

MULTIMODAL SPEAKER IDENTIFICATION WITH
AUDIO-VIDEO PROCESSING

by

736750

Alper Kanak

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

August, 2003

136750
ZC. YÜKSEKÖĞRETİM KURULU
DOKÜMANTASYON MERKEZİ

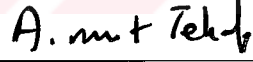
Koç University
Graduate School of Sciences and Engineering

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

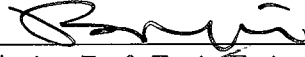
Alper Kanak

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



Assist. Prof. Engin Erzin


Assist. Prof. Yücel Yemez


Prof. Reha Civanlar


Prof. Bülent Sankur

Date: _____



To my family

ABSTRACT

In this thesis we present a multimodal text-dependent speaker identification system. The objective is to improve the recognition performance over conventional unimodal or bimodal schemes. The proposed system decomposes the information existing in a video stream into three modalities: voice, face texture and lip motion. Lip motion between successive frames is first computed in terms of eigenlip coefficients and then encoded as a feature vector. The feature vectors obtained along the whole stream are linearly interpolated to match the rate of the speech signal and then fused with mel frequency cepstral coefficients (MFCC) of the corresponding speech signal. The resulting joint feature vectors are used to train and test a Hidden Markov Model (HMM) based identification system. Face texture images are treated separately in eigenface domain and integrated to the system through decision-fusion. Experimental results are also included for demonstration of the system performance.

ÖZETÇE

Bu tezde, metne bağı çoklu ortamı bir konuşmacı tanıma sistemi tanıtılmıdır. Amaç, geleneksel tek ve çift ortamı tanıma sistemlerinin başarımlını arttırmaktır. Önerilen sistem, bir video akımında bulunan üç temel ortamı birleştirir: ses, yüz dokusu ve dudak hareketi. Video akımının her çerçevesi arasındaki dudak hareketi özduvak katsayıları ile hesaplandıktan sonra bu katsayılar bir öznitelik vektörüne dönüştürülür. Elde edilen öznitelik vektörleri, tüm akım boyunca doğrusal aradeğerlenerek ses işaretinin oranı ile eşleştirildikten sonra mel-frekans kepstral katsayılarla (MFCC) birleştirilir. Sonuçta elde edilen birleşik öznitelik vektörleri, Saklı Markov modeli tabanlı bir tanıma sisteminde eğitim ve sınaama amacıyla kullanılır. Yüz dokusu ise bir özyüz etki yöresinde ayrıca işlenerek karar füzyonu aşamasında sisteme katılır. Deneysel sonuçlar sistem başarımının gösterilmesi için teze eklenmiştir.

ACKNOWLEDGMENTS

First I would like to thank my supervisor Assoc. Prof. Engin Erzin and my co-advisors Prof. A. Murat Tekalp and Assoc. 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 those people who have shared time and given acquisitions for our video database.

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

TABLE OF CONTENTS

List of Tables	ix
List of Figures	x
Nomenclature	xi
Chapter 1: Introduction	1
1.1 State of the art	3
1.1.1 Unimodal Systems	3
1.1.2 Multimodal systems	4
1.2 System Overview and Contribution	6
Chapter 2: Theoretical Framework	8
2.1 Unimodal Identification	8
2.2 Bimodal fusion	10
2.2.1 Decision Fusion	11
2.2.2 Fusion using M -best Likelihoods	12
2.2.3 Data Fusion	13
2.2.4 Bayesian Decision	14
Chapter 3: Unimodal Speaker Identification	18
3.1 Audio-Only Speaker Identification	18
3.1.1 Speaker Identification Using HMMs	19
3.1.2 The Three Basic Problems of HMM	21
3.1.3 Audio Features	22
3.2 Face-only Speaker Identification	23
3.2.1 Eigenface Method	23

3.2.2	Identification from Face Sequences	25
3.2.3	Discussion	26
Chapter 4:	Multimodal Speaker Identification	28
4.1	Optical flow for Motion Analysis	29
4.1.1	Optical Flow Computation	29
4.2	Face and lip detection	30
4.3	Extraction of Lip Features	31
4.3.1	Eigenlips	31
4.4	Feature Fusion by Interpolation and Concatenation	33
4.5	Multimodal Speaker Identification System	34
4.5.1	Normalization	35
4.5.2	Multimodal Fusion	36
4.6	Discussion	37
Chapter 5:	Evaluation of Multimodal Speaker Identification Systems	39
5.1	Database and Test Environment	39
5.2	Performance of the Bimodal Bayesian Decision Fusion	41
5.3	Performance of the Multimodal Speaker Identification System	43
Chapter 6:	Conclusions	52
	Bibliography	55
	Vita	59

LIST OF TABLES

2.1	Finding EER and the threshold value that achieves EER using unimodal likelihood ratio test.	15
2.2	Bimodal Bayesian Decision Algorithm.	16
5.1	Speaker identification results of Bayesian Decision Fusion scenario.	43
5.2	Modality abbreviations for multimodal scenarios.	44
5.3	Speaker identification results of proposed multimodal system for the name scenario.	45
5.4	Speaker identification results of proposed multimodal system for the digit scenario.	46

LIST OF FIGURES

2.1	A typical ROC curve presenting EER, FAR and FRR.	10
2.2	A typical ROC surface	17
3.1	A Hidden Markov Model with three emitting states and continuous output distributions	20
4.1	Lip and face detection performance	32
4.2	(a) Optical flow vectors (b) Accumulated vector magnitudes (brighter regions correspond to fast-moving parts) (c) Thresholded image after vector accumulation (d) Elimination of isolated small regions and filling up small holes by applying morphological operations	33
4.3	Multimodal speaker identification scheme.	35
5.1	Data acquisition system in Koç University.	39
5.2	Sample subjects from the MVGL-SID database.	40
5.3	Bimodal Bayesian decision system.	42
5.4	Receiving operating curves for visual-only scenarios	47
5.5	Receiving operating curves for bimodal decision Fusion at various acoustic noise levels (Name scenario)	48
5.6	Receiving operating curves for bimodal decision Fusion at various acoustic noise levels (Digit scenario)	49
5.7	Receiving operating curves for the proposed multimodal system at various acoustic noise levels (Name scenario)	50
5.8	Receiving operating curves for the proposed multimodal system at various acoustic noise levels (Digit scenario)	51

NOMENCLATURE

HMM	Hidden Markov Model
MFCC	Mel-Frequency Cepstral Coefficient
DCT	Discrete Cosine Transform
PCA	Principal Component Analysis
FA	False Accept
FAR	False Acceptance Rate
FR	False Reject
FRR	False Reject Rate
EER	Equal Error Rate
ROC	Receiver Operating Curve
LPC	Linear Prediction Coding
HTK	Hidden Markov Model Toolkit
LS	Least-Squares

Chapter 1

INTRODUCTION

Biometric person recognition technologies include recognition of faces, fingerprints, voice, signature strokes, iris and retina scans, and gait. Person recognition in general encompasses two different, but closely related tasks: Identification and verification. The former refers to identification of a person from her/his biometric data from a set of candidates, while the latter refers to verification of a person's biometric data. It is generally agreed that no single biometric technology will meet the needs of all potential recognition applications. Although the performance of several of these biometric technologies have been studied individually, there is relatively little work reported in the literature on the fusion of the results of various biometric technologies [1].

A particular problem in multimodal biometric person identification, which has a wide variety of applications, is the speaker identification problem where basically two sources of information exist: audio signal (voice) and video signal. Speaker identification, when performed over audio streams, is probably one of the most natural ways to perform person identification. However, video stream is also an important source of biometric information, in which we have still images of biometric features such as face and also the temporal motion information such as lip movement, which is correlated with the audio stream. Most speaker identification systems rely on audio-only data [2]. Even assuming ideal noiseless conditions, such systems are far from being perfect for high security applications. The same observation is also valid for systems using only visual data, where poor picture quality or changes in lighting and pose can significantly degrade performance [3, 4].

A better alternative is the use of a combination of available modalities in a unified identification scheme. The first question to answer in designing such a unified system is to decide on which modalities to fuse. The word "modality" actually refers to a specific type of information that can be deduced from biometric signals. In this sense, speech, i.e. the

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, face and motion being the major ones. The dominant modality in the motion of a speaking person is naturally the lip movement which is highly correlated with audio whereas gesture (or gait) could also be interpreted as a different but less significant modality in the case of speaker identification. The second problem to address in a multimodal scheme is how to represent the raw biometric data for each modality with a meaningful set of features and, in conjunction with this, to find the best matching metric in the resulting feature space for classification. This step also includes a training phase through which each class is represented with a statistical model or a representative feature set. Curse of dimensionality, computational efficiency, robustness, invariance and discrimination capability are the most important criteria in selection of the feature set and the classification methodology for each modality. The third, and the final issue in a multimodal scheme is how to fuse different biometric signals. Different strategies are possible: One possible way is so-called "early integration" in which modalities are fused at data or feature level whereas in "late integration" decisions or scores resulting from each unimodal classification are combined for final conclusion [3]. When more than two modalities are available, a better alternative, that has not been addressed in the literature, is to make use of both strategies, i.e. to employ early integration and/or late integration where appropriate for bimodal fusion of different modality couples.

In this thesis, we will develop a multimodal speaker identification scheme that improves the performance of conventional unimodal systems. In doing this, we will address the issues and problems mentioned above in the previous paragraph. 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 address the unimodal identification problem, more specifically audio-only and face-only identification respectively; the problem of selecting appropriate feature set and the classification metric for each of the two modalities is considered in this chapter. The question "how to fuse" is addressed in Chapter 4. In this chapter, we describe a bimodal identification scheme that integrates the audio information with the lip motion modality. The overall multimodal scheme that incorporates finally the face texture is again presented in Chapter 4. Experimental results are given in Chapter 5 and the conclusions in

Chapter 6.

1.1 State of the art

A multimodal identification system can be thought of as integration of separate unimodal schemes; in our case these modalities are speech, face and lip movement. The choice of features to represent each modality and that of individual classification methodologies have a lot to do with previous research on especially speech-only and face-only recognition schemes. In this section we will briefly summarize the relevant unimodal and multimodal fusion literature so as to position our work with the others.

1.1.1 Unimodal Systems

Speaker identification through speech or voice appears to be one of the most natural and mature fields of biometric technology. In the last two decades strong and effective statistical tools have been developed, mainly for speech recognition, such as Hidden Markov Models (HMM) and Artificial Neural Networks (ANN) [5]. The same statistical tools are also used effectively for speaker identification problem [2]. In general we can consider two possible scenarios for speaker identification; text-dependent or text-independent. In the text-dependent scenario user-customized passwords are used for the identification task, where statistical tools such as HMM characterize temporal properties of the audio stream as well as the voice. In the text-independent scenario as the input is any free-text speech signal, voice can be characterized statistically using Gaussian Mixture Models (GMM), which are single state HMM structures. The spectral features of the speech signal need to be extracted to represent speech/voice in a statistical recognition system. The mel-scaled cepstral coefficients (MFCC) are known to be robust and effective features to represent the speech signal for speech recognition and speaker identification tasks [2].

The choice of features in face recognition is more controversial as compared to speech. Two broad categories exist in the literature: Geometry-based and intensity-based features [6]. Geometry-based features are in general more immune to changes in lighting and pose. However, they require image analysis for accurate extraction and localization of face features such as eyes, mouth, nose and eyebrows, that brings in certain robustness and computational problems. Thus techniques based on extraction of geometric features usually impose

constraints and assumptions on the general appearance and orientation of the face to be detected and recognized. Although they seem to be invariant to lighting and orientation, variations in these conditions perturb already the analysis task itself. On the other hand, intensity-based features are much easier to obtain. Since relevant techniques work simply on intensity values, they do not involve any analysis or localization task for the identification process. Although these techniques are very robust and computationally efficient, they are in general quite sensitive to lighting conditions and pose. A remedy for this drawback is the normalization of the lighting and pose prior to the identification phase, that requires as well, though not as intensively as geometric feature-based techniques, an image analysis process. Among many others, three popular approaches exist for the use of intensity-based features: Eigenface technique, elastic matching and neural networks [7]. Elastic matching handles better variations in lighting and pose but in turn computational cost is high. In general eigenface and elastic matching outperform neural net-based systems. Due to its efficiency, eigenface technique seems to be more preferable of these three approaches for practical implementations. The drawback associated with the invariance issues may not be very severe when the acquisitions are performed in relatively controlled environments for lighting and pose. Otherwise, such variations should be taken into account during the training of the eigenface classifier, or computationally less efficient techniques such as elastic matching should be employed. Another alternative is not to rely totally on the face image information but to support the identification process with other modalities such as speech if available; hence the need for multimodal identification schemes.

1.1.2 Multimodal systems

Existing multimodal speaker identification systems are mostly bimodal, integrating audio and face information as in [8, 9, 10], audio and lip information as in [11, 12, 13, 14, 15] or face and lip shape as in [16]. In [10], Sanderson et.al. present an audio-visual person verification system that integrates voice and face modalities and compares concatenative data-fusion with adaptive and non-adaptive decision fusion techniques, where adaptation takes into account the acoustic noise level of speech signal. Later in [8], enhanced PCA for face representation and fusion using SVMs and confidence measures are presented. Another audio-visual person identification system proposed in [9] uses a Maximum Likelihood Linear

Transformation (MLLT) based data-fusion technique. These related works do not address lip-motion as a biometric modality for person identification and they all do emphasize on the performance of data and decision fusion in separate. In an eigenface-based person identification system, Kittler et.al. use the lip-shape to classify face images to enhance the face recognition performance [16].

