area. In addition to the explicit motion features on the outer contour, simplistic lip shape features are also extracted and concatenated with the contour-based motion feature vector to find out the contribution of geometric lip information. We have performed an additional investigation on the motion estimation techniques, namely optical flow and block matching, to determine the superior one in the recognition sense.

In our first experiments, it is observed that the motion features computed by the block matching technique provide better recognition rates as compared to the optical flow. As mentioned in Section 6.2, the theory and mathematical assumptions behind the optical flow computation may result inaccurate motion vectors, not only in length but also in direction. However, the optical flow is computationally less expensive than the block matching. Secondly, we have shown that for speaker/speech recognition, grid-based dense lip motion features are superior and more robust compared to contour-based lip motion features. This shows the importance of the skin region even if some erroneous vectors show up. We have also concluded that explicit lip motion is useful in addition to lip intensity and/or geometry. Explicit lip motion fused 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.

Recall that before applying the two-stage discrimination analysis, we first transform the motion data into DCT domain. This transformation 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.

For optimal lip motion feature representation, we have introduced a novel 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. It is interesting to see that after spatial Bayesian discrimination, a small set of DCT features that possess more discriminative power is formed regardless of their energy or coefficient index. 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 is beneficial for speaker identification, especially under the digit scenario. The LDA maps a given high dimensional feature vector to a subspace of reduced dimension that best describes the discrimination among classes. In speaker identification, the LDA is able to well discriminate among classes, i.e., different speakers, however, it collapses in speech-reading scenario since the classes are now some isolated phrases. In other words, the LDA cannot effectively reduce the feature dimension and the reduced feature vector is unable to catch the uttered phrases from the lip motion.

Apart from the efforts to maximize the unimodal performance of the explicit lip motion modality, we have fused the lip motion features with audio and lip texture to build a reliable and robust system that is able to cope with the real-life problems. The audio features are composed of the MFCCs along with the first and second derivatives whereas the lip texture features are the 2D-DCT coefficients of the gray-level lip images. Since the reliability of each independent source of information (lip motion, audio, lip texture) may vary under different light and acoustic conditions, our multimodal decision fusion strategy significantly improves the overall performance. The RWS decision fusion rule with the given reliability measures provides better results than the product rule as it introduces *a priori* information on the modality reliability. It is interesting that a successful system has been built for speaker identification without using the face modality, which is usually considered as indispensable in this problem.

There are further issues to be addressed. First, the lip region should be detected in a fully automatic way to allow complete implementations. There exist a number of ways to detect and segment the lip region however they usually suffer from translational/rotational invariance. The main concern behind the lip segmentation problem is to extract the lip from the mouth image. When lip motion analysis is the primary issue, lip images should be registered and carefully extracted using a static reference point, for instance the center

point in the image. Otherwise, the lip motion analysis cannot be carried out correctly. Second, the training and test database should be enriched both in terms of total population and variety for a more reliable performance analysis. The variety in database refers mainly to changing environmental conditions such as lighting and background, and to including video sequences where the head of the speaker may undergo arbitrary rigid motion. This would allow us to better measure the tolerance of our system to head rotation and changing illumination. Third, more modalities such as face, iris, can be added to the current system to obtain more robust solutions. All these issues should be further investigated.

Appendix A

SPEAKER IDENTIFICATION: NAME SCENARIO

POPULATION NAME LIST[1]

1. Müge Pirtini	21. Ozan Özkan	41. Engin Akın
2. Meral Robens	22. Murat Bahadır Soydan	42. Doğa Aydın
3. Işılay Talay	23. Mehmet Tuğrul Tekin	43. Uğur Dirim
4. Kıvılcım Büyükhatipoğlu	24. Umut Küçükkabak	44. Mustafa Yiğit
5. Leyla Mizrahi	25. Özgür Kaya	45. Uğur Üçgül
6. Selcan İşçi	26. Can James	46. Can Kızılkale
7. İdil Kokal	27. Buğra Tarı	47. Tanır Özçelebi
8. Ali Selim Aytuna	28. Eren Kalkan	48. Mustafa Can Filibeli
9. Ferda Ofli	29. Yücel Yemez	49. Mehmet Emre Yavuz
10. Abdullah Memiş	30. Ferit Ozan Akgül	50. Alper Kanak
11. Çağlar Ataman	31. Ömer Faruk Kurt	
12. Bülent Öktem	32. Seçkin Bayrak	
13. Ali Ekşim	33. Cihan Oruç	
14. Seçkin Kepenek	34. Davut Otar	
15. Alper Tolga Kocataş	35. Doruk Kayaalp	
16. Engin Erzin	36. Cengiz Ulubaş	
17. Baran Atılgan	37. Egemen Şentin	
18. Tahir Çelebi	38. Aykun Haddeler	
19. Ulaş Kemal Ayaz	39. Emre Yanık	
20. Birhan Güzel	40. Emir Erel	

