Evaluation and Analysis of Hybrid Intelligent Pattern Recognition Techniques for Speaker Identification

Noor Almaadeed

A thesis submitted in partial fulfilment
of the requirements for the degree of

Doctor of Philosophy

Brunel University

June 2014

Approved by _____

Chairperson of Supervisory Committee

_____

_____

_____

Program
Authorised

to Offer Degree _____

Date _____

I would like to dedicate this thesis to my loving parents and family
for their support during this long doctorate journey.

Acknowledgements

The author wishes to express sincere appreciation to Prof. Amar Aggoun for his assistance in the preparation of this manuscript. In addition, special thanks to Prof. Abbes Amira whose familiarity with the needs and ideas of the class was helpful during the early programming phase of this undertaking. Thanks also to the members of the council for their valuable input.

Author's Publications

- Conference Article (published)
  - **Noor Almaadeed**, Amar Aggoun, Abbes Amira: "Audio-Visual Feature Fusion for Speaker Identification," in *Proceedings of the 19th International Conference on Neural Information Processing (ICONIP): Part I*, 2012, Doha, Qatar, pp. 56-67.
- Journals Articles (under review)
  - **Noor Almaadeed**, Amar Aggoun, Abbes Amira: "Speaker Identification using Multimodal Neural Networks and Wavelet Analysis," in *IET Biometrics*.
  - **Noor Almaadeed**, Amar Aggoun, Abbes Amira: "Text-independent Speaker Identification using Vowel Formants," in *Journal of Signal Processing Systems*.
  - **Noor Almaadeed**, Amar Aggoun, Abbes Amira: "Evaluation and Analysis of Audio-Visual Speaker Identification," in *Pattern recognition*.
- Conference Article (under review)
  - **Noor Almaadeed**, Amar Aggoun, Abbes Amira: "Vowel Formants in Speaker Identification," in *Proceedings of the Qatar Foundation Annual Research Conference (QF-ARC)* , to be held on November 24-25, 2013, Doha, Qatar.

Abstract

The rapid momentum of the technology progress in the recent years has led to a tremendous rise in the use of biometric authentication systems. The objective of this research is to investigate the problem of identifying a speaker from its voice regardless of the content (i.e. text-independent), and to design efficient methods of combining face and voice in producing a robust authentication system.

A novel approach towards speaker identification is developed using wavelet analysis, and multiple neural networks including Probabilistic Neural Network (PNN), General Regressive Neural Network (GRNN) and Radial Basis Function-Neural Network (RBF NN) with the AND voting scheme. This approach is tested on GRID and VidTIMIT corpora and comprehensive test results have been validated with state-of-the-art approaches. The system was found to be competitive and it improved the recognition rate by 15% as compared to the classical Mel-frequency Cepstral Coefficients (MFCC), and reduced the recognition time by 40% compared to Back Propagation Neural Network (BPNN), Gaussian Mixture Models (GMM) and Principal Component Analysis (PCA).

Another novel approach using vowel formant analysis is implemented using Linear Discriminant Analysis (LDA). Vowel formant based speaker identification is best suitable for real-time implementation and requires only a few bytes of information to be stored for each speaker, making it both storage and time efficient. Tested on GRID and VidTIMIT, the proposed scheme was found to be 85.05% accurate when Linear Predictive Coding (LPC) is used to extract the vowel formants, which is much higher than the accuracy of BPNN and GMM. Since the proposed scheme does not require any training time other than creating a small database of vowel formants, it is faster as well. Further-

more, an increasing number of speakers makes it difficult for BPNN and GMM to sustain their accuracy, but the proposed score-based methodology stays almost linear.

Finally, a novel audio-visual fusion based identification system is implemented using GMM and MFCC for speaker identification and PCA for face recognition. The results of speaker identification and face recognition are fused at different levels, namely the feature, score and decision levels. Both the score-level and decision-level (with OR voting) fusions were shown to outperform the feature-level fusion in terms of accuracy and error resilience. The result is in line with the distinct nature of the two modalities which lose themselves when combined at the feature-level. The GRID and VidTIMIT test results validate that the proposed scheme is one of the best candidates for the fusion of face and voice due to its low computational time and high recognition accuracy.

Contents

Abbreviations

1D : One-Dimensional

2D : Two-Dimensional

3D : Three-Dimensional

AI : Artificial Intelligence

AMM : Active Appearance Model

ANN : Artificial Neural Network

BPA : Back Propagation Algorithm

BPNN : Back Propagation Neural Network

CB : (1) Critical Bandwidth; (2) Codebook

CLAFIC : Class-featuring Information Compression

CPU : Central Processing Unit

DCT : Discrete Cosine Transform

DE : Differential Evolution

DFT : Discrete Fourier Transform

DOG : Derivative of Gaussian

DTW : Dynamic Time Warping

DWPT : Discrete Wavelet Packet Transform

DWT : Discrete Wavelet Transform

EBGM : Elastic Bunch Graph Matching

EER : Equal Error Rate

EFA : Eigenface Approach

EGM : Elastic Graph Matching

EM : Expectation Maximisation

FA : False Acceptance

FAR : False Accept Rate

FB : Filter-Bank

FBG : Face Bunch Graph

FFT : Fast Fourier Transform

FKP : Finger-Knuckle-Print

FLD : Fisher Linear Discriminant

FPR : False Positive Rate

FR : False Rejection

FRR : False Reject Rate

GF : Gaussian-shaped Filter

GMM : Gaussian Mixture Models

GRID : Global Resource Information Database

GRNN : General Regression Neural Network

HHMM : Hierarchical Hidden Markov Models

HMM : Hidden Markov Models

ICA : Independent Component Analysis

IFFT : Inverse Fast Fourier Transform

IMFCC : Inverted Mel-Frequency Cepstral Coefficients

IR : Infrared

KDDA : Kernel Direct Discriminant Analysis

KFD : Kernel Fisherface Discriminants

KICA : Kernel Independent Component Analysis

KL : Karhunen-Loeve (as in KL transform)

KL : Kullback-Leibler (as in KL divergence)

KLDA : Kernel Linear Discriminant Analysis

KPCA : Kernel Principal Component Analysis

K-NN : k-Nearest Neighbour

LC 2D-HMM : Low-Complexity 2D HMM

LDA : Linear Discriminant Analysis

LPC : Linear Predictive Coding

LPCC : Linear Prediction Cepstral Coefficients

LVQ : Learning Vector Quantisation

MAP : Maximum A Posteriori

MFCC : Mel-Frequency Cepstral Coefficients

MF-PLP : Mel-Frequency Perceptual Linear Predictive

MLED : Maximum Likelihood Eigen Decomposition

MLP : Multi-Layer Perception

MNN : Multimodal Neural Network

MSE : Mean Square Error

M-PSOM : Multiple Parametric Self-Organising Maps

NFL : Nearest Feature Line

NIST : National Institute of Standards and Technology

NN : Neural Network

NNC : Neural Network Committee

P2D-HMM : Pseudo 2D HMM

PCA : Principal Component Analysis

PDF : Probability Density Function

PLP : Perceptual Linear Predictive

PNN : Probabilistic Neural Network

QMF : Quadrature Mirror Filters

RBF : Radial Basis Functions

RFID : Radio-Frequency Identification

ROC : Receiver Operating Characteristic

SBC : Sub-Band Coding

SIFT : Scale Invariant Feature Transform

SNR : Signal-to-Noise Ratio

SOM : Self-Organising Map

SSR : Signal-to-Signal Ratio

STT : Shape Trace Transform

SV : Speaker Verification

SVM : Support Vector Machines

TIMIT : Texas Instruments / Massachusetts Institute of Technology

TPR : True Positive Rate

UBM : Universal Background Model

VidTIMIT : Video TIMIT

VQ : Vector Quantisation

WP : Wavelet Packet

WPD : Wavelet Packet Decomposition

WPT : Wavelet Packet Transform

WSBC : Wavelet Sub-Band Coding

ZCR : Zero Crossing Rate

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The twentieth century has seen great advances in science and technology. The increasing demand for more reliable and convenient security systems has generated a renewed interest in human identification based on biometric identifiers such as face, fingerprints, iris, voice and gait. Speaker identification is a biometric classification task aimed at identifying a person from his or her voice [12]. Establishing human identity reliably and conveniently has become a major challenge for a modern-day society [13]. The taxonomy of speaker identification involves an in-depth knowledge of a set of diverse fields including machine learning, pattern recognition and signal processing.

Speaker identification is as old as the computer itself and has accumulated more than fifty years of progress with the assumption that a human voice is unique to each individual and therefore can be used as identification. During this period, a lot of common tasks for speaker identification were identified. Automated speech recognition has widely been studied around the world from both commercial and security perspectives. There is a large market for such systems that include the automation of certain services that would otherwise require an operator, and speech transcription services for multiple purposes.