Another recent trend in bimodal fusion literature is to enforce the audio-only identification with the visual lip motion information. The information inherent in lip movement, which is a natural by-product of the speaking act, has so far been exploited mostly for the speech recognition problem, establishing a one-to-one correspondence with the phonemes of speech and the visemes of lip movement. It is quite natural to assume that lip movement would also characterize the identity of an individual as well as what the individual is speaking. In [17], it was demonstrated that lip movement also contained information about a person's identity. Lip movements while uttering the same phrase vary significantly from individual to individual, but they remain relatively consistent for the same person. Only few articles published so far in the literature incorporate lip information for the speaker identification problem [11, 12, 14, 15]. Although these works demonstrate some improvement over unimodal techniques, they use a decision-fusion strategy and hence do not fully exploit the mutual dependency between lip movement and speech.

In [14], audio features composed of cepstral coefficients are combined with visual features representing the motion of lip contours, to achieve speaker identification. The combination, or so-called multisensor data fusion, is done using principal component analysis or linear discriminant analysis which are accepted as pixel-based representation techniques. The implementation given in [15] uses the time variation of the lip height and width as visual features and the LPC coefficients as the audio features. The visual and audio features are then combined to form a single feature vector, with weighting that depends on the acoustic background noise. The weighting is chosen so that the weight assigned to the audio features gets smaller as the acoustic background noise level becomes higher. In order to match the sequence of extracted features to the database, dynamic time warping is performed. If the distance between the captured features and the prestored features after dynamic time warping falls below a prescribed threshold, a match is declared and the user is identified. The threshold is chosen so that the false acceptance rate and the false rejection rate are

approximately equal.

Both systems reported in [14] and [15] take an early integration approach, i.e., the audio features and the visual features are integrated before they are feed into the matching algorithm. In [18], a late integration approach is taken, in which the visual and audio features are first matched separately, and then the scores of the two matching modules are combined together to form the final decision for person identification.

The only work in the literature that addresses a multimodal speaker identification system, using speech, face and lip motion is the one presented in [19]. In this paper, the information coming from voice, lip-motion and face modalities are assumed to be independent of each other and thus the multimodal fusion is achieved by a simple decision mechanism. The face-only module involves a quite deal of image analysis to normalize and to extract salient features of the face whereas the lip movement is represented by DCT coefficients of the corresponding optical flow vectors in the lip region. Face and lip features are then stored as biometric templates and classified through a set of algorithms so-called synergetic computer. The acoustic information on the other hand is represented by cepstral coefficients that are then classified by vector quantization using a minimum distance classifier.

1.2 System Overview and Contribution

In this thesis we propose a robust text-dependent multimodal speaker identification scheme using speech, lip motion and face texture. An early integration of audio and visual features takes place by feature-fusion of speech and lip motion. The fused feature vector includes the lip motion features that are characterized by eigenlip coefficients transformed into an eigenspace domain and the speech features that are represented by mel frequency cepstral coefficients. The visual texture information, i.e. face images, is expressed in eigenface domain and integrated to the system through decision-fusion.

The thesis has two main contributions to the multimodal speaker identification problem:

1. Three modalities, i.e. speech, lip motion and face texture, are integrated to achieve a multimodal identification system in which a joint data/decision fusion scheme is used so as to exploit the correlations existing between different modalities.
2. A probabilistic framework is presented for decision fusion of independent modalities,

that uses M -best likelihoods resulting separately from each modality in proportion to the reliability of the individual classification task.



Chapter 2

THEORETICAL FRAMEWORK

A major problem of the biometric identification is the time varying nature of some key modalities, such as voice, face, etc. One possible solution to cope with this limitation is to combine several biometrics in a multimodal identity verification system. In a decision fusion approach, an identification system needs to fuse the partial decisions coming from different individual modalities [20].

The fusion of redundant information from different sources can reduce overall uncertainty and increase the accuracy of a classification system. Fusion can take place at two different stages in the recognition process. In early integration techniques the data is combined and then recognition is performed on this combined data. The most common method of early fusion is to concatenate the feature vectors from the different modes. This technique involves aligning and synchronizing the data so as to form one combined data stream. This fusion technique is called data fusion and can be implemented at the feature or signal level. In late integration techniques, the decisions which take the form of some sort of score or classification of each stream (for example a posterior probability or log likelihood) from each of independent classifiers are combined to produce a classification of the sequence. This kind of fusion is called decision fusion [3].

2.1 Unimodal Identification

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

$$\lambda^* = \arg \max_{\lambda_n} P(\lambda_n|\mathbf{f}) \quad (2.1)$$

Since $P(\lambda_n|\mathbf{f})$ is usually difficult to compute, one can rewrite (2.1) in terms of class-conditional probabilities. Using Bayes Rule, we have

$$P(\lambda_n|\mathbf{f}) = \frac{P(\mathbf{f}|\lambda_n)P(\lambda_n)}{P(\mathbf{f})} \quad (2.2)$$

Since $P(\mathbf{f})$ is class independent and assuming equally likely class distribution, $[P(\lambda_n) = \frac{1}{k} \forall k=1, \dots, N]$, (2.1) is equivalent to

$$\lambda^* = \arg \max_{\lambda_n} P(\mathbf{f}|\lambda_n) \quad (2.3)$$

Computation of class-conditional probabilities $P(\mathbf{f}|\lambda_n)$ needs a prior modelling step, through which a probability density function of feature vectors is estimated for each class by using available training data. This modelling step is also referred to as training phase.

In a speaker identification scheme, a reject mechanism is also required due to possible impostor identity claims. The class-conditional probability, or the likelihood, $P(\mathbf{f}|\lambda^*)$ in (2.3) in fact gives a measure of how likely the feature vector \mathbf{f} results from class λ^* . A possible reject strategy is thus to apply a constant threshold: if the resulting likelihood is larger than a predetermined threshold, the speaker's claim is accepted otherwise a reject decision is given. However, the optimal threshold value also depends on the likelihood of the claim being an impostor; thus rather than the likelihood itself, a likelihood ratio $\rho(\mathbf{f}|\lambda_n)$ in log domain is used for the accept or reject decision [21]:

$$\rho(\mathbf{f}|\lambda_n) = \log \frac{P(\mathbf{f}|\lambda_n)}{P(\mathbf{f}|\lambda_{\bar{n}})} = \log P(\mathbf{f}|\lambda_n) - \log P(\mathbf{f}|\lambda_{\bar{n}}) \quad (2.4)$$

where $\lambda_{\bar{n}}$ denotes the impostor class or anti-class for λ_n . Ideally, the impostor class model should be constructed by using all possible impostor observations for class n , which is practically infeasible to achieve. Thus in practice, two approaches are usually employed for approximating the impostor class model. The first one uses the universal background model which is estimated by using all available training data regardless of which class they belong to. The second approach is more accurate but less efficient and referred to as background model which cover all training data but those belonging to the underlying class n . Both approximations yield a kind of average model and thus the likelihood of being an impostor is expected to decrease as the unknown feature vector gets further to this average model in the feature space. The final decision strategy can be stated as follows:

$$\begin{array}{ll} \text{if } \rho(\mathbf{f}|\lambda^*) \geq \tau & \text{accept} \\ \text{otherwise} & \text{reject} \end{array} \quad (2.5)$$

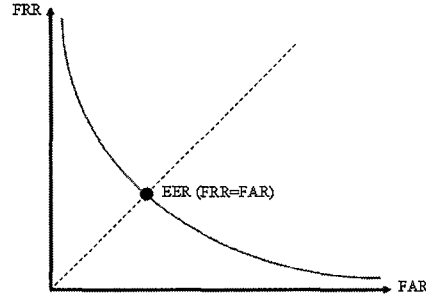


Figure 2.1: A typical ROC curve presenting EER, FAR and FRR.

where τ is the optimal threshold which is usually determined experimentally to achieve the desired false accept or false reject rate [22].

False accept and false reject are defined to accept an impostor (FA) and to reject a true client (FR) where they are the main performance variables for the person identification system. False accept rate (FAR) and false reject rate (FRR) are computed as:

$$\text{FAR} = \frac{\text{number of FA}}{\text{number of impostor claims}} \quad (2.6)$$

$$\text{FRR} = \frac{\text{number of FR}}{\text{number of client claims}} \quad (2.7)$$

One can easily observe that FAR and FRR are indirectly proportional for varying threshold values τ . Equal error rate (EER) is used as another important performance hacker, where it yields the operating point FAR equals FRR with the proper selection of threshold value τ . A sample receiving operating curve (ROC) is depicted in Fig 2.1, which plots FAR vs FRR for varying values of the threshold τ . EER is also located on this sample ROC curve.

2.2 Bimodal fusion

When two or more modalities exist, the selection of the appropriate fusion technique, whether data or decision fusion, should take into account how these modalities are correlated to each other.

2.2.1 Decision Fusion

Let \mathbf{f}_1 and \mathbf{f}_2 represent the unknown feature vectors corresponding to two different modalities. Then the joint class-conditional probability is given by [23]

$$P(\mathbf{f}_1, \mathbf{f}_2 | \lambda_n) = \frac{1}{2} [P(\mathbf{f}_1 | \lambda_n, \mathbf{f}_2)P(\mathbf{f}_2 | \lambda_n) + P(\mathbf{f}_2 | \lambda_n, \mathbf{f}_1)P(\mathbf{f}_1 | \lambda_n)] \quad (2.8)$$

If \mathbf{f}_1 and \mathbf{f}_2 are independent, we can write

$$\begin{aligned} P(\mathbf{f}_1 | \lambda_n, \mathbf{f}_2) &= P(\mathbf{f}_1 | \lambda_n) \\ P(\mathbf{f}_2 | \lambda_n, \mathbf{f}_1) &= P(\mathbf{f}_2 | \lambda_n) \end{aligned}$$

and thereby (2.8) reduces to the product of separate likelihoods:

$$P(\mathbf{f}_1, \mathbf{f}_2 | \lambda_n) = P(\mathbf{f}_1 | \lambda_n)P(\mathbf{f}_2 | \lambda_n) \quad (2.9)$$

Equation (2.9) can then be expressed in terms of log-likelihood ratios as the sum of the individual ratios:

$$\rho(\mathbf{f}_1, \mathbf{f}_2 | \lambda_n) = \rho(\mathbf{f}_1 | \lambda_n) + \rho(\mathbf{f}_2 | \lambda_n) \quad (2.10)$$

One critical issue here is that individual class-conditional probabilities, and the log-likelihood ratios as well, usually results in values with different ranges, with different means and variances. Thus prior to the fusion process, a common practice is to apply a normalization on resulting likelihoods, such as sigmoid normalization, and that is basically why decision fusion is sometimes referred to as *opinion* fusion.

Another issue is reliability of each likelihood contributing to final decision, that is not necessarily equal. One source of information may be more noisy than the other depending on the acquisition equipment and environment, or one modality may be more discriminative than the other. Thus commonly, a weighted sum of likelihoods is used:

$$\rho(\mathbf{f}_1, \mathbf{f}_2 | \lambda_n) = \omega_1 \rho(\mathbf{f}_1 | \lambda_n) + \omega_2 \rho(\mathbf{f}_2 | \lambda_n) \quad (2.11)$$

where ω_1 and ω_2 are weighting coefficients to be determined. There are various methods to estimate these coefficients which are ideally feature and class dependent such as noise estimation or measuring the experimental or statistical discriminative capability of each decision [24].

Looking back to (2.8), if \mathbf{f}_1 and \mathbf{f}_2 are correlated, we observe that (2.9) is no longer valid. In this case, (2.8) can be rewritten as

$$P(\mathbf{f}_1, \mathbf{f}_2 | \lambda_n) = \alpha_n P(\mathbf{f}_1 | \lambda_n) + \beta_n P(\mathbf{f}_2 | \lambda_n) \quad (2.12)$$

where

$$\begin{aligned} \alpha_n &= \frac{1}{2} P(\mathbf{f}_1 | \lambda_n, \mathbf{f}_2) \\ \beta_n &= \frac{1}{2} P(\mathbf{f}_2 | \lambda_n, \mathbf{f}_1) \end{aligned}$$

Exact computation of the joint feature densities $P(\mathbf{f}_1, \mathbf{f}_2 | \lambda_n)$ as formulated above is difficult to obtain since it requires a large amount of training data. A rough approximation of the weighting parameters α_n and β_n can be obtained by assuming that these parameters are only class-dependent and through the use of associative maps as described in [23].

2.2.2 Fusion using M -best Likelihoods

An important issue in multimodal fusion is to exploit all the information provided by each modality and the corresponding individual identification task. Recall that each identification task results in N likelihood scores for a population of N people. The common strategy is to fuse only the decisions with the best match, i.e. with the highest likelihood score, resulting from individual classifiers. However in the case of multimodal fusion, not only the highest score but the others as well may also carry useful information. Thus a better strategy is to let all the scores contribute to the final multimodal decision in proportion to their confidence levels.