[1]There are also out of population names in imposter recordings, these are: *Işıl Yıldırım, Sinem Bozkurt, Erhan Deniz, Suzay Özkan, Uğur Çelikyurt, Arda Gezdur, Harun Dericioğlu, Meral Turhan, Sezen Cürgül*

Appendix B

SPEECH-READING: LIST OF PHRASES

1. Ali Selim Aytuna
2. Ali Ekşim
3. Alper Kanak
4. Alper Tolga Kocataş
5. Abdullah Memiş
6. Baran Atılgan
7. Birhan Güzel
8. Bülent Öktem
9. Buğra Tarı
10. Çağlar Ataman
11. Can James
12. Can Kızılkale
13. Cihan Oruç
14. Cengiz Ulubaş
15. Davut Otar
16. Engin Erzin
17. Eren Kalkan
18. Emre Yanık
19. Umut Küçükkabak
20. Yücel Yemez
21. Ferit Ozan Akgül
22. Ferda Ofli
23. Işılay Talay
24. Mustafa Can Filibeli
25. Müge Pirtini
26. Meral Robens
27. Murat Bahadır Soydan
28. Mehmet Tuğrul Tekin
29. Mehmet Emre Yavuz
30. Özgür Kaya
31. Ozan Özkan
32. Seçkin Bayrak
33. Seçkin Kepenek
34. Tahir Çelebi
35. Tanır Özçelebi

BIBLIOGRAPHY

[1] J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, September 1997.

[2] Y. Yan J. Zhang and M. Lades, "Face recognition: Eigenface, elastic matching, and neural nets," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1423–1435, September 1997.

[3] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 586–591, September 1991.

[4] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On Combining Classifiers," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, March 1998.

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

[6] M. J. Tomlinson, M. J. Russell, and N. M. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 1996 (ICASSP'96)*, vol. II, pp. 821–824, 1996.

[7] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent Advances in the Automatic Recognition of Audio-Visual Speech," *Proc. of the IEEE*, vol. 91, no. 9, pp. 1306–1326, September 2003.

[8] S. W. Foo, Y. Lian, and L. Dong, "Recognition of Visual Speech Elements Using Adaptively Boosted Hidden Markov Models," *IEEE Trans. on Circuits and Systems For Video Technology*, vol. 14, no. 5, pp. 693–705, May 2004.

[9] T. Chen, "Audiovisual Speech Processing," *IEEE Signal Processing Magazine*, vol. 18, pp. 9–21, January 2001.

[10] L. G. Da Silveira, J. Facon, and D. L. Borges, "Visual speech recognition: a solution from feature extraction to words classification," *Proc. of the XVI Brazilian Symp. on Computer Graphics and Image Processing (SIBGRAPI 2003)*, pp. 399–405, 2003.

[11] 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. on Systems, Man and Cybernetics - Part A: Systems and Humans*, vol. 34, no. 4, pp. 564–570, July 2004.

[12] X. Zhang, C. C. Broun, R. M. Mersereau, and M. A. Clements, "Automatic Speechreading with Applications to Human-Computer Interfaces," *EURASIP Journal on Applied Signal Processing*, pp. 1228–1247, 2002.

[13] S. L. Wang, W. H. Lau, S. H. Leung, and H. Yan, "A real-time automatic lipreading system," *Proc. of the 2004 Int. Symp. on Circuits and Systems (ISCAS 2004)*, vol. 2, pp. 101–104, 2004.