Speaker identification systems have always suffered loss of accuracy compared to biometric systems based on physical biometrics such as fingerprints and retinal scan. Therefore, recent research has improved speaker identification systems

through multiple strategies of which multimodal biometrics is well known. Multimodal biometric systems are increasingly becoming popular since they usually demonstrate a superior performance over their unimodal counterparts [14]. But the choice of the modalities in identification systems are not influenced just by their error rate criteria, but also by the speed of operation. Usually the first modality (the one that is processed and checked first) is selected such that it works very fast. Then one or more subsequent modalities (which are slower but more accurate) are combined with its results. This type of combination, in most cases, has proven to be more effective and reliable.

Biometrics-based products are being used for security and authentication over the phone or locally on the doorstep to grant access to some vital information. Traditionally, these security and authentication systems are either based on Radio-Frequency Identification (RFID) tags or some mechanical or electronic lock systems [15] and perform authentication using physical biometrics such as fingerprints, iris, hand recognition and face recognition. Other behavioural biometrics-based systems include voice, signature and gait recognition.

Researchers over the past forty years have realised that single-biometric systems are prone to much greater risk of being deceived and tricked. One of the ways to overcome the problems with these systems is to combine multiple biometrics. From the early 1990s, with aggressive research in speaker identification for use in commercial products, researchers' attention has shifted between physical and behavioural, and then towards a combination of both at the feature or score-level to enhance the identification or authentication process [16].

Systems based on physical biometrics are more powerful and accurate than the ones based on behaviour. However, they are more rigid, invasive and may impose a burden on the user [16]. For example, iris scanning requires the person being scanned to be in a certain posture to help the scanner detect his eye. If he is not at the right angle, he might end up trying for a few minutes before the system grants him access. Furthermore, a fingerprint system may have hard time identifying a subject with fingers worn out due to a particular work environment. Behavioural approaches, on the other hand, although relatively less powerful, work seamlessly, and are more flexible compared to physical biometrics-based approaches.

The multimodal biometric approach aims to solve these problems by relying on more than a single biometric to identify the given user [16]. The robustness of this approach is established by combining the patterns' data from two biometric sources. The system becomes more reliable and effective: should the detection of any of the biometrics fail, the other is still available. These dual approaches started emerging in the 1990s and led to related products on the market.

Speaker identification is the process of finding out who a person is by comparing his or her blueprint (e.g. voice, face) to a list of registered users created during an enrollment stage (reading, recording and analysis) [17]. The system extracts unique features from each recording and saves it as a template. There is a difference between speaker identification (which is the identification of the person speaking) and speech recognition (which is the recognition of the content of the speech). In addition, there is a difference between the act of authentication (also known as speaker verification) and identification.

During the identification phase, the system tries to match the current template with those available in the database. The current template is assigned to a particular user if its score or distance makes it very close to that of a registered user. This phase is also called the testing phase. In a speaker verification system, when a subject claims to be a certain speaker, the system verifies the authenticity of this claim by matching the template generated with that of the target speaker stored in the database. If the distance of the current template is below a certain threshold, the current claimant is accepted. Otherwise, he is rejected [17]. The system selects the speaker based on the highest matching score or based on the shortest distance between the current template and the enrolled templates of various speakers.

In a multimodal biometric system, the risk of fooling the system is greatly reduced. Moreover, we can implement the required level of security by tuning the threshold as high or low as we want for each of the biometrics (behavioural or physical) [16].

To improve speaker identification systems, multiple strategies have been designed and tested. The possible points of improvement in this process are:

1. Improving quality of voice through environmental control;

2. Implementing a feature extraction algorithm;

3. Implementing a learning and pattern-recognition algorithm; and

4. Fusion of other biometrics with voice.

Environmental control is out of the scope of our research. Speaker identification systems should be able to work with noise pollution, background noise and voiced and unvoiced audio streams. For a few systems these preconditions may not exist, but for the others, improving the hardware for recording presents a preferable solution.

Recent improvements in wavelet-based spectral analysis [18; 19; 20] have shown great improvement in signal analysis compared to the classic model of MFCC [2; 21], which requires more computation and memory. This is an area that deserves further investigation.

Machine learning and Artificial Intelligence (AI) are considered in the third step, i.e., pattern recognition based on extracted feature sets. Machine learning and AI concern the construction of systems that can learn from data. In general, these algorithms process large amounts of data, discover patterns in data, and construct predictive models. The core of machine learning deals with representation of data instances, and their generalisation, which is a task performed on unseen data instances. Machine learning has been widely used in a number of fields like computer vision, language processing, recommender systems, etc. There is a wide variety of machine learning methods, and these fields are continuously evolving with better learning algorithms. Our second area of improvement is in this context.

A strategy using more than one biometric feature, often known as multimodal biometrics, is considered as the third area for the improvement of speaker identification systems. It combines more than one means of identifying a speaker (in this case, voice and face) and therefore, is more reliable and robust than unimodal systems.

During the course of this research, these three areas of improvement were greatly explored with the purpose of producing more accurate and robust speaker identification algorithms. In essence, improving real time response, accuracy and scalability were the main motivations of this research.

This chapter serves as an introduction to aspects of the problem of speaker identification. It highlights the basic architecture of a speaker identification system and describes the functionality of each step involved in this process. Subsequently, the goals of this doctoral thesis are outlined, followed by the main contributions.

1.2 Research Challenges in Speaker Identification

Over the last fifty years, speaker identification systems have evolved from scientific curiosities to popular commercial products sold in the market. Speaker identification has been fully embraced in Automated Phone Banking and Automated Help Centres where automated voices guide users towards specific options to perform a certain task once they are authenticated. People tend to speak naturally over these phones and they often want a quick recognition and the perception of a human-like response on the other end. If the user feels that this process is taking too long, or its authentication/recognition is totally incorrect, then he may lose interest.

However, as speaker identification technology has not yet reached maturity, researchers need to investigate various factors such as noise, the speed at which words are spoken in human conversation, physical and psychological factors causing two utterances of the same words interpreted as two unique words, etc. In the current systems, it is difficult to perform an accurate analysis and respond quickly, even if a speaker has been trained for voice identification.

Physical and behavioural biometrics can be combined to yield more accurate identification of the test data [16]. Fusing two or more modalities together has the advantage of strengthening and overcoming the weaknesses of each. This integrated approach increases the probability of correctly identifying a given speaker. Fusion can occur at the following various levels:

1. Sensor-level fusion;
2. Feature-level fusion; and
3. Decision or score-level fusion.

Fusion based approaches generally consider one of these strategies to increase the identification rate. Sensor-level fusion is carried out at the very beginning of the entire process. Information from multiple sensors is merged or fused together before further processing. Subsequent pre-processing, modelling, training and testing are carried out on this merged information. Many methods have been proposed in this context. The fusion of hand and face biometrics at feature extraction level has been proposed, and Infrared (IR) based face recognition has been fused at the feature-level, showing a substantial improvement in recognition performance.

In feature-level, the fusion is carried out after data pre-processing and feature extraction from multiple modalities. These features are then combined and fused together to form a single matrix. The fused feature set is then used for the modelling, training and finally, testing of the system. There have been immense research on multimodal biometrics over the last few years; however, fusion at the feature-level is a rather underexplored area. It is also somewhat hard to implement in reality since different modalities usually have feature sets that are incompatible.

Decision or score-level fusion is similar to the bagging or boosting techniques in machine learning and data mining. Multiple classifiers are presented with the same training and test data, and the classification is done on the basis of their combined effective classification result.

1.3 Research Objectives

The main goals of this doctoral thesis are summarised as follows:

1. To develop robust and real-time algorithms for speaker identification using multi-resolution statistical approaches;

2. To investigate and evaluate new hybrid intelligent methods using DWT and neural networks;

3. To develop a robust real-time speaker identification system using a vowel formant-based approach;

4. To investigate and evaluate different feature extraction, classification and fusion techniques for audio-visual speaker identification; and

5. To increase the performance and reduce the execution time of a speaker identification system.

1.4 Overall Contribution

In this doctoral research, we made the following contributions in improving the state of research and development in speaker identification:

1. We implemented a fast text-independent speaker identification system based on DWT and parallel neural networks including PNN, RBF-NN and GRNN. Tests conducted on the GRID[1] [2] audio-visual corpus showed that the system gained competitive accuracy and identification time compared to contemporary approaches.

2. We implemented a novel method based on formant analysis applicable for highly scalable speaker identification systems with minimised database storage of features for each speaker. Tests conducted with the proposed score-based scheme affirm a net reduction of 50% in storage data without compromising the accuracy of speaker identification.

3. We proposed, designed, implemented and tested a novel investigation leading towards feature-level and decision-level fusion of audio and face data to construct hybrid Gaussian Mixture Models (GMM) and PCA models for speaker identification. Test results confirmed a 15% improvement in accuracy over the contemporary methods in the same field.

Figure 1.1 outlines a summary of the contributions of this thesis.

1.5 Thesis Organisation

This section gives a brief outline of this thesis. A concise description of each chapter follows.

[1]GRID (short for *Global Resource Information Database*) is a large multi-talker audio-visual sentence corpus to support joint computational-behavioural studies in speech perception.