When the total population N is very large however, processing all N likelihoods resulting from each modality identification becomes computationally inefficient, which is in fact not necessary at all. The likelihood ratios are mostly too small to really contribute to the decision and therefore can be neglected without any significant information loss. The strategy that we use is to arrange the list of scores resulting from each individual identification process in descending order and then let only the M -best matches, e.g. $M = 3$, contribute to the final multimodal decision.

Let \mathbf{M}_1 and \mathbf{M}_2 be the sets that separately include the M -best model matches of the

features \mathbf{f}_1 and \mathbf{f}_2 , respectively. The weighted sum of likelihoods is given as:

$$\rho(\mathbf{f}_1, \mathbf{f}_2|\lambda_n) = \begin{cases} \omega_1\rho(\mathbf{f}_1|\lambda_n) + \omega_2\rho(\mathbf{f}_2|\lambda_n) & \text{if } \lambda_n \in \mathbf{M}_1 \text{ and } \lambda_n \in \mathbf{M}_2, \\ \omega_1\rho(\mathbf{f}_1|\lambda_n) & \text{if } \lambda_n \in \mathbf{M}_1 \text{ and } \lambda_n \notin \mathbf{M}_2, \\ \omega_2\rho(\mathbf{f}_2|\lambda_n) & \text{if } \lambda_n \notin \mathbf{M}_1 \text{ and } \lambda_n \in \mathbf{M}_2. \end{cases}$$

The final weighted likelihood includes contributions from one or two sources depending on the presence of the source in the M -best list.

2.2.3 Data Fusion

In the early stages of processing when modalities are at signal or feature level, one can combine different modalities into one signal or feature. We call this information combining as data fusion. Data fusion is generally considered when the sources of information are correlated to each other, either in spectral or temporal domain. Data fusion can be simply the concatenation of different information sources if there is only a temporal correlation between sources, or one can further process the fused information to remove spectral correlations, if any.

In order to compute joint class-conditional feature probabilities $P(\mathbf{f}_1, \mathbf{f}_2|\lambda_n)$ by employing data fusion, the concatenated feature density functions must be directly computed:

$$P(\mathbf{f}_1, \mathbf{f}_2|\lambda_n) = P(\mathbf{f}_{12}|\lambda_n) \quad (2.13)$$

where $\mathbf{f}_{12} = [\mathbf{f}_1, \mathbf{f}_2]$.

As it is expected, data-fusion-based methods better exploit the temporal correlation of audio-video streams for robust performances, especially in the presence of environmental noise. But such systems do not always guarantee the overall performance to remain at least as good as the unimodal performance under low noise levels. On the other hand, the problem of curse of dimensionality may arise, that should be handled carefully and that may result in performance degradation. Data fusion generally is considered more appropriate for closely coupled and synchronized modalities, such as speech and lip movements. However, such a system tends not to generalize as well if it consists of modes that differ substantially in the time scale characteristics of their features, as is the case with speech and gesture input. Modelling complexity, computational intensity, and training difficulty typically are other problems associated with the data fusion approach. Due to the high dimensionality

of input features and high degree of freedom of system models, a large amount of training data is also required for building this type of system.

2.2.4 Bayesian Decision

In the unimodal scenario, a decision is taken based on the log likelihood ratio test as stated in Equation 2.5. In this decision mechanism, the desired FAR and FRR values could be reached by setting a proper threshold τ . It's also stated that a ROC curve that plots FAR vs FRR, could be extracted for varying values of τ . If one tries to find the proper threshold τ^* that achieves EER, that is FAR equals FRR operating point, ROC curve could be traced by incremental threshold values until the EER point is reached. A possible algorithm is given in Table 2.1 that finds the threshold value τ^* and EER value.

Once a likelihood ratio test threshold τ is set, one can claim that if the log likelihood ratio $\rho(\mathbf{f}|\lambda^*)$ is much larger or much smaller than τ , the confidence of the decision is stronger. Hence the absolute value of the difference between the likelihood ratio $\rho(\mathbf{f}|\lambda^*)$ and the threshold τ could be used as a measure of confidence ($C_{\mathbf{f}}$),

$$C_{\mathbf{f}} = |\rho(\mathbf{f}|\lambda^*) - \tau|. \quad (2.14)$$

In the bimodal scenario, the confidence measure could be used beneficially in the decision fusion if we have enough a priori information on the two different modality streams. Let us define a bimodal scenario with two different modalities, \mathbf{f}_1 and \mathbf{f}_2 . There are two streams of log likelihood ratios $\rho(\mathbf{f}_1|\lambda^*)$ and $\rho(\mathbf{f}_2|\lambda^*)$, correspondingly. If we have a priori information such that the identification performance of first modality \mathbf{f}_1 is much better than the second modality \mathbf{f}_2 under controlled conditions (such as low acoustic noise, frontal face stream, etc.), then the weighted use of the decision that is coming from first modality would be beneficial for the decision fusion. Keeping this fact in mind a Bayesian decision system is built. In this system decision is taken in two stages. In the first stage, a decision is taken if the the confidence measure $C_{\mathbf{f}_1}$, that is coming from the reliable modality \mathbf{f}_1 , is high enough, otherwise a decision is taken with respect to the second modality \mathbf{f}_2 in the second stage. The algorithm for the bimodal Bayesian decision fusion that finds the EER and the corresponding thresholds is given in Table 2.2.

In speaker identification systems, performance of voice identification is superior when

```

for each  $\tau$  in a threshold range
{
  init TA = FA = FR = TR = 0
  for  $k = 1$  to Number of subjects
  {
    if  $\rho(\mathbf{f}^k|\lambda^*) \geq \tau$ 
      if  $\mathbf{f}^k \in \lambda^*$ 
        True Accept TA++
      else
        False Accept FA++
    else if  $\rho(\mathbf{f}^k|\lambda^*) < \tau$ 
      if  $\mathbf{f}^k \in \lambda^*$ 
        False Reject FR++
      else
        True Reject TR++
  }
  FAR=FA/Number of subjects
  FRR=FR/Number of subjects
  if FAR=FRR
     $\tau^* = \tau$ , EER = FAR
}

```

Table 2.1: Finding EER and the threshold value that achieves EER using unimodal likelihood ratio test.

there is no acoustic noise. But this is not the case in general. In such noisy environments, audio-visual stream may help us as the second source of information. Although visual data are sensitive to rotation, image size and light conditions, the overall performance of the audio-visual system is superior than audio-only performance when there is heavy acoustic noise in the environment. In our bimodal Bayesian decision fusion algorithm, practically, \mathbf{f}_1 refers to voice stream and \mathbf{f}_2 refers to audio-visual stream. A typical ROC curve for this scheme is depicted in Fig 2.2. In this figure FAR and FRR represents two different surfaces, where their intersection forms an EER line in this space. An optimal EER could be picked as the minimum on this line.

```

n = 0
for each  $\tau_1$  in a threshold range
{
  TA=FA=TR=FR=0,  $\Gamma[\ ] = \{0\}$ 
  for each  $\tau_2$  in  $\tau_1 - \delta \leq \tau_2 \leq \tau_1 + \delta$  such that  $\delta$  runs from 0 to  $\Delta$ 
  {
    for k = 1 to Number of subjects
      if  $\rho(f_1^k|\lambda^*) \geq \tau_1 + \delta$ 
        if  $f_1^k \in \lambda^*$ 
          True Accept TA++
        else
          False Accept FA++
      else if  $\tau_1 - \delta \leq \rho(f_1^k|\lambda^*) \leq \tau_1 + \delta$ 
        if  $\rho(f_2^k|\lambda^*) \geq \tau_2$ 
          if  $f_2^k \in \lambda^*$ 
            True Accept TA++
          else
            False Accept FA++
        else  $\Rightarrow \rho(f_2^k|\lambda^*) < \tau_2$ 
          if  $f_2^k \in \lambda^*$ 
            False Reject FR++
          else
            True Reject TR++
      else  $\Rightarrow \rho(f_1^k|\lambda^*) < \tau_1 - \delta$ 
        if  $f_1^k \in \lambda^*$ 
          False Reject FR++
        else
          True Reject TR++
    }
  FAR=FA/Number of subjects
  FRR=FR/Number of subjects
  if FAR=FRR {
     $\Gamma[n] = [\tau_1, \tau_2]$ 
    ER[n++] = FAR
  }
}
EER =  $\min_i ER[i]$ 
 $\Gamma^* = \Gamma[\arg \min_i (ER[i])]$ 
}

```

Table 2.2: Bimodal Bayesian Decision Algorithm.

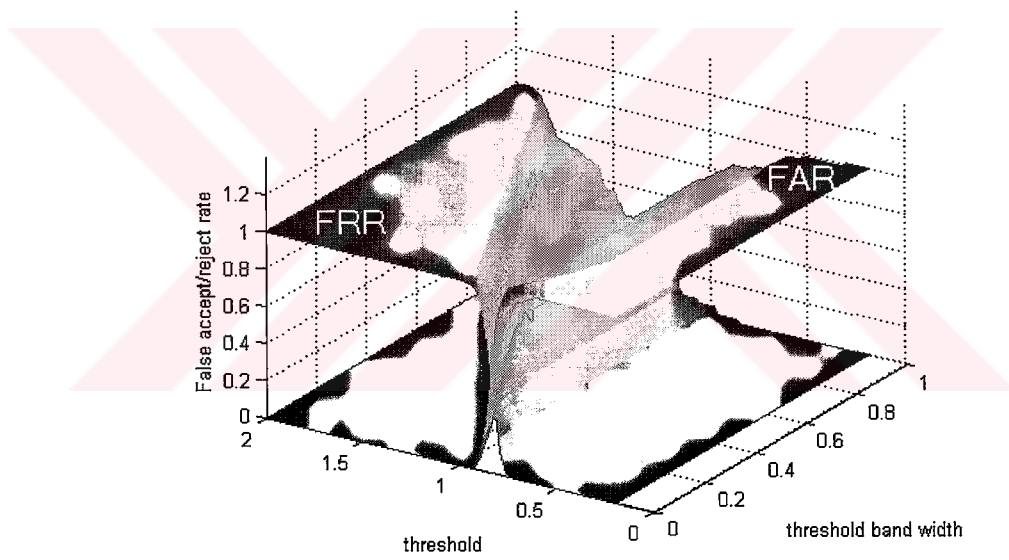


Figure 2.2: A typical ROC surface

Chapter 3

UNIMODAL SPEAKER IDENTIFICATION

Speaker identification through speech and face is one of the natural and mature technology that tries to mimic human's perception system for person identification. In the last two decades strong and effective tools have been developed for both speaker and face recognition, such as Hidden Markov Models (HMM), Artificial Neural Networks (ANN), Eigenface approach and elastic matching [5, 7]. In this chapter the theory of HMM for speaker recognition and the theory of Eigenface for face recognition is covered. The proposed multimodal speaker identification systems will be based on these audio and video based unimodal systems.

3.1 Audio-Only Speaker Identification

Template-based approaches to speaker identification suffer from variation in speech signal's spectral properties. The number of reference patterns needed to cover all the variations during a speech is too hard to handle. That's why statistical approaches stand more efficient in speaker identification problems.

Today the Hidden Markov Model (HMM) based statistical approaches are dominating. HMM 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 speaker identification problem.

3.1.1 Speaker Identification Using HMMs

Hidden Markov Models [25] are reliable structures to model human hearing system, and thus they are widely used for speech recognition and speaker identification problems [5, 25, 2, 26]. The temporal characterization of an audio-video stream can successfully be modelled 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 $\mathbf{S} = \{s_1, s_2, \dots, s_N\}$.
- $\mathbf{A} = \{a_{ij}\}$ is the matrix of state transition probabilities where a_{ij} is the probability of making a transition from state i to j , such that $a_{ij} = P(q_{\tau+1} = \mathbf{S}_j | q_{\tau} = \mathbf{S}_i)$, where q_{τ} is the state at time τ . The state transition probabilities are assumed to be time independent.
- $\mathbf{B} = \{b_j(\mathbf{f})\}$ is the vector of observation probabilities associated with each emitting state j , with $b_j(\mathbf{f}) = P(\mathbf{f} | q_{\tau} = \mathbf{S}_j)$.
- $\mathbf{\Pi} = \{\pi_i\}$ is the vector with the initial state probabilities of entering the model at state i such that $\pi_i = P(q_1 = \mathbf{S}_i)$.