[14] J. F. G. Perez, A. F. Frangi, E. L. Solano, and K. Lukas, "Lip reading for robust speech recognition on embedded devices," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 2005 (ICASSP'05)*, vol. I, pp. 473–476, 2005.

[15] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of Visual Features for Lipreading," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 198–213, February 2002.

[16] P. S. Aleksic, J. J. Williams, Z. Wu, and A. K. Katsaggelos, "Audio-Visual Speech Recognition Using MPEG-4 Compliant Visual Features," *EURASIP Journal on Applied Signal Processing*, pp. 1213–1227, 2002.

[17] S. Dupont and J. Luettin, "Audio-Visual Speech Modeling for Continuous Speech Recognition," *IEEE Trans. on Multimedia*, vol. 2, no. 3, pp. 141–151, September 2000.

[18] E. Erzin, Y. Yemez, and A. M. Tekalp, "Multimodal speaker identification using an

adaptive classifier cascade based on modality reliability," *to appear in IEEE Transactions on Multimedia*, 2004.

[19] N. A. Fox and R. B. Reilly, "Robust Multi-modal Person Identification with Tolerance of Facial Expression," *2004 IEEE Int. Conf. on Systems, Man and Cybernetics*, vol. 1, pp. 580–585, 2004.

[20] C. C. Broun, X. Zhang, R. M. Mersereau, and M. Clements, "Automatic speechreading with application to speaker verification," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 2002 (ICASSP'02)*, vol. I, pp. 685–688, 2002.

[21] L. L. Mok, W. H. Lau, S. H. Leung, S. L. Wang, and H. Yan, "Lip features selection with application to person authentication," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 2004 (ICASSP'04)*, vol. III, pp. 397–400, 2004.

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

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

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

[25] R. W. Frischholz and U. Dieckmann, "BioID: A multimodal biometric identification system," *Journal of IEEE Computer*, vol. 33, no. 2, pp. 64–68, February 2000.

[26] C. Sanderson and K. K. Paliwal, "Noise compensation in a person verification system using face and multiple speech features," *Pattern Recognition*, vol. 36, no. 2, pp. 293–302, February 2003.

[27] R. Brunelli and D. Falavigna, "Person identification using multiple clues," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 17, pp. 955–966, 1995.

[28] U. V. Chaudhari, G. N. Ramaswamy, G. Potamianos, and C. Neti, "Information fusion and decision cascading for audio-visual speaker recognition based on time-varying stream reliability prediction," *Proc. of the Int. Conf. on Multimedia & Expo 2003 (ICME 2003)*, vol. 3, pp. 9–12, July 2003.

[29] M. R. Civanlar and T. Chen, "Password-free network security through joint use of audio and video," *Proceedings of SPIE Photonic*, pp. 120–125, November 1996.

[30] D. D. Zhang, *Automated Biometrics*, Kluwer Academic Publishers, 2000.

[31] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach, Englewood Cliffs NJ*, Prentice-Hall Inc., 1982.

[32] P. Verlinde and G. Chollet, "Combining vocal and visual cues in an identity verification system using k-nn based classifiers," *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, December 1998.

[33] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

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

[35] J.-M. Odobez and P. Bouthemy, "Robust Multiresolution Estimation of Parametric Motion Models," *Journal of Visual Communication and Image Representation*, vol. 6, no. 4, pp. 348–365, December 1995.

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

[37] B. Lucas and T. Kanade, "An iterative image restoration technique with an application to stereo vision," *Proc. of the DARPA IU Workshop*, pp. 121–130, 1981.

[38] J.-Y. Bouguet, "Pyramidal implementation of the Lucas Kanade feature tracker: Description of the algorithm," *Intel Corporation, Microprocessor Research Labs*, 2000.

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

[40] H. E. Çetingül, Y. Yemez, E. Erzin, and A. M. Tekalp, "Robust lip-motion features for speaker identification," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 2005 (ICASSP'05)*, vol. I, pp. 509–512, March 2005.

[41] 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, June 2002.

[42] S-H. Leung, S-L. Wang, and W-H. Lau, "Lip Iimage Segmentation Using Fuzzy Clustering Incorporating an Elliptic Shape Function," *IEEE Trans. on Image Processing*, vol. 13, no. 1, pp. 51–62, January 2004.