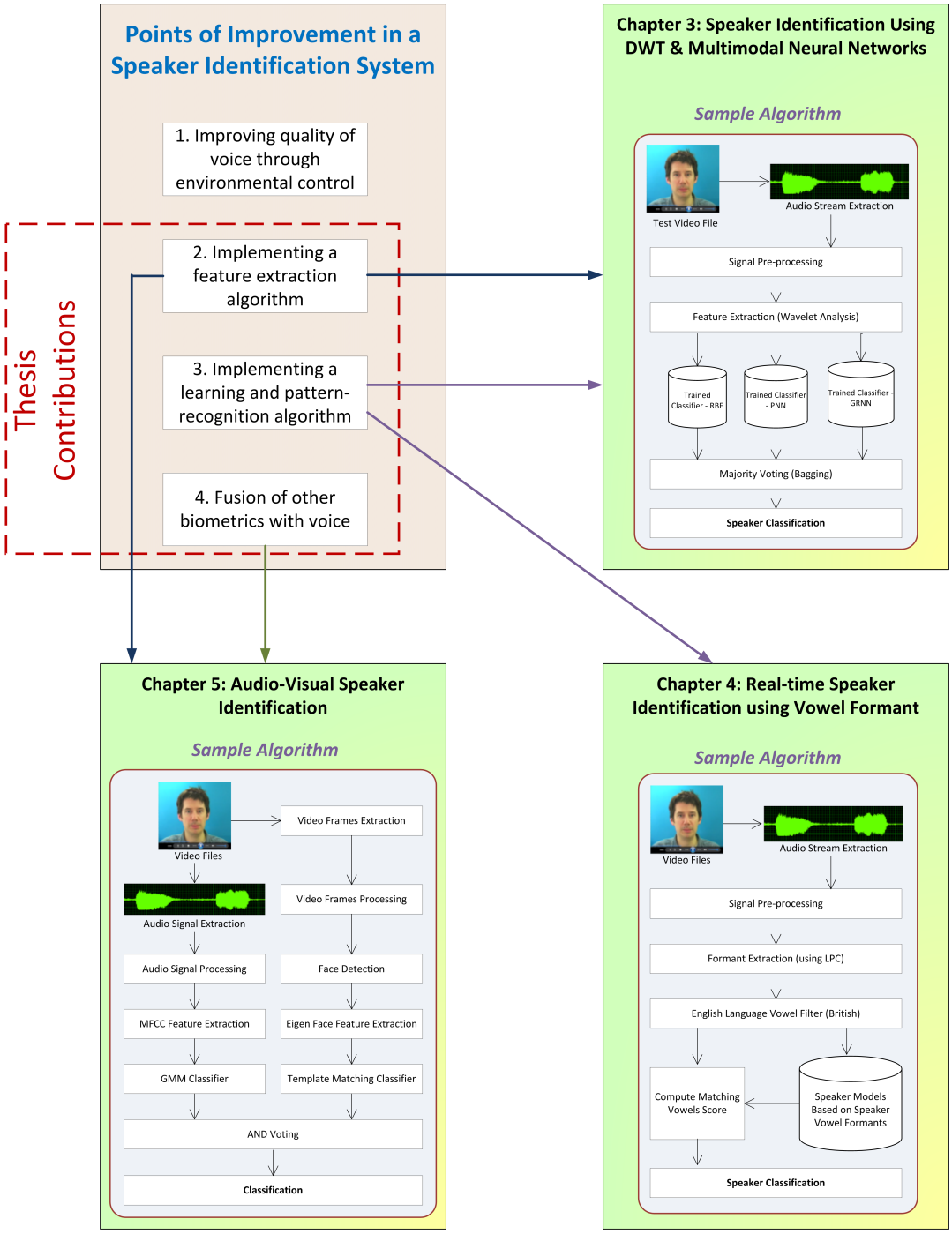

Figure 1.1: Summary of thesis contributions.

Chapter 1 serves as an introduction to the problem of speaker identification and its various aspects. It highlights the major challenges in various phases of speaker identification, and the numerous factors that make it an active research problem. Subsequently, the motivation behind this research and the research objectives of this doctoral research are outlined, followed by our major research contributions. The thesis organisation sums up the entire work in the next chapters. Chapter 2 presents a survey of the available approaches in speaker identification systems. Next, it offers general overviews of face recognition technologies and current techniques for fusion of both audio and video for speaker identification. A summary of all these approaches is presented in this chapter, along with relevant citations to prior existing work. Chapter 3 describes the design and implementation of a text-independent multimodal speaker identification system based on wavelet analysis and neural networks that we developed. Wavelet analysis comprises Discrete Wavelet Transform (DWT), Wavelet Packet Transform (WPT), Wavelet Sub-Band Coding (WSBC) and MFCC [22; 23]. The learning module comprises a Generalised Regression Neural Network (GRNN), a Probabilistic Neural Network (PNN) and a Radial Basis Neural Network (RBF-NN), forming decisions through a majority voting scheme. The system was found to be competitive and it improved the identification rate by 15% as compared to the classical MFCC. In addition, it reduced identification time by 40% compared to Back Propagation Neural Network (BPNN), GMM and PCA based on the performance testing conducted on the GRID [2] corpus. We have produced one publishable chapter as a result of this research. In Chapter 4, we describe the design and implementation of a highly scalable speaker identification system based on formant analysis. It discusses the limitations of the MFCC-based approaches and investigates the usefulness of formant analysis to counter these limitations. It begins with an overview of vowel formants. Next, it presents different methods of detecting English vowels in the audio signal using LDA, PCA and MFCC. A score-based algorithm is proposed for speaker identification. The performance of the proposed system is compared with that of GMM [24] and Back Propagation Algorithm (BPA) on the GRID [2] corpus. We have also produced one publishable chapter as a result of this research. In Chapter 5, we apply our research and investigation to propose a novel approach for speaker identification by fusing two

different modalities of face and voice at the feature and decision-levels. Performance testing and benchmarks conducted on the GRID [2] and VidTIMIT[1] [11] audio-visual corpora are reported in a publishable chapter. It evaluates GMM, Eigenface Approach (EFA) and the hybrid of both approaches for feature-level and decision-level fusion. Next, it describes the architecture of the proposed scheme and comprehensively validates the test results in comparison with the research done in previous chapters. The chapter concludes with a description of limitations of the fusion approach and a discussion of future directions. We conclude this thesis in Chapter 6 and present a brief overview of the aims and objectives fulfilled in this work. We provide a brief overview of the limitations of the research described in this thesis and give guidelines for future research.

[1]TIMIT, named after *Texas Instruments* (TI) and *Massachusetts Institute of Technology* (MIT), is a corpus of phonemically and lexically transcribed speech of American English speakers of different sexes and dialects. VidTIMIT (Video TIMIT), one of the derivatives of TIMIT, is a useful tool for research involving audio-visual speech identification.

Chapter 2

Related Work

2.1 Introduction

The field of speaker identification has evolved greatly with the expansion of technology in recent years. There are huge market interests for developing accurate authentication and security systems. In this chapter, we give a critical review of the existing approaches to solve the problem of speaker identification.

Although in the present chapter, we focus mostly on speaker identification technology, many of the algorithms, with a little modification, are also applicable to speaker recognition in general. For example, Section 2.3, which outlines the general architecture of a speaker identification system, feature extraction methodologies and pattern recognition algorithms, can very easily be applied to both speaker verification and speaker identification, and hence to speaker recognition in general. The algorithms developed throughout this doctoral thesis focus primarily on the speaker identification task. These algorithms can be adapted to deal with the sub-problem of speaker authentication with a little code change.

In Section 2.2, we provide an overview of the major differences between text-independent and text-dependent speaker identification systems. In Section 2.3, we highlight the basic architecture of a speaker identification system based on voice biometrics and feature-extraction strategies, and present a literature review of the existing systems based on voice biometrics. In Section 2.4, we discuss an identification system based on visual identity (face), its basic architecture, feature-extraction strategies and we review the literature on the existing ap-

proaches relevant to our research. In Section 2.5, we explain information fusion and its variations, the basic architecture of an identification system based on fusion of face and voice and present a literature review of existing audio-visual approaches. In Section 2.6, we summarise the discussion while focusing on the limitations and drawbacks of the strategies presented in this chapter.

This chapter begins with a literature review of the available speaker identification systems for text-dependent as well as text-independent approaches, followed by the general architecture of a speaker identification system. Next, we offer a general overview of face recognition technologies, followed by an overview of existing techniques for speaker identification employing fusion of both audio and video. This chapter ends with a summary of the limitations of existing approaches and proposes possible solutions.

2.2 Text-independent and Text-dependent Systems

Speaker identification systems are broadly classified into two main categories on the basis of textual information enclosed in the audio signal. On one hand, the text-dependent approach focuses on a pre-defined set of words or sentences which are used to train the speaker identification system for each individual speaker. The same sentences or set of words used in training are uttered for testing in order to complete the speaker identification/verification task. This is a relatively easier approach to implement and test because template-matching algorithms like Principal Component Analysis (PCA) and Vector Quantisation (VQ) are applicable here.