A HMM can now be represented by the compact parameter set $\lambda = (\mathbf{A}, \mathbf{B}, \mathbf{\Pi})$. 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 $\mathbf{F} = (\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_K)$, given the model λ is defined as,

$$P(\mathbf{F}|\lambda) = \sum_{\text{all } \mathbf{q}} P(\mathbf{F}, \mathbf{q}|\lambda), \quad (3.1)$$

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

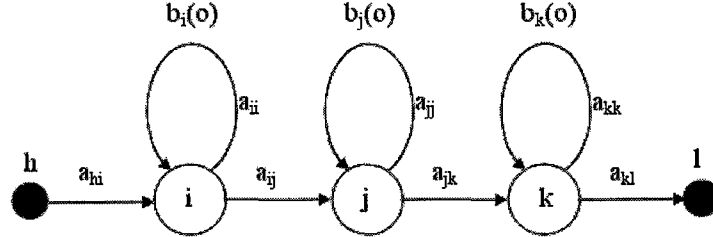


Figure 3.1: A Hidden Markov Model with three emitting states and continuous output distributions

model as,

$$P(\mathbf{F}, \mathbf{q}|\lambda) = P(\mathbf{F}|\mathbf{q}, \lambda)P(\mathbf{q}|\lambda), \quad (3.2)$$

where

$$P(\mathbf{F}|\mathbf{q}, \lambda) = b_{q_1}(\mathbf{f}_1)b_{q_2}(\mathbf{f}_2)\cdots b_{q_K}(\mathbf{f}_K), \quad \text{and}$$

$$P(\mathbf{q}|\lambda) = \pi_{q_1}a_{q_1q_2}a_{q_2q_3}\cdots a_{q_{K-1}q_K}.$$

The resulting likelihood function from Equation 3.1 will be in the form of,

$$P(\mathbf{F}|\lambda) = \sum_{\text{all } \mathbf{q}} \pi_{q_1}b_{q_1}(\mathbf{f}_1)a_{q_1q_2}b_{q_2}(\mathbf{f}_2)a_{q_2q_3}\cdots b_{q_{K-1}}(\mathbf{f}_{K-1})a_{q_{K-1}q_K}b_{q_K}(\mathbf{f}_K), \quad (3.3)$$

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

$$b_j(\mathbf{f}_k) = \sum_{l=1}^L \omega_{jl}\mathcal{N}(\mathbf{f}_k; \mu_{jl}\Sigma_{jl}) \quad (3.4)$$

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 $0 < \omega_{jl} \leq 1$.

In this work a word-level continuous-density HMM structure is built for the speaker identification task using the HTK library [27]. Each speaker in the database population is modelled using a separate HMM and is represented with the feature sequence that is extracted over the audio-video stream while uttering the secret phrase. First a world or universal background HMM model $\tilde{\lambda}$ is trained over the whole training data of the population. Then using the world HMM model as the initial state, each HMM associated to a

speaker λ_n is trained over some repetitions of the audio-video utterance of the corresponding speaker.

In the identification process, hypothesis testing is performed between the best match of the population and the world model for the given audio-video utterance of an unknown subject. The subject is either rejected or identified to be the speaker with the best match based on a likelihood ratio test. The likelihood ratio for the identification of n -th person can be derived from Equation 2.4 and is given as,

$$\rho(\mathbf{F}|\lambda_n) = \log P(\mathbf{F}|\lambda_n) - \log P(\mathbf{F}|\tilde{\lambda}). \quad (3.5)$$

3.1.2 The Three Basic Problems of HMM

Given the HMM structure, there are three basic problems that need to be solved to effectively address real-world applications. These are:

- Given the observation sequence $\mathbf{F} = (f_1, f_2, \dots, f_K)$, and a model $\lambda = (\mathbf{A}, \mathbf{B}, \mathbf{\Pi})$, how do we efficiently compute $P(\mathbf{F}|\lambda)$, the probability of the observation sequence, given the model?
- Given the observation sequence and the model λ , how do we choose a corresponding state sequence $q = (q_1, q_2, \dots, q_K)$ that is optimal in some sense (best explains the observations)?
- How do we adjust the model parameters $\lambda = (\mathbf{A}, \mathbf{B}, \mathbf{\Pi})$, to maximize $P(\mathbf{F}|\lambda)$?

First problem is the evaluation problem. We can also view the problem as one of scoring how well a given model matches a given observation sequence. Problem 2 is the one in which we attempt to uncover the hidden part of the model, that is, to find the *correct* state sequence. For practical situations, we usually use an optimality criterion to solve this problem as best as possible. Viterbi algorithm [28] presents an effective solution for this problem. The last problem is the one in which we attempt to optimize the model parameters to best describe how a given observation sequence comes about. The observation sequence used to adjust the model parameters is called a training sequence as it is used in the training phase of the HMM. The training problem is the crucial one for most applications of HMMs, because it

allows us to optimally adapt model parameters to observed training data. The solutions of these three problems can be found in [5].

3.1.3 Audio Features

Feature extraction converts the speech waveform to some type of parametric representation (at a considerably lower information rate) for further analysis and processing which is referred as the signal-processing front end.

The speech signal is a slowly time-varying signal (called quasi-stationary). When examined over a sufficiently short period of time (5 ~ 100 ms), its characteristics are fairly stationary. However, over long periods of time (on the order of 1/5 seconds or more) the signal characteristic change to reflect the different speech sounds being spoken. Therefore, the short-time spectral analysis is the most common way to characterize the speech signal.

A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), and others. Mel frequency cepstral coefficients (MFCC) give good discrimination of speech data; hence they are widely used to represent audio streams in HMM-based speech recognition and speaker identification systems. In our system, the speech signal which is sampled at 16 kHz is analyzed on 25 ms frame basis by frame shifts of 10 ms. Each 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 [5]. The resulting MFCC features are derived using discrete cosine transform over log-scaled filter-bank energies e_i :

$$c_j = \frac{1}{N_M} \sum_{i=1}^{N_M} e_i \cos \left[(i - 0.5) \frac{j\pi}{N_M} \right], \quad \text{for } j = 0, 1, \dots, L - 1 \quad (3.6)$$

where N_M is the number of mel-scaled filter banks and L is the number of MFCC features that are extracted. The MFCC feature vector for the k -th frame is defined as,

$$\mathbf{c}_k = [c_0 \ c_1 \ \dots \ c_{L-1}]^T. \quad (3.7)$$

It has been known that instantaneous changes in the spectrum yields valuable information for the recognition and identification tasks [2, 5]. The first delta MFCC feature vector $\Delta \mathbf{c}_k$

for the k -th frame is defined to incorporate the instantaneous changes in the spectrum,

$$\Delta \mathbf{c}_k = \frac{\sum_{d=-D}^D d \mathbf{c}_{k+d}}{\sum_{d=-D}^D d^2}, \quad (3.8)$$

where the second delta MFCC feature vector $\Delta \Delta \mathbf{c}_k$ is defined as the first delta of $\Delta \mathbf{c}_k$ vector. The audio feature vector \mathbf{f}_a^k for the k -th frame is formed as a collection of MFCC feature vectors including the first and the second delta MFCCs [2]:

$$\mathbf{f}_a^k = [\mathbf{c}_k \quad \Delta \mathbf{c}_k \quad \Delta \Delta \mathbf{c}_k]. \quad (3.9)$$

3.2 Face-only Speaker Identification

One of the major tasks in achieving a multimodal speaker identification system is to exploit the visual information in the video signal of a speaking person as well as the audio information. Motion, more specifically lip movement, is one of the modalities that can be extracted from the face sequence. We will consider lip motion as a separate modality in Chapter 4. Now in this chapter, we will focus on the texture modality and describe our methodology to identify the face of a speaking person from a set of still images sampled from her face texture sequence. Our methodology will be based on the well known eigenface technique and the theoretical framework presented in Chapter 2.1.

3.2.1 Eigenface Method

The eigenface technique [4], or more generally principal component analysis [29], has proven itself as an effective and powerful tool for recognition of still faces. The core idea is to reduce the dimensionality of the problem by obtaining a smaller set of features than the original dataset of intensities. Every image is expressed as a linear combination of some basis vectors, i.e. eigenimages that best describe the variation of intensities from their mean. These basis vectors define an eigenspace with reduced dimension.

The eigenspace of face images is calculated by identifying the eigenvectors of the covariance matrix derived from a set of training images. The eigenvectors corresponding to non-zero eigenvalues of the covariance matrix form an orthonormal basis for the N -dimensional eigenspace. The mathematical procedure is as follows:

Each image is first stored in a vector of size N :

$$\mathbf{x}^j = [x_1^j \cdots x_N^j] \quad (3.10)$$

The image vectors are then mean centered by subtracting the mean from each image vector:

$$\bar{\mathbf{x}}^j = \mathbf{x}^j - \mathbf{m} \quad (3.11)$$

where

$$\mathbf{m} = \frac{1}{M} \sum_{j=1}^M \mathbf{x}^j \quad (3.12)$$

The vectors $\bar{\mathbf{x}}^j$ are combined, side-by-side, to create a data matrix of size $N \times M$, where M is the number of images:

$$\mathbf{X} = [\bar{\mathbf{x}}^1 | \bar{\mathbf{x}}^2 | \dots | \bar{\mathbf{x}}^M] \quad (3.13)$$

The $N \times N$ covariance matrix $\mathbf{\Omega}$ of the data matrix \mathbf{X} given by

$$\mathbf{\Omega} = \mathbf{X} \mathbf{X}^T \quad (3.14)$$

has up to N eigenvectors \mathbf{v}_i associated with N eigenvalues such that

$$\mathbf{\Omega} \mathbf{V} = \mathbf{\Lambda} \mathbf{V} \quad (3.15)$$

where \mathbf{V} is the $N \times N$ matrix of eigenvectors:

$$\mathbf{V} = [\mathbf{v}_1 | \mathbf{v}_2 | \dots | \mathbf{v}_N] \quad (3.16)$$

and $\mathbf{\Lambda}$ is the $N \times N$ diagonal matrix of the associated eigenvalues. The eigenvectors are sorted, high to low, according to their associated eigenvalues. The eigenvectors with the largest p eigenvalues, $p \ll N$, are the eigenfaces, i.e. the basis vectors of the eigenspace of dimension p . When a given image is projected onto this lower dimensional eigenspace, a set of p eigenface coefficients is obtained, that gives a parameterization for the distribution of the signal.

The classification is performed in the eigenface feature domain. Once the eigenspace is created by using all the images in the training set as described above, each image $\bar{\mathbf{x}}^j$ of the training set is projected onto the eigenspace by

$$\mathbf{f}^j = \mathbf{V}^T \bar{\mathbf{x}}^j \quad (3.17)$$

and the resulting projection \mathbf{f}^j becomes the feature vector representing the corresponding speaker class λ_j , assuming that each speaker class has one single face image in the training set.

In order to classify an unknown face image \mathbf{y} , the image is first centered by subtracting the mean image evaluated in Eq. 3.12 and then projected onto the same eigenspace defined by V :

$$\mathbf{f} = \mathbf{V}^T(\mathbf{y} - m) \quad (3.18)$$

The test image \mathbf{y} is assigned to the class λ^* with the feature vector \mathbf{f}^j that is found to be the closest to \mathbf{f} in the feature space:

$$\lambda^* = \arg \min_{\lambda_j} \|\mathbf{f} - \mathbf{f}^j\| \quad (3.19)$$

where $\|\cdot\|$ is the Euclidean distance metric.

3.2.2 Identification from Face Sequences

In the case of speaker identification, rather than a single image, a sequence of face images is available for an unknown speaker to be recognized. A number of images, say K_1 , can be sampled from this sequence and can be used to enforce the identification process. The eigenface coefficients w_l , $l = 1, \dots, p$, when computed for every image i of a given sequence, constitute the face texture feature vector (subscript t denotes the word *texture*) that we will denote by \mathbf{f}_t^i , $i = 1, \dots, K_1$:

$$\mathbf{f}_t^i = [\omega_1, \omega_2, \dots, \omega_p]. \quad (3.20)$$

The face images in the training set are all used first to obtain the eigenspace. Note that the training set contains a number of images as well, say K_2 , from each speaker class λ_n . Let \mathbf{f}_t^{jn} , $j = 1, \dots, K_2$, denote the feature vectors of these images belonging to the class λ_n in the training set. Then the minimum distance d_n between these two sets of feature vectors can be used for hypothesis testing of the unknown person with the speaker class λ_n :

$$d_n = \min_{i,j} \|\mathbf{f}_t^i - \mathbf{f}_t^{jn}\| \quad (3.21)$$

The distance metric defined in (3.21) can also be modelled with probabilistic approach by making use of Gibbs distribution [30]: Given the face texture feature vector set $\{\mathbf{f}_t^i\}$, the class conditional probability of the feature set can be written as

$$P(\{\mathbf{f}_t^i\}|\lambda_n) = \frac{1}{\kappa} e^{-d_n/\sigma} \quad (3.22)$$

where $\kappa = \sum_d e^{-d/\sigma}$ and σ is the decay coefficient of the Gibbs distribution function.

The log likelihood ratio as defined in Equation 2.2 requires the definition of a universal background class. For this, we will adapt the faceness measure defined by the authors in [4]. According to this measure, the distance of the eigenspace origin to the eigenface coefficient vector for a given image determines how likely the image is a face. Thus in our case, the eigenspace origin will be used as the representative feature vector of the face universal background class. By using (3.22), we can now define the log likelihood ratio:

$$\rho(\{\mathbf{f}_t^i\}|\lambda_n) = \frac{\tilde{d} - d_n}{\sigma} \quad (3.23)$$

where \tilde{d} is the distance of the feature vector \mathbf{f}_t^i (that yields the minimum distance d_n) to the universal background model. The constant σ can normally be set to 1. But we will use this constant later in Chapter 4 during decision fusion for normalization of individual likelihood scores. Note for the moment that log likelihood ratio ρ_t for face texture takes values in the interval $[0, \tilde{d}_{\max}/\sigma]$, assuming $\tilde{d} > d_n$, and \tilde{d}_{\max} is the maximum value of \tilde{d} that can be determined experimentally using the training data.