[43] T. Wakasugi, M. Nishiura, and K. Fukui, "Robust lip contour extraction using separability of multi-dimensional distributions," *Proc. of 6th IEEE Int. Conf. on Automatic Face and Gesture Recognition (FGR'04)*, pp. 415–420, May 2004.

[44] S. L. Wang, W. H. Lau, S. H. Leung, and A. W. C. Liew, "Lip segmentation with the presence of beards," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 2004 (ICASSP'04)*, vol. III, pp. 529–532, 2004.

[45] N. Eveno, A. Caplier, and P.-Y. Coulon, "Accurate and Quasi-Automatic Lip Tracking," *IEEE Trans. on Circuits and Systems For Video Technology*, vol. 14, no. 5, pp. 706–715, May 2004.

[46] X. Zhang, R. M. Mersereau, M. A. Clements, and C. C. Broun, "Visual speech feature extraction for improved speech recognition," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 2002 (ICASSP'02)*, pp. 1993–1996, 2002.

[47] A. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy, "A couple hmm

for audio-visual speech recognition," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 2002 (ICASSP'02)*, pp. 2013–2016, 2002.

[48] J. Luettin, N. Thacker, , and S. Beet, "Statistical lip modelling for visual speech recognition," *Proceedings of the 8th European Signal Processing Conference*, pp. 10–13, 1996.

[49] N. Eveno, A. Caplier, and P.-Y. Coulon, "Jumping Snakes and Parametric Model for Lip Segmentation," *Proc. of the Int. Conf. on Image Processing 2003 (ICIP 2003)*, September 2003.

[50] C. C. Chibelushi, J. S. Mason, and F. Deravi, "Integration of acoustic and visual speech for speaker recognition," *Proc. 3rd Euro. Conf. Speech Communication and Technology*, vol. 1, pp. 157–160, 1993.

[51] T. J. Wark, S. Sridharan, and V. Chandran, "An approach to statistical lip modeling for speaker identification via chromatic feature extraction," *Int. Conf. on Pattern Recognition*, vol. 1, pp. 123–125, 1998.

[52] X. Huang, A. Acero, and H-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall, 2001.

[53] A. M. Martinez and A. C. Kak, "PCA versus LDA," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, February 2001.

[54] L. Rabiner and B-H. Juang, *Fundemantals of Speech Recognition*, Prentice Hall, 1993.

[55] H. E. Çetingül, Y. Yemez, E. Erzin, and A. M. Tekalp, "Discriminative lip-motion features for biometric speaker identification," *Proc. of the Int. Conf. on Image Processing 2004 (ICIP 2004)*, pp. 2023–2026, October 2004.

[56] H. Altınçay and M. Demirekler, "Undesirable effects of output normalization in multiple classifier systems," *Pattern Recognition Letters*, vol. 24, pp. 1163–1170, 2003.

[57] E. Erzin, Y. Yemez, and A. M. Tekalp, *DSP in Mobile and Vehicular Systems*, chapter Joint Audio-Video Processing for Robust Biometric Speaker Identification in Car, Kluwer Academic Publishers, October 2004.

VITA

HASAN ERTAN ÇETİNGÜL was born in İzmir, Turkey on January 1, 1980. He received his B.Sc. degree in Electrical and Electronics Engineering and his minor in General Management from Middle East Technical University, Ankara, Turkey, in 2003. From August 2003 to July 2005, he worked as a teaching and research assistant in Koç University, İstanbul, Turkey. At Koç University, he focused on *Multimodal Signal Processing for Speaker/Speech Recognition*, which has been supported by TUBİTAK and European FP6 Network of Excellence SIMILAR project. He also worked as a visiting researcher at Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland on August-September 2004 for SIMILAR project. He has submitted 2 journal publications and published several papers about *Lip Motion in Biometric Speaker Identification* in the following conferences: SİU'04 (Kuşadası, Turkey), MMSP'04 (Siena, Italy), ICIP'04(Singapore), ICASSP'05 (Philadelphia, USA), SİU'05 (Kayseri, Turkey).