On the other hand, text-independent systems (also known as "text-free" speaker identification systems) do not limit the set of words used for training or testing of a given speaker. In the text-dependent system VidTIMIT [25], the identification phrases are known a priori. For instance, a speaker may be asked to read a random sequence of numbers [26]. These systems focus on speaker characteristics that are independent of the sounds/frequencies of the spoken words. These systems are the object of active research these days because of their applicability in diverse security and verification systems. Text independence implies

greater complexity in terms of variability of signal and consequently makes the speaker identification task more challenging and worth researching.

Generally speaking, speaker identification is implemented in two stages. At first, features are extracted from the speech. Then the extracted features are used for speaker classification [27]. The advantage of text-dependency lies in that the sentence used for identification does not need to be very long, it can simply be a word or an utterance. Unlike the text-independent system, shorter sentences can increase classification speed. Some observers have noticed that the performance of text-independent systems lags behind that of text-dependent systems. However, Markel and Davis [28] achieved excellent results with a linguistically unconstrained database of unrehearsed speech. Using voice pitch and Linear Predictive Coding (LPC) [29; 30; 31] reflection coefficients in their model, they reached 2% identification error and 4% verification error rates for 40-second segments of input speech. Results were not nearly as good with shorter input speech segments even though the system avoided operational problems of microphone degradation, acoustic noise and channel distortion. In text-independent identification of nine male speakers over a radio channel (in a test performed at the acoustics firm of Bolt, Beranek and Newman), the best performance had a 30% error rate for input speech segments of about two seconds.

The variability in phonetics presents itself as a major derogatory factor when it comes to accuracy of text-independent systems. There can be changes in the acoustic environment due to technical factors like transducer effects, channel variability, etc. Also, there can be "within-class" variations, which can arise from any changes in a speaker's mood, health, etc. These unwanted factors can lead to session variability, which is defined as the variation in between different recordings of the same speaker [32; 33]. It is one of the most taxing problems in any speaker identification system.

In this chapter, we give an overview of speaker identification, face recognition and fusion of face and voice technologies from the last three decades.

2.3 Speaker Identification Based on Voice

Traditional speaker identification systems are based solely on voice signal. A speech signal is captured in digital format through a microphone or the data is made available through a recorded audio file. The analogue signals are pre-processed to remove silence and noise, followed by a feature extraction phase in which speaker-specific spectral features across the data set of speakers are extracted to form the speaker models or data models. A learning algorithm is then trained on the data model, followed by the classification of unseen/seen signals into uniquely named speakers.

A speaker identification system is incomplete without the following two components:

1. Training (or enrollment) phase; and
2. Testing (or verification) phase.

In the training phase, speech models corresponding to different speakers are created. The speakers' voices are recorded and a number of features are extracted to create a voice print or template. In the testing phase, a speech sample from the current speaker is compared against the previously formed voice print and a decision is made. Identification systems compare the utterance against multiple voice prints and hence, requires more time than verification.

There are four main implementation steps for a speaker identification system, as given below:

1. Signal pre-processing;
2. Feature extraction;
3. Model training; and
4. Speaker classification.

Once the raw audio (whether recorded or in real-time) is obtained, a front-end pre-processing is first required. There are several methods of speech activity detection, which are employed so that the non-speech parts of the signal are

curtailed and the background noises are suppressed. Then features that contain speaker-specific information are obtained from the signal. There are no general set of features that can apply to all speakers. However, the speech spectrum shape encodes information about the speaker's vocal tract shape and glottal source using formants (resonances) and pitch harmonics, respectively [34]. These types of features are extracted from the audio signal and transformed to feature vectors. These vectors then serve as speaker models. The selection of the modelling techniques to be employed depends on the type of speech, level of desired accuracy, storage capabilities, computational requirements, etc [35]. More details on different modelling techniques and their use in speaker classification are discussed in the following sections.

2.3.1 Feature Extraction Techniques

This section presents the feature extraction methodologies that were used in speaker identification over the years. Feature extraction employs techniques for revealing the anatomical configuration of a target speaker. In such techniques, two kinds of aspects are usually useful for extraction from speech: acoustic and spectral. The acoustic aspects reflect both anatomy, like the size and shape of throat and mouth, and learnt behavioural patterns, like voice pitch, speaking style and accent. In spectral aspects, the use of features related to Cepstral analysis such as Mel-Frequency Cepstral Coefficients (MFCC) [36] and Linear Predictive Cepstral Coefficients (LPCC) [37] is most widespread. Similarly, very few techniques place emphasis on the characteristics of vocal cord vibration like the harmonic intensity information and fundamental frequency.

Research indicates that spectral features such as Mel-frequency Cepstral Coefficients (MFCC) are more efficient than acoustic features and are used widely due to their ability to represent the speech spectrum in a compact form. MFCC showed superior identification rate of 99.55% in a clean environment; however, the identification rate is only 60% in a noisy environment for the same data set. The reduction in the identification rate in the noisy environment is due to the fact that MFCC accepts a stationary signal within a given time domain whereas speech is a non-stationary signal. The drawback of MFCC is that it is not im-

mune to noise and thus cannot contain all the information in a speech in a given environment.

Since the late 1980s, speaker identification has been one of the applications that started using the wavelet concept to pre-process and extract the features of the signal to classify or identify speakers by varying the decomposition and regularity of the wavelet [38]. Several notable techniques and methods have been developed. The Discrete Wavelet Transform (DWT) [39], Wavelet Packet (WP) and Wavelet Packet Transform (WPT) [19] are some of the techniques that have been developed in the last few years to address speech feature extraction.

Automatic speaker identification has numerous applications in interactive voice recognition and biometric authentication. There are several widely used techniques available to automatically identify a speaker. This section will introduce signal representation forms, starting with Fourier transformation theory, followed by a relevant example application of MFCC. Then, an alternative called the multi-resolution analysis and wavelet theory will be discussed.

2.3.1.1 Short-time Fourier Transformation

An arbitrary signal given by some function can be represented in many forms for better understanding of the signal. Speech is a signal denoted by $f(t)$ that can be represented in different forms. The Fourier transform is one way of representing the signal and it is a frequency representation of the signal.

To study a non-periodic signal over the time, the given signal is divided into smaller signals and Fourier analysis is performed for each of them. This technique is called the short-time Fourier transform [40]. The signal is fragmented into smaller signals by a window function w,

$$F(s) = \int_{-\infty}^{\infty} f(t)w(t-\tau)e^{-j2\pi ts}dt. \tag{2.1}$$

Simultaneous analysis of signals in both the time domain and the frequency domain provides better understanding of the signal. The short-time Fourier transform lays the foundation for joint time-frequency analysis. The window size of the short-time Fourier transform needs to be determined up front to study the signal. If the window is too narrow, then the frequency resolution is poor, and

if the window is too wide, then the time resolution is poor. Short-time Fourier transform, when followed by various filter bank smoothing techniques, is a major component of spectral analysis. Bark-scale or Mel-scale frequency banks are usually employed to model the human auditory framework. Finally, the smoothed spectrum is transformed into Cepstral coefficients.

2.3.1.2 Mel-Frequency Cepstral Coefficients

One of the parameters employed most often for speech identification is MFCC. Figure 2.1 outlines the MFCC process of extraction and its major steps.

First, the speech spectrum is flattened to limit the dynamic range using a process of pre-emphasis. This is accompanied by using a first order filter.

Second, Hamming windowing is used to slab the signal into layers of frames with reduced edge discontinuity. A 20-30 ms length is used normally to achieve a balance between temporal and spectral resolution. Each frame has a time shift of approximately 10 ms, during which the changes in articulation configuration are negligible.

Third, the conversion process of a speech signal into a frequency domain is performed (DFT), which carries significant speaker information.

Fourth, Mel-scale band pass filtering is used to analyse the frequency resolution of auditory systems.

Fifth, the nonlinear compression of energy is approximated by sub-bank energy compression. At this stage, a log operation is performed to make the energy Gaussian distributed.

Sixth, the spectral information is transformed to the Cepstral domain. Here, the information is dominated by fewer coefficients. It is particularly important to note that MFCC (see Figure 2.1) and other short-time based features are extracted by employing a short-time window while reliance on adjacent frames is not considered.