The log likelihood ratio in (3.23) is computed for each class λ_n . In the case of unimodal face-only identification, the class that gives the maximum likelihood is the best match, and if this best-match ratio remains above a certain threshold, the unknown speaker can be assigned to the corresponding class, otherwise rejected. These likelihood ratios will later be used in the multimodal fusion process in Chapter 4.

3.2.3 Discussion

As an appearance-based approach, eigenface recognition method has several advantages. First raw intensity data are used directly for learning and recognition without any significant low-level or mid-level processing. Second, no knowledge of geometry of faces is required which increases the complexity of the algorithm. On the other hand, data compression is achieved by the low-dimensional subspace representation. These advantages reflect the power of eigenface approach in ease of implementation. However, the experimental results also demonstrate some serious limitations of eigenface representation method for face recognition under different conditions.

First, the method is very sensitive to scale, therefore, a low-level preprocessing is still necessary for scale normalization. Secondly, since the eigenface representation is a pixel-

based approach, in a least-squared sense, its recognition rate decreases under varying pose and illumination. Third, though the eigenface approach is shown to be robust when dealing with expression and glasses, these experiments were made only with frontal views. The problem can be far more difficult when there exists extreme change in pose as well as in expression and disguise. Fourth, since the face images tested in the experiments are taken with different backgrounds, this will seriously deteriorate the recognition performance. In such cases, a segmentation process has to be considered.

Additionally, the eigenface recognition method bears some common disadvantages due to its “appearance-based” nature. First, learning is very time-consuming, which makes it difficult to update the face database. Second, recognition is efficient only when the number of training data is large enough. The variations in pose and illumination in the training dataset also make the system more reliable and robust.

In controlled environments, as in our case, the consequences of the above disadvantages of the eigenface technique may not be very dramatic. In other cases however, eigenface-only identification is not reliable alone; hence the need for using other available modalities, such as audio and motion, in a multimodal fusion scheme to improve the overall system reliability. Instead of a single face image, using a number of images from the same individual, as in our case, may also serve for enforcing the identification process under changing light and pose conditions.

Chapter 4

MULTIMODAL SPEAKER IDENTIFICATION

No speaker identification system is error free. The reason for this may be various. Errors, i.e. false accept or false alarms, may source from inadequate acquisition, noise interference or due to discrimination incapability of the selected features. The motivation behind a multimodal system is to compensate such errors specific to a given modality and to enforce the overall decision. When audio is missing or corrupted by noise, visual information can be used as the dominant modality or visa versa.

There has been considerable research on speaker recognition with audio-only features. In Chapter 3 we have presented an audio-only HMM-based speaker identification system. However, the performance of such a unimodal system can be improved with the integration of visual data, which is relatively a new research area necessitating new fusion and feature representation techniques. In this chapter, we will address the fusion of audio information with lip movement. Lip movement, being part of the visual data, is highly correlated with speech and carries useful information about the identity of a speaking individual. Afterwards, a multimodal fusion and identification scheme will be proposed that integrates the face texture modality to this audio-lip bimodal system.

Two problems will be addressed for the audio-lip bimodal system: What features to use for representing lip movement and how to fuse them with audio information. In this respect, we will propose an eigenlip based feature representation technique in which lip images are transformed onto a lip space obtained by training a huge set of lip frames corresponding to each speaking individual. The advantage of this representation as compared to geometric techniques in the literature is its high correlation with the audio features. In order to extract lip frames during a speaking act, we propose a face and lip detection mechanism which is based on optical flow method for motion analysis of image sequences.

4.1 Optical flow for Motion Analysis

A fundamental problem in the processing of image sequences is the measurement of optical flow (or image velocity). The optical flow is a vector field which is defined as the apparent motion of the brightness pattern [33]. The goal is to compute the 2-D motion field from spatiotemporal patterns of image intensity. The optical flow in this sense is an approximation of the motion field which can be computed from time-varying image sequences. The error of this approximation is small at points with high spatial gradient and exactly zero only for translational motion or for any rigid motion such that the illumination direction is parallel to the angular velocity. Once this approximated motion field is computed, the measurements of image velocity can be used for a wide variety of tasks.

4.1.1 Optical Flow Computation

Optical flow computation techniques devised by the computer vision community can roughly be divided into two major classes: *differential techniques* and *matching techniques*. Differential techniques are based on the spatial and temporal variations of the image brightness at all pixels, and can be regarded as methods for computing optical flow. Matching techniques, instead, estimate the disparity of special image points between successive frames [33]. Differential techniques compute velocity from spatiotemporal derivatives of image intensity or filtered versions of the image (using low-pass or band-pass filters).

Optical flow is defined as an apparent motion of image brightness, $I(x, y, t)$, that changes in time to provide an image sequence, then two main assumptions can be made:

1. Brightness $I(x, y, t)$ depends on coordinates x, y in greater part of the image.
2. Brightness of every point of a moving or static object does not change in time.

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

$$I(x + dx, y + dy, t + dt) = I(x, y, t) + \frac{\partial I}{\partial x} dx + \frac{\partial I}{\partial y} dy + \frac{\partial I}{\partial t} dt + \dots \quad (4.1)$$

According to the second assumption, Equation: 4.1 becomes:

$$I(x + dx, y + dy, t + dt) = I(x, y, t) \quad (4.2)$$

and

$$\frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt + \dots = 0 \quad (4.3)$$

If we divide Equation 4.3 by dt and define $\frac{dx}{dt} = u$ and $\frac{dy}{dt} = v$, we obtain the following equation which is called *optical flow constraint equation*:

$$\frac{\partial I}{\partial t} = \frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v \quad (4.4)$$

Here u and v are components of optical flow field in x and y coordinates respectively. Since Equation 4.4 has more than one solution, more constraints are required.

Lucas-Kanade Method

Using the optical flow equation for group of adjacent pixels and assuming that all of them have the same velocity, we can make a system of linear equations. In a non-singular system for two pixels we can compute a velocity vector to solve the system. However, combining equations for more than two pixels is more effective. We might get a system that has no solution; yet we can solve it roughly, using the least square method. We will use weighted combination of equations. This method involves the solution of 2x2 linear system.

$$\begin{aligned} \sum_{x,y} W(x,y)I_xI_yu + \sum_{x,y} W(x,y)I_y^2v &= -\sum_{x,y} W(x,y)I_yI_t \\ \sum_{x,y} W(x,y)I_x^2u + \sum_{x,y} W(x,y)I_xI_yv &= -\sum_{x,y} W(x,y)I_xI_t \end{aligned}$$

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.

4.2 Face and lip detection

The first step in extracting visual features is to detect face and lip regions. We assume that the acquired images contain the face of a speaking person with a stationary background. A possibility here would be using a simple change detection algorithm. Such simple algorithms are computationally attractive; however they are usually very sensitive to noise,

changing light and possible small camera movements. Thus we propose an optical flow based detection technique that gives more accurate and reliable results. Optical flow vectors are first computed between successive frames of the video sequence [34], as described in Section 4.1.1. The magnitudes of these vectors are accumulated in a buffer and then thresholded. The rectangular region enclosing the pixels survived after thresholding gives the face frame. Once face is detected, then in this region we search for the lip, assuming that the lip constitutes the largest portion of the face that dominates the overall movement. A second thresholding of optical flow vector magnitudes in the detected face region, followed by morphological processing to fill up small holes and eliminating small isolated regions clears out the most moving parts, possibly the lip area. Around the center of gravity of these partial lip regions, we construct a fixed size window frame that we label as the lip region. The average error is observed as 11 pixel among our current database. In Fig. 4.1, we demonstrate the performance of our detection method on a video sequence from our current database. The stages of lip and face detection process is shown in Fig. 4.2

The described lip detection technique relies only on the motion information which is much more reliable as compared to texture which may show large discrepancies in terms of color, brightness and shape from one person to another. Thus, the resulting accuracy of localization is not as high as other detection techniques incorporating also texture information. But in turn, the technique seems quite general and robust.

4.3 Extraction of Lip Features

In this section, we will consider the so-called “eigenlip” representation as a visual representation methods to characterize lip movement.

4.3.1 Eigenlips

An efficient alternative to optical flow based representation technique is the eigenlip technique [32]. Eigenlips, just like eigenfaces, are appearance-based or pixel-based features that can be used to characterize the appearance of the lip of a speaking individual. Obtaining principal components of a lip image, i.e. eigenlips, can be thought of as the eigenvalue problem that we have briefly described in Section 3.2.1. Each lip image extracted from the video signal is represented by a set of eigenlip coefficients. These eigenlip coefficients, when



Figure 4.1: Lip and face detection performance

computed for every frame of the lip sequence, constitute the feature vector that can be used in place of the lip feature vector.

As opposed to geometry based features, each eigenlip feature vector represents only the appearance, i.e. the lip texture; the motion information is only inherent in the eigenlip sequence that should further be exploited by modelling the temporal relations between eigenlips of successive lip frames.

The advantage of the eigenlip approach is that it works simply on intensity values. This improves the robustness and the computational efficiency of the overall scheme as compared to techniques that require more sophisticated methods such as lip tracking for extraction of some geometric features, e.g. lip contours as in [11]. With the eigenlip approach, it suffices to employ a simple lip detection process for extracting lip frames from face images. The disadvantage of this approach is that it is generally sensitive to translation, rotation and lighting conditions, though small rigid motions of the head and small changes in illumination

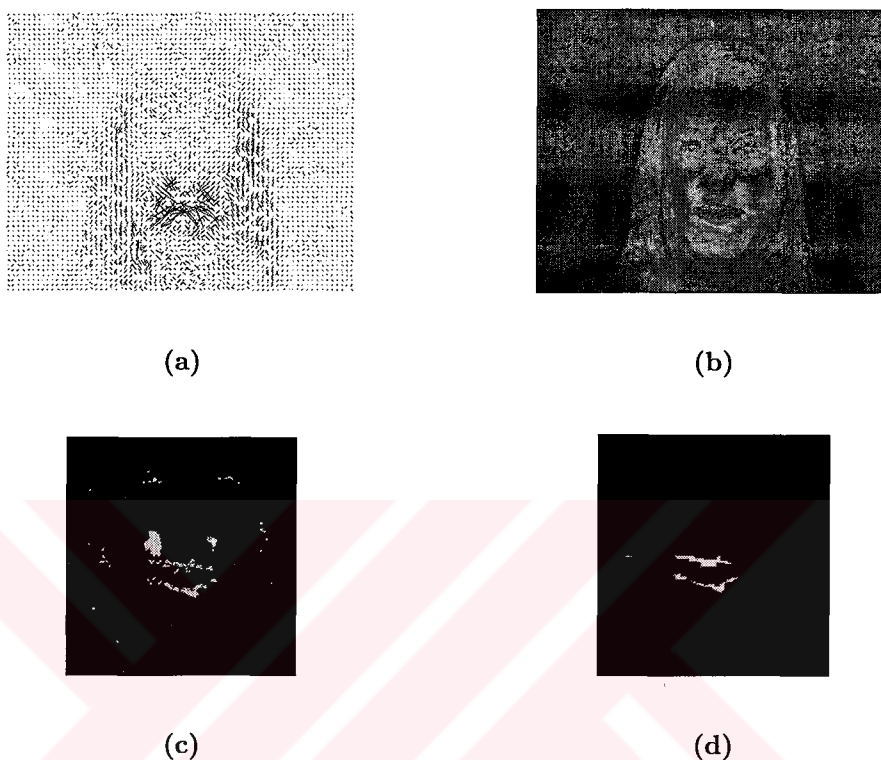


Figure 4.2: (a) Optical flow vectors (b) Accumulated vector magnitudes (brighter regions correspond to fast-moving parts) (c) Thresholded image after vector accumulation (d) Elimination of isolated small regions and filling up small holes by applying morphological operations

can be tolerated up to a certain measure.

4.4 Feature Fusion by Interpolation and Concatenation

Recalling that speech and lip movement are highly correlated, the proposed audio-motion fusion scheme is based on the early integration model where the integration is performed in the feature space to form a composite feature vector of audio and lip motion features. Classification is implemented by using these composite vectors. The audio features f_a and the motion features f_m are combined to form the joint audio-motion features. That will better exploit the temporal correlation of audio-video streams for robust performance.

As the audio features are extracted at a rate of 100 fps and the lip motion features are

extracted at a rate of 15 fps, a rate synchronization should be performed prior to the data fusion. Let the audio and the visual motion features be represented at time instants $k\frac{1}{100}$ and $i\frac{1}{15}$ seconds, respectively, i.e.,

$$\mathbf{f}_a^k = \mathbf{f}_a(k\frac{1}{100}) \quad \text{for } k = 0, 1, 2, \dots \quad (4.5)$$

$$\mathbf{f}_m^i = \mathbf{f}_m(i\frac{1}{15}) \quad \text{for } i = 0, 1, 2, \dots \quad (4.6)$$

The visual motion features can be computed using linear interpolation over the \mathbf{f}_m^i sequence to match the 100 fps rate,

$$\tilde{\mathbf{f}}_m^k = \tilde{\mathbf{f}}_m(k\frac{1}{100}) = (1 - \alpha_k)\mathbf{f}_m^{i^*} + \alpha_k\mathbf{f}_m^{i^*+1}, \quad (4.7)$$

where $i^* = \lfloor \frac{3k}{20} \rfloor$ and $\alpha_k = \frac{3k}{20} - i^*$. Hence the joint audio-motion feature \mathbf{f}_{am}^k is formed by combining the MFCCs, the first and second delta MFCCs and the interpolated lip motion features $\tilde{\mathbf{f}}_m^k$ for the k -th audio-visual frame:

$$\mathbf{f}_{am}^k = [\mathbf{f}_a^k \quad \tilde{\mathbf{f}}_m^k]. \quad (4.8)$$

4.5 Multimodal Speaker Identification System

As observed from Fig.4.3, the proposed overall scheme consists of two independent identification tasks: One performed with audio-motion information fused in feature space and the other with face-only texture features. Assuming that face texture is uncorrelated with speech and lip movement, the two individual decisions obtained in this way are combined by late integration. The fusion of audio and motion features is basically a data fusion process and the joint feature vector \mathbf{f}_{am}^k , for every audio-visual frame k , is the concatenation of the audio features \mathbf{f}_a^k and the interpolated motion features $\tilde{\mathbf{f}}_m^k$ as given in Eq. 4.8. The audio features are MFCCs and the motion features are eigenlip coefficients, as explained in Chapter 4. The HMM-based classifier described in Section 3 is first trained with these concatenated feature vectors extracted from the training dataset for each speaker class λ_n . For identification of an unknown speaker, the HMM-based classifier results in N likelihood ratios $\{\rho_{am}(\lambda_n)\}$, where λ_n denotes the n th HMM speaker class model and N is the number of the population. These ratios can then be combined with the N likelihood ratios $\{\rho_t(\lambda_n)\}$ provided separately by the face identification process presented in Chapter 3.

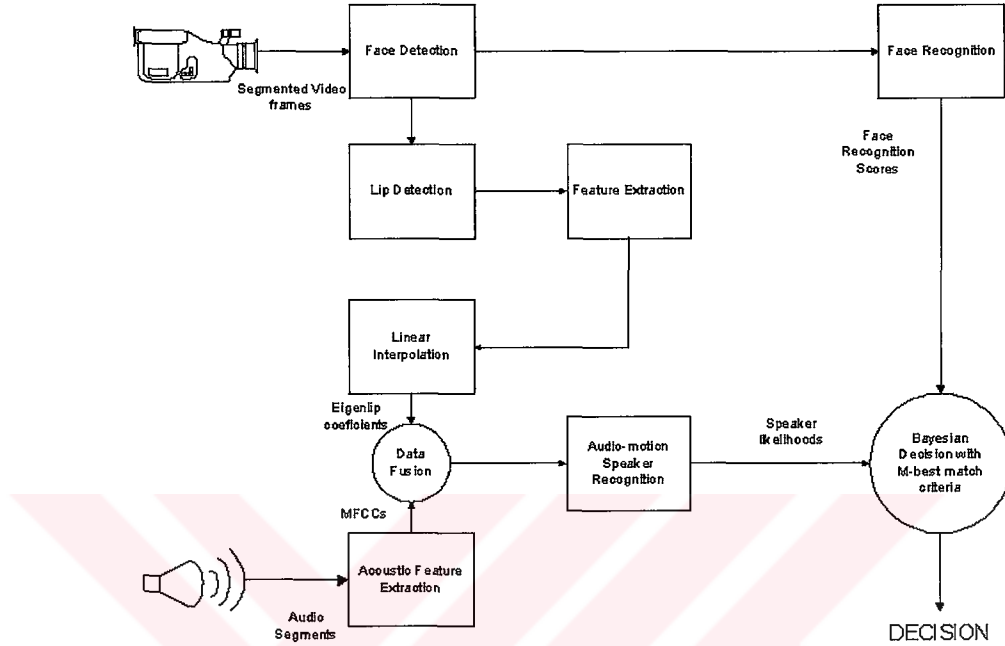


Figure 4.3: Multimodal speaker identification scheme.

4.5.1 Normalization

One issue to be addressed before proceeding with details of the overall system is normalization of the likelihood ratios resulting from different modalities. This is necessary to be able to use Equation 2.2 for decision fusion of audio-motion and texture likelihood scores, that we denote ρ_{am} and ρ_t respectively. Each of these scores covers a different range and thus they have to be normalized so as to be in the same scale. Recall from Chapter 3.2.2 that the value of ρ_t ranges in the interval $[0, \tilde{d}_{\max}/\sigma]$ and the choice of σ in the Gibbs distribution function was arbitrary (see Eq 3.22). Similarly, recalling Equation 2.4, ρ_{am} takes values in the interval $[0, \rho_{\max}]$ where ρ_{\max} and \tilde{d}_{\max} are to be determined experimentally by using the training data. Thus for normalization it is sufficient to choose σ as

$$\sigma = \frac{\tilde{d}_{\max}}{\rho_{\max}} \quad (4.9)$$

4.5.2 Multimodal Fusion

With the normalized likelihood ratios $\{\rho_{am}(\lambda_n)\}_{n=1}^N$ and $\{\rho_t(\lambda_n)\}_{n=1}^N$ in hand, we can now compute the likelihood ratios $\{\rho_{amt}(\lambda_n)\}_{n=1}^N$ of audio-motion-texture fusion by weighted summation as in Eq. 2.11. For each speaker class λ_n we obtain the following log likelihood ratio:

$$\rho_{amt}(\lambda_n) = \omega_{am}\rho_{am}(\lambda_n) + \omega_t\rho_t(\lambda_n) \quad (4.10)$$

where the weights ω_{am} and ω_t determine the contribution of each modality. Then the unknown speaker can be assigned to the class $\hat{\lambda}$ with the highest likelihood score:

$$\hat{\lambda} = \arg \max_{\lambda_n} \rho_{amt}(\lambda_n) \quad (4.11)$$

and recalling Equation 2.5 the corresponding accept-reject strategy is as follows:

$$\begin{array}{ll} \text{if } \rho(\hat{\lambda}) \geq \tau & \text{accept} \\ \text{otherwise} & \text{reject} \end{array} \quad (4.12)$$

The weights ω_{am} and ω_t in (4.10) should each normally reflect the reliability of the corresponding likelihood score. As stated before, there are different, adaptive or nonadaptive, ways of determining these weights. The most straightforward and efficient way of doing this comes with the assumption that reliability and discrimination capability are closely related issues since the feature sets are selected accordingly and if the feature set corresponding to a modality fails to discriminate and classify a given speaker among different classes, one may conclude that there is something wrong with the data itself and thus that it is not reliable. An easy way of measuring how much a given feature set is discriminative is to measure the difference between the two highest likelihood scores:

$$\begin{aligned} \omega_{am} &= |\rho_{am}(\lambda_i) - \rho_{am}(\lambda_j)| \\ \omega_t &= |\rho_t(\lambda_r) - \rho_t(\lambda_s)| \end{aligned}$$

where (λ_i, λ_j) and (λ_r, λ_s) are the best two matches for audio-motion modality and face modalities, respectively. Note that $\omega_{am} + \omega_t$ is not necessarily equal to 1. When the difference between the two highest likelihood scores for both modalities comes out to be very small, the weights ω_{am} and ω_t , and thus the likelihood score becomes also very small and the unknown speaker is rejected regardless of the individual likelihood ratios.

In multimodal speaker identification, M -best likelihood mechanism is used to increase performance. The details of decision fusion using M -best likelihoods is given in Section 2.2.2.

4.6 Discussion

In our HMM-based speaker identification system, joint use of the lip sequence and the audio signal of a speaking individual with early integration of audio and visual features is the key point that considers the correlation between lip motion and speech. We have considered the eigenlip technique which is an appearance-based approach, to implement visual feature representation.

The appearance-based eigenlip technique is an effective and computationally efficient method; however it characterizes the lip texture rather than the motion. As our multimodal system integrates the face texture as a separate modality, using lip texture here seems in fact redundant and the lip movement is indirectly taken into account during the HMM-based classification phase. Moreover, eigenlips are very sensitive to lighting conditions and pose (rotation, translation and scale).

The information in a video signal can be decomposed into three source of information. The first one is the audio signal which is extracted from the most natural act of speaking. The second is the texture information extracted from face or a portion of face. And the third one is the motion characteristics of the speech. All these sources involve spatial and temporal characteristics that can be exploited for any identification or recognition task. For instance, the temporal changes around the lip area of a speaking individual can be represented in terms of intensity itself or with some geometric sources of information such as optical flow motion vectors. Image intensities may be used as a correlated source of information with audio data if it is thought as a temporal sequence of images during a speech or an uncorrelated source if some of frames are used independently as done in our face-only scheme.

The advantage of using audio and visual sources jointly is the flexibility it provides in case a problem occurs in one of these sources. The fused system must still perform well if the audio or video of an individual is missing or very noisy. In our final proposed system, the feature concatenation of eigenlips and audio data is combined with texture-based face recognition at decision-fusion level. Although this scheme performs well with our database,

a separate audio-only identification task could be added to the system through decision fusion as well; in case the video is missing or noisy, the audio-only identification scores alone may remain still reliable. Such a system may increase the overall performance at the cost of system complexity.



Chapter 5

EVALUATION OF MULTIMODAL SPEAKER IDENTIFICATION SYSTEMS

In this chapter after presenting a brief introduction to the database and to the test environment, we will evaluate the performances of the multimodal speaker identification systems. Considering unimodal, bimodal and multimodal systems, the EER and ROC characteristics will be presented at varying levels of acoustic noise conditions.

5.1 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 in MVGL can be seen in Fig 5.1. The speaker identification database (MVGL-SID) includes 50 subjects where 8 of them are females. Each subject in the database utters 10 repetitions of her/his name and the fixed six-digit number. A set of impostor data is also collected with each subject uttering five different names from the population. The training and testing are performed over two independent data sets. A view of the variation in our database is presented in Fig 5.2.

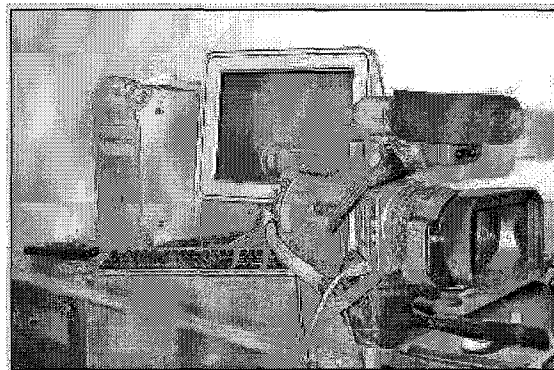


Figure 5.1: Data acquisition system in Koç University.



Figure 5.2: Sample subjects from the MVGL-SID database.

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 3.0, where each speaker is represented by a 6-state left-to-right HMM structure. The acquired video data is first split into segments of secret phrase utterances. The audio and visual streams are then separated into two parallel streams, where the visual stream has gray-level video frames of size 720×576 pixels containing the frontal view of a speaker's head at a rate of 15 fps and the audio stream has 16 kHz sampling rate. The acoustic noise, which is added to the speech signal to observe the identification performance under adverse conditions, is picked to be a mixture of office and babble noise. The audio stream is processed over 10 msec frames centered on 25 msec Hamming window. The MFCC feature vector, \mathbf{c}_k , is formed from 13 cepstral coefficients including the 0th gain coefficient using 26 mel frequency bins. The resulting audio feature vector, \mathbf{f}_a^k of size 39, includes the MFCC vector along with the first and the second delta MFCC vectors.

Each video stream is at most 1 second in duration and results in 15 individual face

and lip frames of sizes 370×460 and 120×128 , respectively. The motion feature vectors \mathbf{f}_m^i , which are used in both training and testing of the HMM-based classifier, are obtained as described in Chapter 4 with $p = 20$. As for the extraction of face feature vectors, an eigenspace of dimension $r = 20$ is computed using 5 pictures from each video sequence of the training set.

5.2 Performance of the Bimodal Bayesian Decision Fusion

In the bimodal Bayesian decision fusion algorithm, which is presented in Section 2.2.4, two sources of information are used. The more reliable information source \mathbf{f}_1 is taken to be the audio stream, which is known to out-perform under noise-free environments but its performance degrades rapidly under noisy conditions. The second source of information \mathbf{f}_2 is taken to be the eigenlip based audio-visual stream fused with the face texture.

The flow of the bimodal Bayesian decision fusion system is given in Fig 5.3. The proposed scheme consists of two independent identification tasks performed with audio-only and audio-motion-texture features. For the final decision a Bayesian classifier is incorporated to combine the two decisions obtained in this way. The likelihood ratios of audio-motion data fusion process, ρ_{am} , and the face identification task, ρ_t are described in Chapter 4. The likelihood score of audio-motion-texture fusion, ρ_ℓ , is obtained by the weighted average of the two individual likelihood ratios,

$$\rho_\ell = G\rho_{am} + (1 - G)\rho_t \quad (5.1)$$

where the weight G , $0 \leq G \leq 1$ determines the contribution of each modality. Note that for $G = 0$, the second source of information turns out to be only face texture and similarly for $G = 1$, it turns out to be only audio-motion stream.