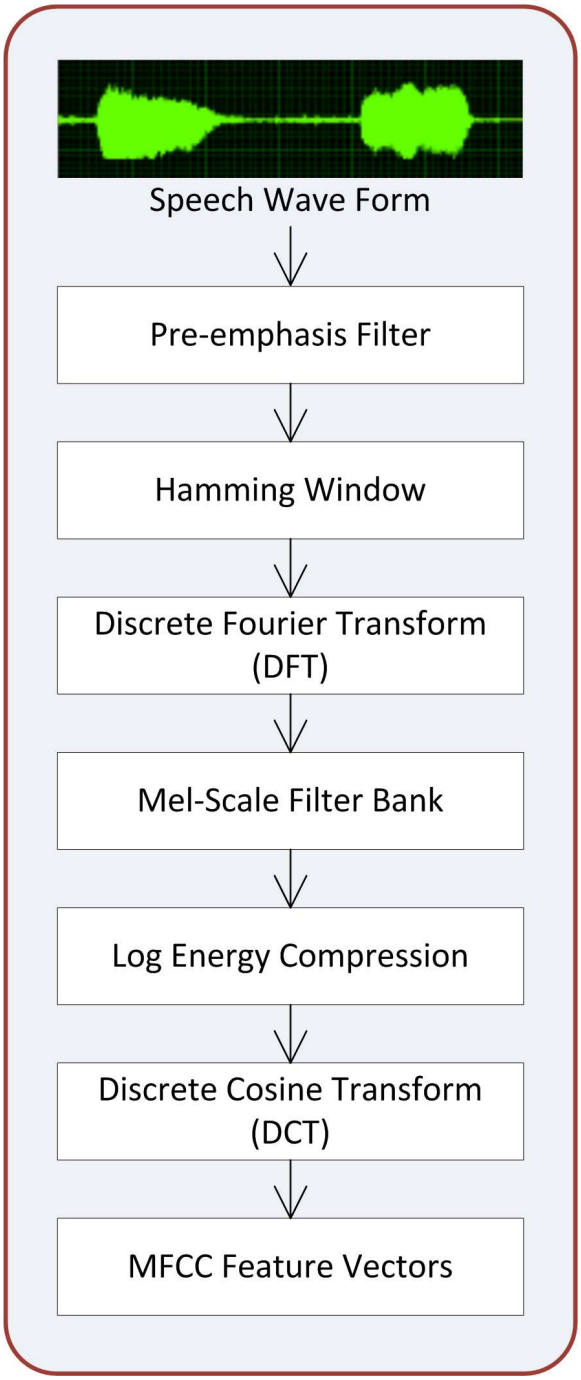

Figure 2.1: MFCC feature extraction process.

2.3.1.3 Multi-Resolution Analysis and Wavelets

The fundamental motivation for using wavelets lies in the analysis of signals in a scalable manner. The wavelet method splits up the given signal into a set of smaller signals and examines different frequencies of the signal with different resolutions. For instance, high frequency components benefit from delicate time-resolution, even if the frequency resolution is poor. The opposite is true for low frequencies. The window width can be altered while computing the transform for every spectral component. Furthermore, wavelets are good for estimating data that have sharp discontinuities. An example of how to demonstrate the signal in the wavelet domain using a short-time Fourier transform can be seen in Figure 2.2. Wavelets are a class of functions used to localise a given function in both space and scaling. A family of wavelets can be constructed from a function $\psi(t)$, called a mother wavelet, which is confined in a finite interval with zero average:

$$\int_{-\infty}^{\infty} \psi(t)dt = 0. \tag{2.2}$$

It is normalised ($\|\psi(t)\| = 1$) and centered in the neighbourhood of $t = 0$. A set of wavelets formed by the mother wavelets, called daughter wavelets and denoted $\psi(\tau, s, t)$, are formed by translating s and scaling the mother wavelets.

$$\psi_{t,s}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-\tau}{s}\right). \tag{2.3}$$

A signal's specific characteristics as well as the type of the application in hand affect the choice of wavelet. Wavelet families can have a wide array of differences in different properties. The support of the wavelet in time and frequency is an important factor. The decay rate, the symmetric properties, vanishing moments, etc. are all accounted for when choosing the appropriate wavelet.

Morlet, Meyer and Mexican Hat wavelets are a few mother wavelets widely used in various applications. These wavelets are symmetric in shape, and they are chosen based on their shape and ability to analyse a signal in a particular application. The Morlet wavelet is appropriate when there is a need for continuous analysis since it does not require any scaling function. The wavelet can be real or complex-valued. The scaling function of the Meyer wavelet, and the wavelet

itself, are defined in the frequency domain. The Mexican Hat wavelet is a special form of derivative of Gaussian (DOG) wavelets. It is proportional to the second derivative of the Gaussian Probability Density Function (PDF).

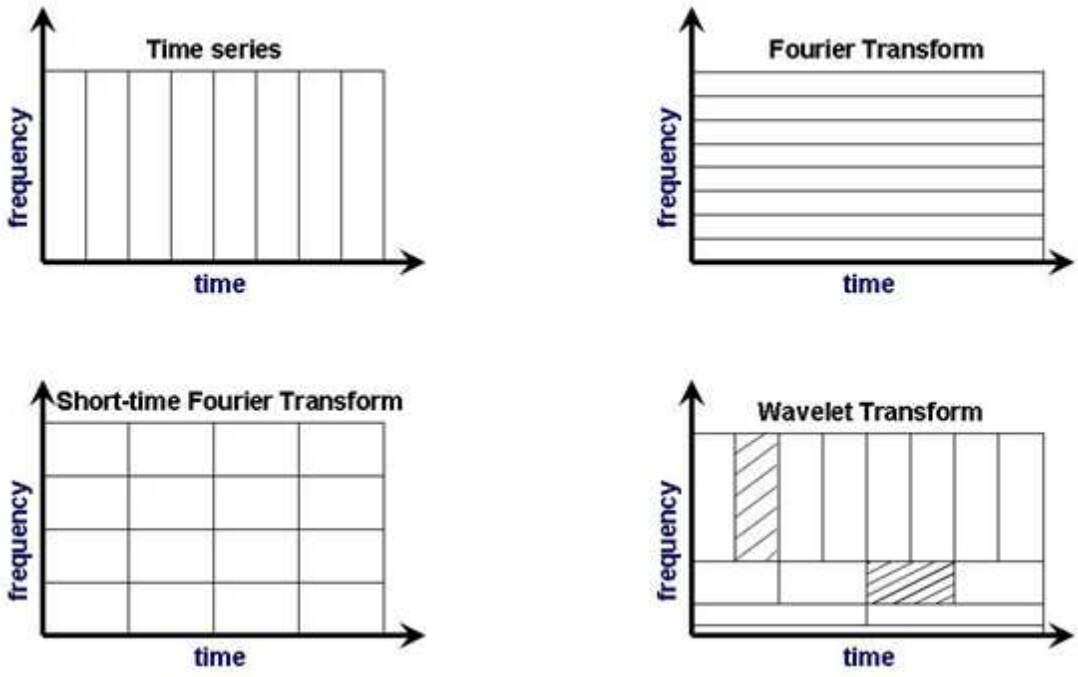

Figure 2.2: Time Series, Fourier Transform, Short-time Fourier Transform and Wavelet Transform.

2.3.1.4 Discrete Wavelet Transform

The DWT is a sampled version of the continuous wavelet transform. The DWT is easy to implement, requires little resource and is faster. In this context, digital signal filtering is a very helpful technique applied to generate a time-scale representation. Compared to the continuous wavelet transform, digital signal correlation computes a correlation between a wavelet at different scales, and the signal with the proper scale (or the frequency) is used as a measure of similarity. The continuous wavelet transform is computed by changing the scale of the

analysis window, shifting the window in time, multiplying by the signal and integrating over all times. In the discrete case, filters of different cut-off frequencies are used to analyse the signal at different scales. The signal is passed through a series of high-pass and low-pass filters to analyse the high and low frequencies, respectively.

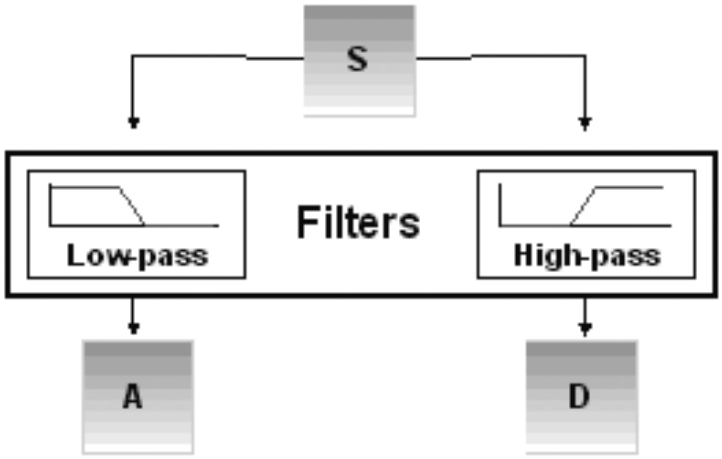

Figure 2.3: Wavelet filter [1].

As the basic model indicates in Figure 2.3, the given signal is decomposed into approximation A and detail signal D. Signal A carries low frequencies whereas signal D carries the high frequencies of the given signal. The filtering operations change the resolution of the signal, which is a measure of the amount of detail information in the signal while the upsampling and downsampling (subsampling) operations change the scale. The role of the latter is to reduce the sampling rate or remove some signal samples.