The identification results are shown in Table 5.1, where we observe the equal error rates at varying levels of acoustic noise for name scenario. In the training phase 5 repetitions of each name utterance are used and in the testing phase each subject utters 5 repetitions of her/his name and they also utter 5 different names to cover the imposter data. The first two rows display the equal error rates obtained for audio-only and audio fused with lip motion (audio-motion). The third row presents the equal error rate for the face texture only identification system that is based on the eigenface method. Finally, the last five rows

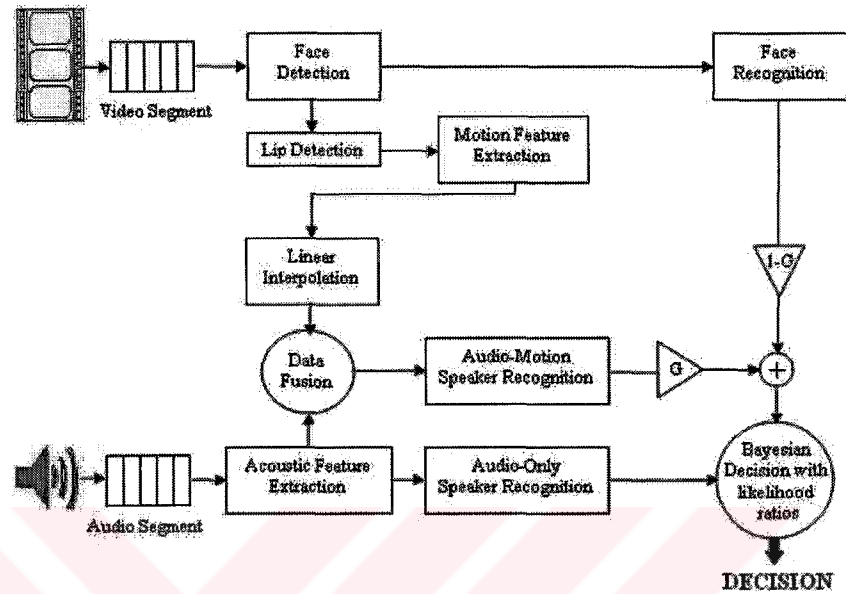


Figure 5.3: Bimodal Bayesian decision system.

display the equal error rates obtained after the Bayesian decision fusion of the audio-only and the audio-motion-texture identification results, at varying values of G . The best equal error rate results are obtained when G is 0.75, that is when audio-motion and texture-only schemes have 75% and 25% contributions to the decision fusion of likelihood ratios, respectively.

In the audio-only case the identification performance degrades rapidly with decreasing SNR. However, when lip motion is fused with audio, the identification performance improves at the low SNR values due to the correlation existing between lip movement and speech. But some performance degradation is observed at high SNR values. This is mostly due to the uncertainty introduced by the lip motion modality. The Bayesian decision fusion is introduced to overcome this performance degradation at high SNR levels. The overall performance is improved significantly using the Bayesian decision fusion at all SNR levels. Thus this system seems less sensitive to noise level and the incorporation of the Bayesian classifier guarantees the overall performance to remain at least as good as the audio-only performance [36].

EER (%)							
Source Modality	Noise Level (dB SNR)						
	clean	25	20	15	10	5	0
Audio	2.3	3.0	4.3	7.0	10.4	19.0	23.6
Audio-Lip	8.4	11.4	12.2	13.0	13.9	14.5	14.7
Eigenface	4.1						
Bayesian Decision Fusion							
G = 1.00	1.8	2.5	4.2	6.8	8.9	9.9	11.2
G = 0.75	1.5	2.6	3.9	6.5	8.7	9.3	9.8
G = 0.50	1.7	2.8	4.0	6.7	9.0	13.1	13.2
G = 0.25	1.7	2.8	4.2	6.8	9.7	14.1	14.1
G = 0.00	1.8	2.9	4.2	7.0	10.2	14.5	14.6

Table 5.1: Speaker identification results of Bayesian Decision Fusion scenario.

5.3 Performance of the Multimodal Speaker Identification System

The proposed multimodal identification system includes audio, lip-motion and face texture modalities, where audio and lip-motion are considered for data fusion and all possible likelihood streams are considered for decision fusion. The M -best scores contribute to the final multimodal decision fusion with weighted confidence levels as described in Chapter 4.

Performance evaluations are done over two scenarios; each subject either utters her/his name or fixed 6-digit number (348572). We have used 5 repetitions for training, 10 repetitions for testing in which 5 of them are collected as impostor test data for the name scenario. For the digit scenario 4 repetitions are used as training data and 6 repetitions are used as test data. Note that in digit scenario all the utterances not belonging to the subject are used as impostor data.

For the proposed multimodal system we have used different combinations of modalities to find out an optimal fusion strategy. While presenting the performance figures, the abbreviations in Table 5.2 are used to easily follow the various fusion strategies.

The performance of the multimodal identification system is presented in terms of equal error rates and some selected ROC curves at varying levels of acoustic noise. The unimodal (video-only and audio-only), bimodal (audio-visual data fusion) and multimodal equal error

<i>Symbol</i>	<i>Description</i>
A	Audio-only scenario
F	Face-only scenario (Eigenface)
L	Eigenlip scenario (automatic lip region detection)
L _H	Eigenlip scenario (hand labelled)
+	Multimodal M-best decision fusion
⊕	Data fusion at feature level

Table 5.2: Modality abbreviations for multimodal scenarios.

rates are displayed on the same table to better observe the improvement obtained by multimodal identification system. For the purpose of checking the lip detection performance we run our system for lip frames extracted either by hand-labelling or by applying our optical flow based detection algorithm. The results of name and digit scenarios are given in Table 5.3 and Table 5.4, respectively.

In Table 5.3 and Table 5.4, the first four rows display the equal error rates obtained for unimodal scenarios (audio-only and video-only). Next two rows display the equal error rates obtained for bimodal scenarios (audio fused with lip motion). Finally the last nine row presents the equal error rates for the scenarios designed according to the proposed multimodal system which is based on M -best match criteria.

In the audio-only case the identification performance degrades rapidly with decreasing SNR. However, by fusing eigenlip features with MFCCs, the identification performance improves significantly at the low SNR values, due to the correlation existing between lip movement and speech. But for high SNR levels, an improvement is not observed after fusing visual features. In such cases audio-only scenario performs better than the fused systems. The ROCs of audio-only and audio-visual scenarios for varying acoustic noise levels are given in Figures 5.5 and 5.6.

The overall performance is further improved at all noise levels using the multimodal Bayesian decision fusion which is introduced in Chapter 4. Thus the multimodal system seems less sensitive to noise level and the incorporation of the M -best likelihoods significantly improves the overall system for both scenarios. This significant improvement is the result of contributing the scores of different identification tasks where each modality carries useful

EER (%)							
Source Modality	Noise Level (dB SNR)						
	clean	25	20	15	10	5	0
Unimodal							
A	2.3	3.0	4.3	7.0	10.4	19.0	23.6
F	6.5						
L	10.4						
L _H	7.2						
Bimodal Data Fusion							
L ⊕ A	8.4	11.4	12.2	13.0	13.9	14.5	14.7
L _H ⊕ A	4.9	5.8	6.0	6.8	8.2	9.4	10.0
Multimodal Fusion							
L ⊕ A + F	7.0	11.0	12.0	12.8	12.9	13.2	13.8
A + L ⊕ A + F	2.7	3.1	4.5	5.8	7.7	10.5	11.8
A + L + L ⊕ A + F	2.5	3.0	4.1	5.2	7.3	9.1	9.7
A + L + F	2.7	3.6	5.3	6.5	10.4	18.6	19.2
A + F	2.5	3.1	4.2	6.3	9.5	18.6	23.0
A + L + L ⊕ A	2.8	3.2	5.4	6.0	9.5	11.3	12.2
A + L ⊕ A	2.9	3.2	5.3	6.1	9.7	12.0	12.6
A + L	4.1	5.6	7.2	9.2	12.6	20.9	21.1
L + F	8.8						

Table 5.3: Speaker identification results of proposed multimodal system for the name scenario.

information about the speaking individual.

In Figure 5.7 and 5.8, the ROC curves for top running fusion combinations are presented in loglog scale for a better comparison with the audio-only ROC curve. The decision fusion between the audio-motion and the face texture (L ⊕ A + F) out-performs the decision fusion of audio, lip-motion and face texture (L + A + F). Hence this observation supports the fact that the correlation between audio and lip-motion helps to better discriminate speakers. Although the decision fusion between audio, lip-motion, audio-motion and face texture (A + L + L ⊕ A + F) performs slightly better than the candidate system L ⊕ A + F, this performance difference is not that significant.

EER (%)							
Source Modality	Noise Level (dB SNR)						
	clean	25	20	15	10	5	0
Unimodal							
A	3.3	3.8	5.4	8.4	12.6	21.3	25.9
F	8.2						
L	15.4						
L _H	13.8						
Bimodal Data Fusion							
L \uplus A	10.1	12.8	13.9	14.8	15.9	17.2	17.8
L _H \uplus A	7.7	8.0	8.3	9.1	11.0	12.4	13.1
Multimodal Fusion							
L \uplus A+F	8.3	11.7	13.3	14.1	14.9	16.8	17.2
A+L \uplus A+F	3.3	4.0	5.1	6.4	8.9	10.5	12.1
A+L+L \uplus A+F	3.4	3.9	4.7	5.9	8.4	9.5	10.5
A+L+F	3.4	3.9	5.3	7.9	12.4	18.3	20.9
A+F	3.6	4.1	5.0	6.3	11.5	18.5	23.5
A+L+L \uplus A	3.5	4.1	5.3	6.6	9.6	11.3	12.9
A+L \uplus A	3.5	4.2	5.5	7.2	10.0	11.7	13.5
A+L	4.4	4.9	7.7	9.8	14.6	21.5	23.2
L+F	13.6						

Table 5.4: Speaker identification results of proposed multimodal system for the digit scenario.

The overall performance of the name scenario is better than the fixed 6-digit number scenario in which identification task is expected to be harder since all impostor speakers utter the same 6-digit number. This is mainly due to the different sensitivities of these two scenarios to the true client since in the name scenario case the HMM structure models not only the personal biometric voice and lip movements but also the voice and lip movements corresponding to the speech content.

The lip detection performance can be evaluated from the visual-only EERs. In our experiments we have used optical flow based face and lip detection method. In order to check the detection performance we also run our unimodal and bimodal system for lip

regions which are extracted by hand-labelling. The detection performance can be observed in Figure 5.4.

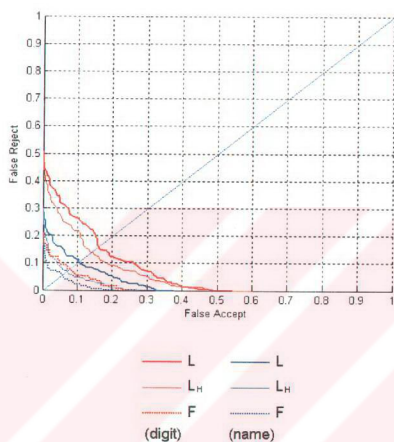
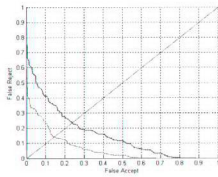
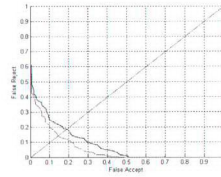


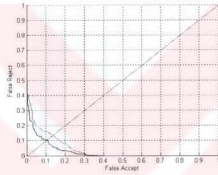
Figure 5.4: Receiving operating curves for visual-only scenarios



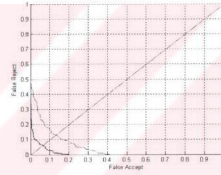
(0 dB)



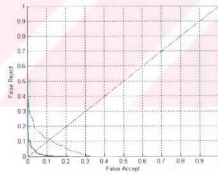
(5 dB)



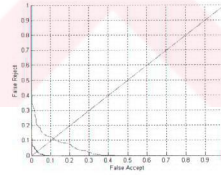
(10 dB)



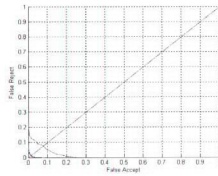
(15 dB)



(20 dB)



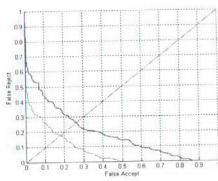
(25 dB)



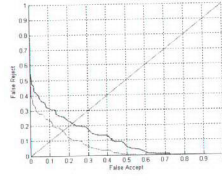
(clean)

— A
 - - LWA

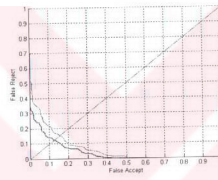
Figure 5.5: Receiving operating curves for bimodal decision Fusion at various acoustic noise levels (Name scenario)



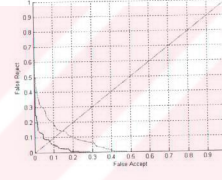
(0 dB)



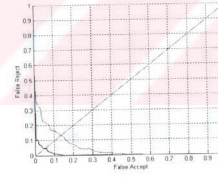
(5 dB)



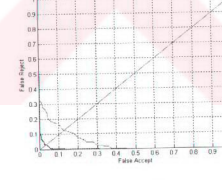
(10 dB)



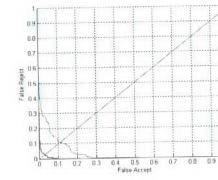
(15 dB)



(20 dB)



(25 dB)



(clean)

— A
— LWA

Figure 5.6: Receiving operating curves for bimodal decision Fusion at various acoustic noise levels (Digit scenario)