Subsampling by two refers to dropping every other sample of the signal. Subsampling by a factor n reduces the number of samples in the signal n times. These signals A and D can result in an increased overall sample length. For example, if a source signal S has length k, it will yield two sequences of length k each, making the total length $2k$. There is a better way of performing such decomposition. By looking carefully at the computation, we can keep only one out of every two

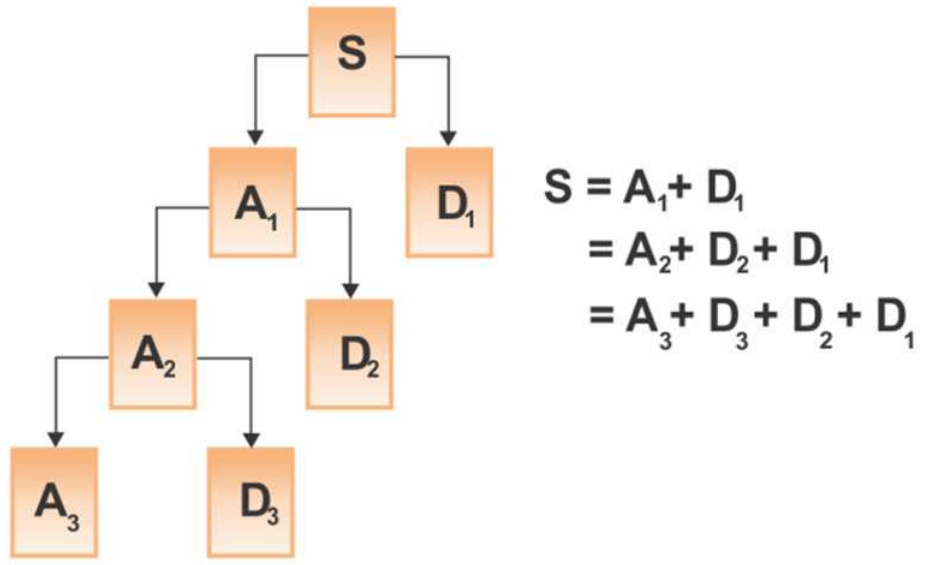

Figure 2.4: Wavelet analysis [1].

points (i.e. downsample) in each of the two k-length samples, which provides enough information. As shown in Figure 2.3, we produce two sequences called A and D.

2.3.1.5 Wavelet Packet

The wavelet packet method is a generalisation of wavelet decomposition that offers a richer range of possibilities for signal analysis [41; 42]. A wavelet is a function that looks like a small wave of the baseline. The wavelet basis is generated by stretching out and moving the wavelet to fit and cover all parts of the signal in different scales. A wavelet packet is an integrable modulated waveform, which is localised in position and frequency. It is usually assigned with the parameters frequency, scale and position. Wavelet packets can be used to expand a signal multiple times. The main characteristic of wavelets is the possibility to provide a multiresolution analysis of the image in the form of coefficient matrices. In wavelet analysis, a signal is split into an approximation and detail. The approximation itself is then split into a second-level approximation, and the process is repeated. For n-level decomposition, there are $n+1$ possible ways to decompose the signal.

In wavelet packet analysis, we can split both the details and the approximations. Figure 2.5 shows a depiction of a Wavelet Packet Decomposition (WPD) tree.

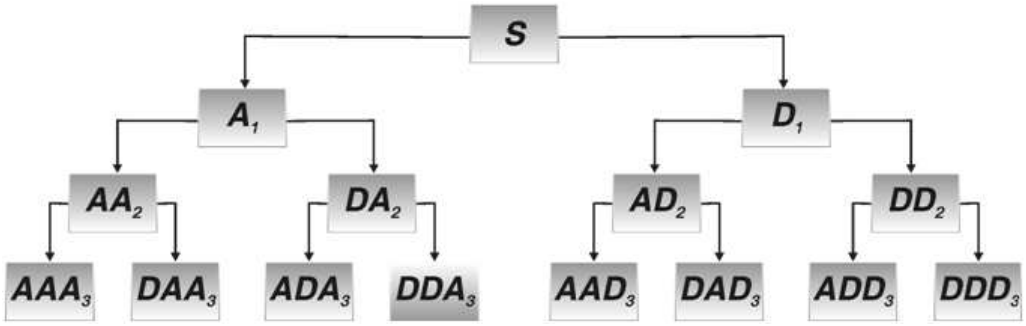

Figure 2.5: Wavelet decomposition tree [1] [2].

The wavelet decomposition tree is a part of the complete binary tree. For instance, wavelet packet analysis allows the signal S to be represented as A1 + AAD3 + DAD3 + DD2 (Figure 2.5). Usual wavelet analysis cannot facilitate this. We can choose one out of three of the possible encoding methods. An entropy-based criterion can be employed to choose the best decomposition for the signal. Every node of a decomposition tree can measure the information upon executing a split.

2.3.1.6 Wavelet Sub-band Coding

Speech and image compression applications have given rise to another method called sub-band coding (SBC), which was proposed in [43] using a special class of filters called quadrature mirror filters (QMF). The sub-band coding of speech [43; 44] spurred a detailed study of critically sampled filter banks. The advent of QMF in 1976 [43] allowed a signal to be split into two down-sampled sub-band signals and then reconstructed without aliasing. SBC can enable data reduction by removing redundant information in frequencies which are concealed. The result differs from the original signal, but if the discarded information is chosen carefully, the difference will not be noticeable or objectionable.

In this research, we simulate the fixed wavelet packed tree SBC proposed in Figure 2.6. The advantage of SBC is to model the human auditory system and

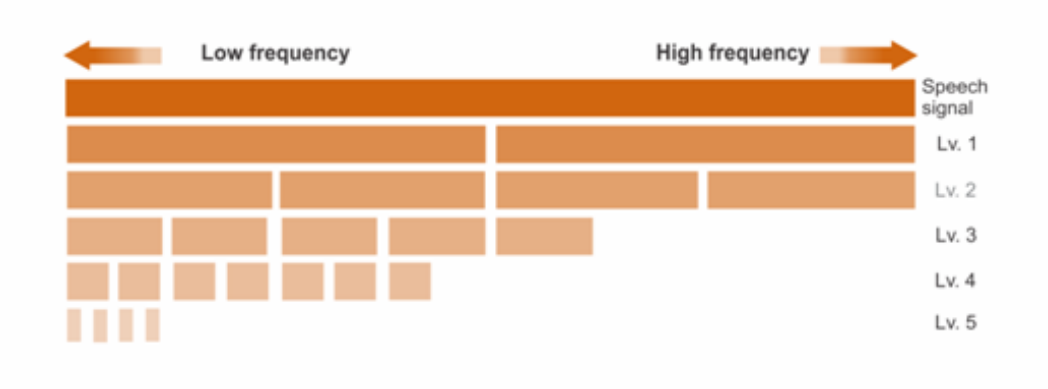

Figure 2.6: Wavelet packet in the Sub-band Coding (SBC) Domain [1].

decrease the number of parameters in the whole WPT which will reduce the time of speaker identification.

2.3.1.7 Discrete Wavelet with Irregular Decomposition

An irregular decomposition scheme based on WPT can significantly improve the performance of a speaker identification system. This procedure uses the energy of the speakers' utterances, which can eventually lead to speaker recognition. The energy distribution of conventional WPT illustrates that voice energies appear both in the low frequency region and a higher frequency region. So the decomposition concentrates on these regions. The resolution on lesser energy regions is decreased. Experimental results [1] have shown that a part of energy is conspicuous in the specific frequency region. It indicates that the energy distribution of speakers' utterances is uneven and the analysis can be focused on energy-centralised parts to prevent unnecessary operations.

The irregular decomposition method shown in Figure 2.7 can provide a better identification rate than conventional WPT and WPT in Mel scale. Using this method, the computational load can be eased off without sacrificing the identification rate. Since the energy in the speakers' pronunciation concentrates in specific regions, the decomposition is detailed on them. Apart from DWT, conventional WPT and WPT in Mel Scale, the irregular decomposition method has

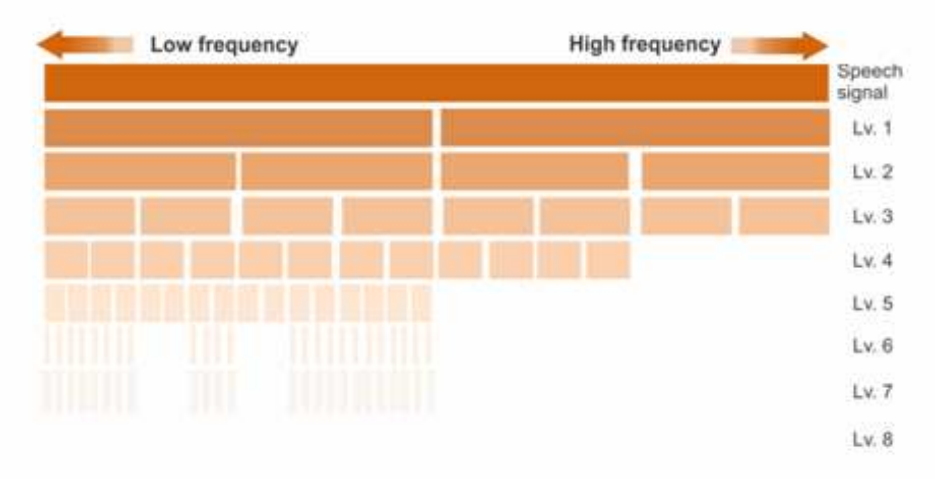

Figure 2.7: Discrete wavelet with irregular decomposition [1].

an improved identification rate without increasing the extracting time.

2.3.2 Speaker Models

The purpose of speaker modelling techniques is to identify unique patterns in a set of speakers or speaker models and help match features of a new instance. The speaker models contain enhanced speaker-specific information at a compressed rate [1]. There are multiple speech signals per speaker and during training speaker models are built using the specific voice features extracted from the current speaker. In the testing phase, the speaker model is compared to the current speaker model for identification or verification purposes [45]. Three main types of modelling techniques were mentioned, namely: template matching, stochastic modelling, and neural networks.

In template matching, a simple feature template from the frame of a speech can represent the speaker model. The distance between the input feature vector and the templates modelled in the system database is measured to calculate the matching score in determining the identity of the speaker. The Euclidean distance is commonly used for this purpose. There are a few universal template-matching systems that apply to all types of features, but the others are applicable to specific

types of feature representation schemes. These systems use example templates as prototypes and hence, are not hard to train. Template matching systems can become computationally cumbersome if the number of prototypes get too high. Lookup techniques can prove to be effective for identification if only a small-size alphabet is used to extract tokens.

The statistics of the feature vectors are used to obtain classifiers in stochastic matching. Usually, the variance and correlations of the individual features, and the average feature vectors are utilised. A large class of modelling challenges has successfully been resolved by employing neural networks as well. These networks, although assembled from simple elements, can learn from and adapt to many intricate and ambiguous problem scenarios.

2.3.2.1 Dynamic Time Warping

DTW is one of the most dependable and widely used template based methods for text-dependent speaker identification systems. It enables an automated system to prearrange multiple dissimilar sequences and search for the optimal matches that exist among them. Independent of some of the specific variations, a similarity-measure can be computed by warping these sequences non-linearly in time. Time-series classification often makes use of this sequence alignment.

DTW is a technique that uses dynamic programming to process text-dependent input feature vectors to remove the effect of speech rate variability by the speakers. The matching score obtains the speaker model saved in the database and compares it to the feature vectors frame by frame to identify the speaker [46]. The major optimisations to the DTW algorithm arise from observations on the nature of paths through a DTW grid, such as the monotonic, continuity and boundary conditions, as well as the condition of the adjustment window.

2.3.2.2 Vector Quantisation

Vector quantisation is an efficient way to compress large training vectors by using codebooks. Codebooks contain the numerical representation of features that are speaker-specific. During the training phase, the feature vectors obtained for a given speaker are clustered to create their codebook. In the test phase, input utterances are vector quantised and the VQ distortion that is calculated over the

entire utterance is used to determine the identity of the speaker. There are many types of codebook generation algorithms but the most well-known and widely applied one is the K-means algorithm [47].

The steps of the K-means algorithm are:

1. Cluster the vectors based on attributes into k partitions of centroids.
2. Assign each feature vector to the centroid that is nearest to it.
3. Calculate the position of the k centroids using the means of the distances between the features and the centroids.
4. Repeat steps 2 and 3 until the position of each centroid no longer changes.

The advantages of K-means clustering lie in its simplicity and ease of computation. However, the K-means method does not guarantee that the classification of speech spacing is optimal.

2.3.2.3 HMM or Single-state GMM

Single state Hidden Markov Model (HMM)-GMM is a robust parametric model for text-independent speaker identification as reported in [48]. A Gaussian Mixture Model (GMM) is a parametric learning model and it assumes that the process being modelled has the characteristics of a Gaussian process whose parameters do not change over time. This is an excellent assumption as a signal can easily be assumed to be stationary during a Hamming window. This is also a true assumption in the case of a human face where the distance between the eyes, the shape of the nose and the area of the forehead, do not change within a single frame.

A GMM tries to capture the underlying probability distribution governing the instances presented during training of the GMM. Given a test instance it tries to estimate the maximum likelihood of the test instance given each trained GMM model. The model with the maximum likelihood given its GMM is declared as the GMM model of origin. The voice and face GMM models for each speaker are developed separately, and also fused together for this purpose. In this study we present an overview of the GMM used in our system.

Generally, the text-independent speaker identification system is modelled as the statistical speech parameters' distribution model, which uses GMM as the model of each speaker and as the Universal Background Model [48; 49]. A GMM tries to fit a Gaussian distribution around the training samples and estimate the underlying distribution's parameters through an iterative approach called Expectation Maximisation (EM). The likelihood given a Gaussian distribution is given by the equation,

$$N(x|\cup,\varepsilon) = \frac{1}{(2\pi)^{d/2}\sqrt{|\varepsilon|}} + e^{\left(-\frac{1}{2}(x-\mu)^T\varepsilon^{-1}(x-\mu)\right)}, \tag{2.4}$$

where d is the dimension of x, μ is the mean and ε is the covariance matrix of the Gaussian. Usually ε is either diagonal or full. In this study we have experimented with full covariance.

The likelihood of a test pattern given a speaker GMM is,

$$P(x) = \sum_{i=1}^{N} w_i.N\left(X|U_i,\varepsilon_i\right), \tag{2.5}$$

where N is the number of Gaussians and w_i is the weight of Gaussian i, with

$$\sum_{i=1}^{N} w_i = 1 \quad \forall i : w_i \geq 0. \tag{2.6}$$

Initial parameters for each GMM are set using K-means, and estimated to fit the training samples using Expectation Maximisation (EM). Detailed technical specifications of this algorithm have been described by Memon et al. [50].

2.3.2.4 Neural Networks

The ability of neural networks to recognise patterns of different classes makes them suitable for speaker identification. A typical neural network has three main components: the top layer, the hidden layer, which can be one or more layers, and the output layer [51]. Each of the layers contains processing units that represent the interconnected neurons. During the training phase, the weights of the neurons are adjusted using a training algorithm that attempts to minimise the sum of squared difference between the desired and actual values of the output

neurons. The weights are adjusted over several training iterations until the desired sum of squared difference between the desired and actual values is attained [52]. To put it simply, neural networks are used to model patterns between inputs and outputs.

Artificial Neural Network (ANN) performance depends mainly on the size and quality of training samples [53; 54]. When the number of training data is small, not representative of the possible space, standard neural network results are poor. Fuzzy theory has been used successfully in many applications to reduce the dimensionality of feature vector [55].

A neural network consists of multiple perceptrons combined in multiple layers, starting from the input layer, followed by one or more hidden layers and ending at the output layer. Each perceptron has a weight associated with it. These weights are adjusted during training to map the training instances to known target concepts. At the end of the training, a tuned weight matrix that corresponds to a complex function is produced to map inputs to outputs.

The most common types of neural networks include Back Propagation Neural Networks (BPNN) [56] and feed-forward networks. A training input is passed through the network a number of times to adjust the weights accordingly. The iterative process of training the data requires multiple passes through the network to train it correctly. This takes a large amount of time before the network converges to a fine-tuned weight matrix. Therefore, ANNs are notorious for long training times and over- or under-fitting of the training data. This aspect yields a poor performance for unknown test data or for new instances that are not present in the training. For speaker identification, the GRNN was introduced in [57], and the PNN in [58].

When the decisions are strict or the dimensionality is high, combining multiple neural networks can serve as a superb machine learning mechanism [6]. The combination of multiple neural networks eliminates the poor performance due to over- or under-fitting of the training data with individual neural networks because each network has a different level of hypothesis generalisation capability. The combination of multiple neural networks resolves the problem of high identification rates but complicates the method further by increasing the training time.

Jian-Da [51] reported that the Back Propagation Algorithm (BPA) over-fits the training data and has a higher error rate than radial basis neural networks [51]. These two problems were the main motivation for developing a scheme that resolves the under- and over-fitting problems and minimises the training time.

2.3.3 Existing Approaches

This section discusses the main speaker identification systems found in scientific literature. In the beginning of the twenty-first century, GMM was employed by Reynolds et al. [46; 48] for the purpose of robust text-independent speaker identification. GMM has Gaussian components that correspond to several speaker-dependent spectral shapes, which can be exploited in characterising a speaker's identity. This scheme works well on unconstrained conversational speech, especially the short ones, and it is robust to transmission channel (e.g. telephone) degradations. A comprehensive evaluation of this method was conducted using 49 speakers using a conventional telephone speech database. The paper explored, through experiment, the effects of several key issues like initialisation, variance, and model order selection. It was compared to unimodal Gaussian, VQ, tied GM and RBF. It was shown that the implemented identification system based on spectral variability, tested on 49 speakers, achieved a higher accuracy rate (96.80%) using clean speeches than telephone speeches (80.8%).