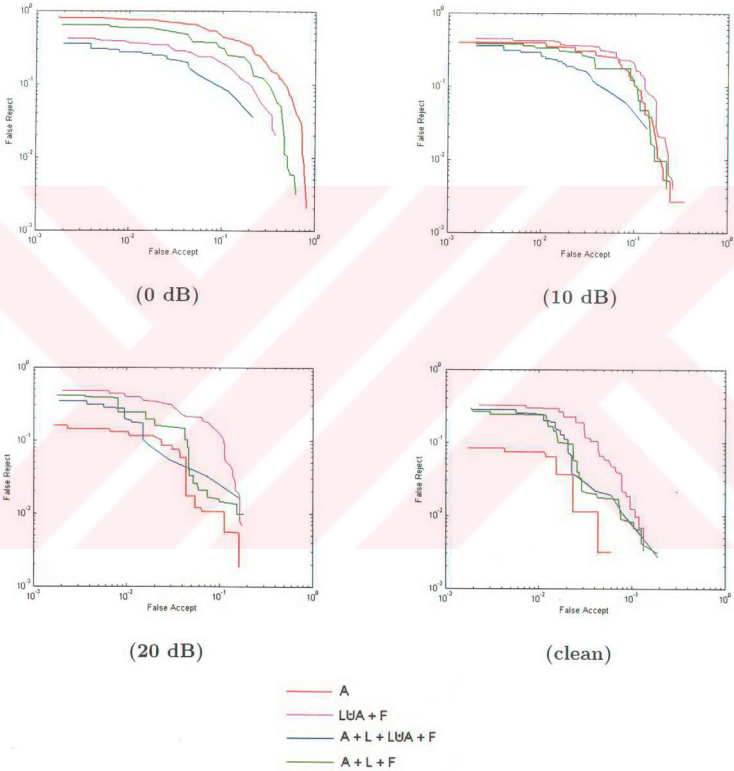


Figure 5.7: Receiving operating curves for the proposed multimodal system at various acoustic noise levels (Name scenario)

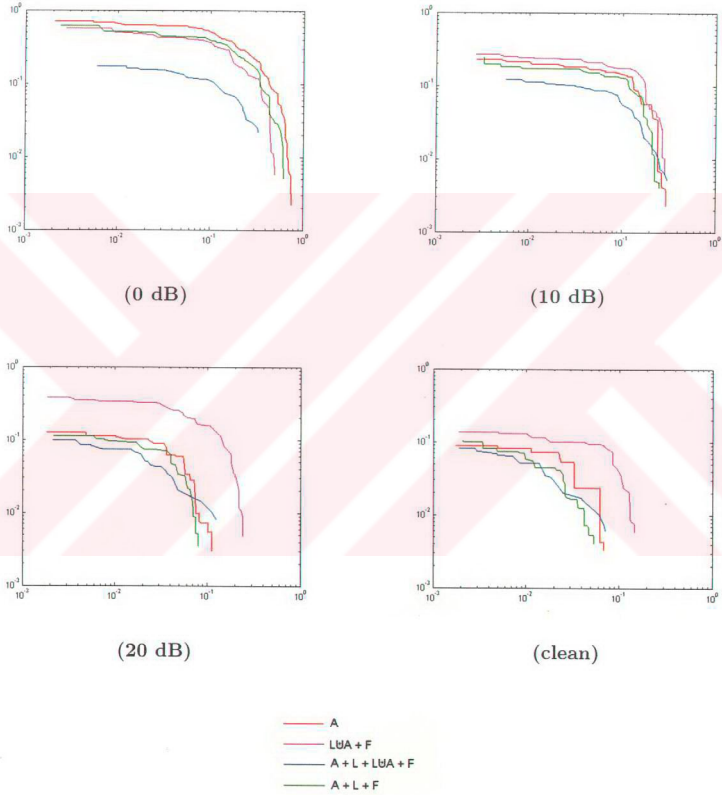


Figure 5.8: Receiving operating curves for the proposed multimodal system at various acoustic noise levels (Digit scenario)

Chapter 6

CONCLUSIONS

Biometric person identification technologies focus on voice, face, iris and retina scans, signature strokes, fingerprint and gait as distinguishing source of personal information. All suggested conventional techniques for unimodal scenarios which are based on these biometrics do not perform well enough to obtain a robust identification system. The promising attempt to build more reliable and robust identification systems appears to be the fusion of individual modalities by means of data or decision fusion. It is clear that there exists a correlation between voice and lip of a speaking individual. The audio information of a speech and the temporal and visual characterization of lip constitute more information about the speaking person. This correlation could be represented by fusing the features of both modalities. On the other hand the visual data of a person may not be correlated with audio data if only a face recognition scheme is taken into consideration. Such modalities do not carry motion information but appearance-based characteristics. In order to integrate the decision resulting from any uncorrelated data source, decision fusion algorithms are used. Decision fusion algorithms do not care about the feature-level fusion but the likelihoods or scores obtained from different identification systems. In recent developments both data and decision fusion mechanisms are used to improve the performance of audio-visual speaker identification systems as we applied a similar strategy in our proposed multimodal system.

Our proposed multimodal system integrates three sources of information to improve the identification performance over unimodal schemes. The data fusion of audio and lip motion information has availed us the possibility of fully exploiting the correlations existing between two modalities. Since the reliability of each individual source of information (audio, lip, face) may vary under different light and acoustic conditions, our multimodal decision fusion strategy which uses audio, motion and texture characteristics of a speaking individual, significantly improves the overall performance. On the other hand the use of confidence

measure which is applied beneficially in the Bayesian decision fusion is another advantage of our system since we have a priori information on different modality streams. The fusion of the decisions with the highest matches (highest likelihood scores) resulting from individual classifiers and weighting these decisions with their confidence levels make the proposed speaker identification system more reliable. We have considered 3-best match scores and checked the difference between these scores in order to observe the reliability of the modality. In our proposed system we have used eigenlip coefficients as the motion features. Since eigenlip coefficients are pixel-based features, the temporal changes around the lip area can be easily represented. Such a representation avoids inevitable robustness problems of the systems relying rather on geometric features that require sophisticated and mostly unreliable image analysis tasks, such as segmentation and lip tracking. The face texture information decoupled from the video stream is also incorporated into the decision fusion mechanism to further improve the performance. The disadvantage of using eigenlips and eigenfaces is their highly dependency on the different light and pose conditions as well as image quality.

In this thesis we give the theoretical framework of fusion techniques which are widely used in speaker identification systems. We have considered the formulation of reject-accept mechanism and background theory for unimodal, bimodal data and decision fusion methods. We give a brief explanation of Bayesian decision fusion approach. We also discuss a more robust decision fusion algorithm which is based on the fusion of M -best likelihoods. Using the M -best likelihoods resulting from each modality is far more computationally effective than applying regular additive decision fusion, especially when the number of subjects in the database is large.

We discuss the theory of Hidden Markov Models which is used in text-dependent speaker identification systems. We briefly explain how HMM models the temporal characterization of an audio or even an audio-video stream. We also present MFCCs as audio features.

Later, we focus on the widely used Principal Component Analysis method which extracts the visual information of a speaking individual. We present the eigenface method which is then adapted to lip images and named as eigenlip approach. We also propose a method of speaker identification from the face sequences in an audio-visual system.

In Chapter 4, we first give the theoretical background of optical flow method. Optical flow vectors are used for the purpose of motion detection which is observed mostly in face and

lip regions of a speaking individual. After the extraction of lip area, the eigenlip coefficients are calculated as the visual feature vectors. The resulting eigenlip-based visual features are then interpolated and fused by audio features by concatenating these two sources of information.

There are further issues to be addressed. First, 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. In this respect, methodologies that would enforce the overall scheme for better invariance to such properties has to be explored. Secondly, the decision fusion mechanism can be improved, noting that there are many other ways of combining information coming separately from audio, motion and texture parts of the video sequence of a speaking person. All these issues should be further investigated.

BIBLIOGRAPHY

- [1] N.K. Ratha, A. Senior, and R.M. Bolle, "Automated biometrics," *ICAPR*, pp. 445–474, May 1997.
- [2] J.P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, September 1997.
- [3] D. D. Zhang, *Automated Biometrics*, Kluwer Academic Publishers, 2000.
- [4] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 586–591, September 1991.
- [5] L. Rabiner and B-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.
- [6] W. Zhao, R. Chellappa, A. Rosenfeld, and P. Phillips, "Face recognition: A literature survey," *UMD CfAR Technical Report*, pp. CAR-TR-948, 2000.
- [7] 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.
- [8] C. Sanderson, S. Bengio, H. Bourlard, J. Mariethoz, R. Collobert, M.F. BenZeghiba, F. Cardinaux, and S. Marcel, "Speech and face based biometric authentication at idiap," *Proc. of the Int. Conf. on Multimedia & Expo 2003 (ICME2003)*, vol. 3, pp. 1–4, July 2003.
- [9] 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 (ICME2003)*, vol. 3, pp. 9–12, July 2003.

- [10] 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.
- [11] 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.
- [12] T. Wark, S. Sridharan, and V. Chandran, "The use of temporal speech and lip information for multi-modal speaker identification via multi-stream HMM's," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 2000 (ICASSP 2000)*, pp. 2389–2392, 2000.
- [13] A. Kanak, E. Erzin, Y. Yemez, and A. M. Tekalp, "Joint audio-video processing for biometric speaker identification," *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing 2003 (ICASSP 2003)*, vol. II, pp. 377–380, 2003.
- [14] 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*, September 1986.
- [15] 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.
- [16] J. Kittler, Y. P. Li, J. Matas, and M. U. Ramos Sanchez, "Lip-shape dependent face verification," *First International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA)*, pp. 61–68, March 1997.
- [17] J. Luettin, N. A. Thacker, and S. W. Beet, "Speaker identification by lipreading," *Proc. Int. Conf. Spoken Language Processing*, p. 6265, October 1996.
- [18] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner, "Acoustic-labial speaker verification," *Pattern Recognition Letters*, vol. 18, no. 9, pp. 853–858, 1997.

- [19] 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.
- [20] B.V. Dasarathy, *Decision Fusion*, IEEE Computer Society Press, 1994.
- [21] P.A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*, Englewood Cliffs NJ, Prentice-Hall Inc., 1982.
- [22] 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.
- [23] L. Wu, S. L. Oviatt, and P. R. Cohen, "Multimodal integration - A statistical model," *IEEE Transactions on Multimedia*, vol. 1, no. 4, pp. 334-341, December 1999.
- [24] H. Altincay and M. Demirekler, "An information theoretic framework for weight estimation in the combination of probabilistic classifiers for speaker identification," *Journal of Speech Communication*, vol. 30, pp. 255-272, 2000.
- [25] 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.
- [26] J. Luettin and S Dupont, "Continuous audio-visual speech recognition," *Technical Report IDIAP*, 1997.
- [27] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK-hidden markov model toolkit v2.1," *Entropic Research, Cambridge*, 1997.
- [28] G.D. Forney, "The viterbi algorithm," *Proc.IEEE*, pp. 61:268-278, March 1973.
- [29] I. T. Jolliffe, "Principal component analysis," *SpringerVerlag*, p. New York, 1986.
- [30] S. Geman and D. Geman, "Stochastic relaxation, gibbs distribution, and the bayesian restoration of images," *IEEE TPAMI*, pp. 6(6):721-741, 1984.

- [31] Hamid Bolouri Zhengjun Pan, Rod Adams, "Image recognition using discrete cosine transforms as dimensionality reduction," *IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing (NSIP01)*, June 2001.
- [32] C. Bregler and Y. Konig, "Eigenlips for robust speech recognition," *Proc. of IEEE Conf. on Acoustics, Speech and Signal Processing*, pp. 669–672, 1994.
- [33] E.Trucco and A.Verry, *Introductory Techniques for 3-D Computer Vision*, Prentice Hall, 1998.
- [34] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *Proc. of 7th International Joint Conference on Artificial Intelligence*, pp. 674–679, 1981.
- [35] X.Liu, T.Chen, and B.V.K. Vijaya Kumar, "Face authentication for multiple subjects using eigenflow," *The Journal of Pattern Recognition Society*, pp. 313–328, December 2001.
- [36] Y. Yemez, A. Kanak, E. Erzin, and A. M. Tekalp, "Multimodal speaker identification with audio-video processing," *Proc. of the Int. Conf. on Image Processing 2003 (ICIP 2003)*, pp. 14–17, September 2003.

VITA

ALPER KANAK was born in Bursa, Turkey on July 3,1978. He received his B.Sc. degree in Control and Computer Engineering from Istanbul Technical University,Istanbul, in 2001. He worked as a software engineer in his last year of undergraduate school for the war simulation project which was supported by Turkish Army. From August 2001 to August 2003, he worked as a teaching and research assistant in Koc University, Turkey and had studied for the "*Multi-Stage and Multi-Modal Signal Processing for Person Identification*" project which was sponsored by TUBITAK, since March 2002. He has published several papers about *Biometric Person Identification* for the following conferences ICASSP2003 (Hong Kong), ICME2003 (Baltimore), ICIP2003 (Barcelona, Spain), SIU2003(Istanbul, Turkey) and Workshop on DSP in Mobile and Vehicular Systems (Nagoya, Japan).