The iterative clustering approach, along with perceptual features, was demonstrated by Revathi et al. [52] in 2009 both for speaker and speech identification. For speaker-independent speech utterances and for text-independent speaker identification, the training models were developed using different training speech formations. The main emphasis was on the use of clustering models. These models were developed for the training signals for the Mel-frequency perceptual linear predictive cepstrum. Speaker identification achieved a 91% accuracy rate, the isolated digit recognition also achieved the same, while continuous speech recognition had the best (99.5%) accuracy rate. The Equal Error Rate (EER) was also reasonably low (9%). Tested on the TIMIT database, this algorithm, through an iterative clustering approach, Perceptual Linear Predictive (PLP) cepstrum and Mel-frequency PLP (MF-PLP) achieved a 91% accuracy

rate with 50 random speakers.

An efficient and modern Speaker Identification (SI) system needs an error-resilient feature extraction component and speaker modelling technique that can extract and exploit generalised depictions of these traits and features. MFCC has been modelled on human auditory systems by many researchers, and made use of standard acoustic feature sets. In 2009, Chakroborty et al. [59] utilised MFCC as an established feature extraction strategy. They combined MFCC and Inverted MFCC (IMFCC) based on Gaussian filters, achieving a 97.42% accuracy rate with 131 subjects from the YOHO database. The authors have shown that IMFCC has the potential of forming quite useful feature sets for speaker identification. The high frequency portion of IMFCC contains valuable additional information. Although Triangular Filters (TF) were previously used in such systems, this work employed Gaussian-shaped Filters (GF). As a result, the sub-band outputs had a high correlation between them. The GF was proven to be the better option over TF. Tested both on microphone and telephone speeches (YOHO and POLYCOST databases, respectively) MFCC and IMFCC demonstrated authentic performance with this novel approach. Over 130 speakers were tested in individual and fused modes.

In 2009, Saeidi et al. [60] achieved a 97% accuracy rate with 34 speakers by applying Kullback-Leibler (KL) divergence. They used a single microphone to record multiple speakers and attempted to identify the speakers from this co-channel scenario. The accuracy depended much on the availability of the Signal-to-Signal Ratio (SSR) approximation. When the SSR was assessed, the identification rates were high. The authors also explored the scenario in the absence of SSR estimation using adapted GMM and KL divergence. They achieved fairly high rates of accuracy (97% or 93% from different trials).

In 2011, Gomez [61] implemented an identification system based on a novel parametric neural network, and achieved 94% accuracy with 40 speakers. The published work presented a novel speaker identification technique which was aimed at high identification accuracy and low imposter acceptance. This method employed a neural network with modifications based on the Self-Organising Map (SOM) in various dimensions. This scheme aimed at improving the imposter rejection rate as well as the overall accuracy rate.

The proposed Multiple Parametric Self-Organising Maps (M-PSOM) method was a classification technique tested on the CSLU (Oregon School of Engineering) speaker corpora. The proposed methodology, designed based on a parametric neural network for the individual speakers, yielded satisfactory results. The earlier approaches used single NNs for the entire SI system. The authors of this work showed that a speaker's acoustic signature could be uniquely represented using their parametric NN. The achievement of an identification rate of 85% and above indicates the usefulness of this feature extraction and modelling method. Table 2.1 covers a number of other such feature extraction and modelling techniques.

This section has presented an overview of the most widely known concepts in the field of ASR. The steps to design an ASR begin with pre-processing, feature extraction, modelling and, finally, comparison using distance methods.

2.4 Face Recognition Technology

Face recognition is the task of identifying a person based on image data (stored or captured live). To verify someone's identity, at first, the image has to be analysed to find the face portion (face detection task) before any features of the individual face can be extracted. Although face recognition has been studied for decades, there are multiple research projects being pursued every year because it is a complex problem encompassing both computer vision and machine learning. This situation gives rise to a number of possibilities for research and improvement. A formal algorithm for face classification first appeared in [62] [63]. The author proposed the collection of users' facial profiles in the form of curves, finding their mean (average/norm) and then classifying other users by their distance from the mean profile.

Although face recognition systems have started to become part of real-world applications [64], they still face numerous challenges, such as variation in light, pose, age, gesture and expression of the speaker [65]. The current research does not aim to deal with these constraints individually, rather, it focuses on using and finding the best candidate algorithms for fusing face and voice modalities to improve the overall performance of speaker identification. The state-of-the-art methods for face recognition from video have a complexity of $O(n^3)$ [66] where

n is the number of frames. Next, we present an overview and illustrations of the sub-processes involved in face recognition.

2.4.1 Sub-processes in Face Recognition

The problem of face recognition can be divided into the following sub-processes in order to make clear the road for improvement of each component.

1. Image acquisition;
2. Pre-processing;
3. Face detection;
4. Feature Extraction; and
5. Face classification.

These sub-processes are illustrated in the specific order they are performed in Figure 2.8.

Image acquisition is the first step of any standard face recognition system. An image can be acquired by a camera, stored on the computer hard drive or extracted from a video containing images in various frames. This research used GRID [2] and VidTIMIT [11] audio-visual corpus to verify and benchmark existing as well as newly proposed models.

GRID [2] and VidTIMIT [11] have video files containing the speaker's upper body image including the head, face, neck, shoulders, abdomen and the background. The speaker keeps his lips moving throughout the provided videos. The image acquisition phase (in our case the video capturing device) can not benefit from a software but more advanced hardware can capture the video or image with high resolution and density and thus contribute towards improvement in face recognition in the later phases.

Image pre-processing involves various techniques to prepare the image data for processing in the later stages. This may include colour normalisation, histogram normalisation, smoothing, filtering, grey scale transformation for dimensional reduction, etc. This research applied image size reduction and grey scale transformation to improve the computation time of the overall system.

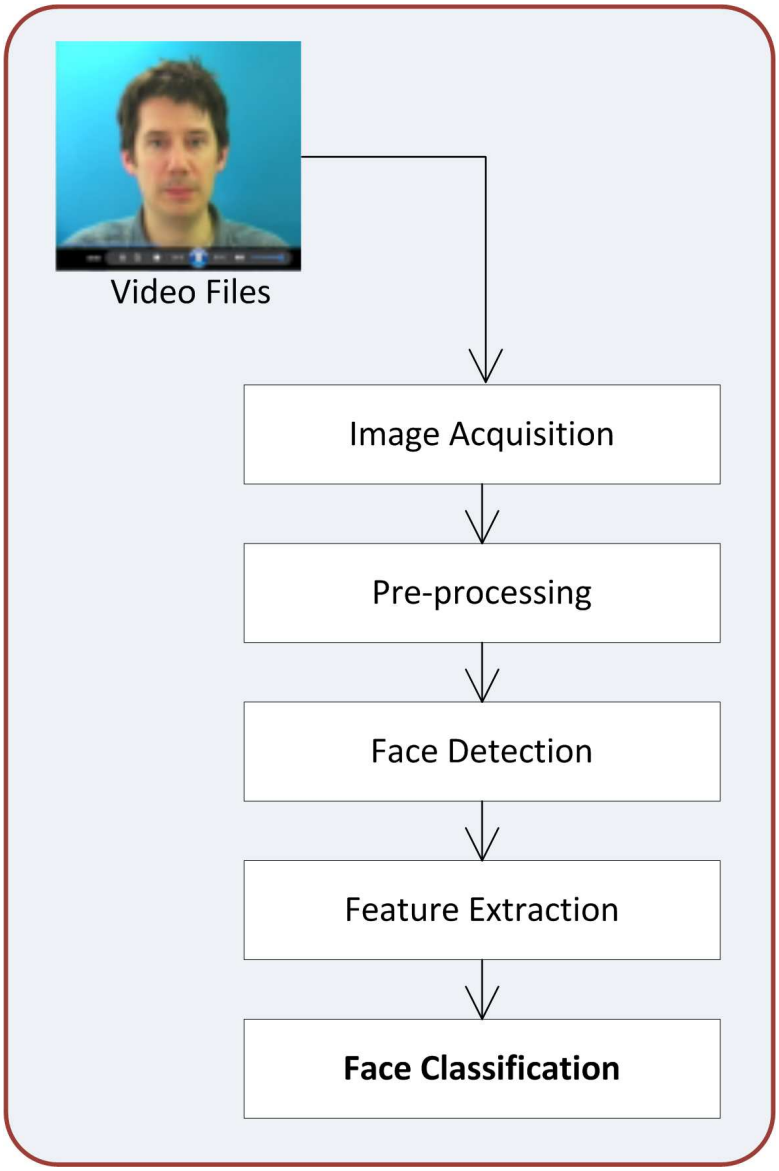

Figure 2.8: Sub-processes in face recognition.

Every human face includes hair, forehead, ears, eyes, nose, cheeks, lips and chin. These objects are also relatively positioned in a specific setting and order on the face. Although the template is the same, the size of each component as well as the relative distances are different from face to face at least on the structural

level. Therefore, the face object can be detected in comparison with other objects based on such templates.

This research utilised a Cascade approach based on the Haar algorithm [67] to detect the face portion in the form of a rectangular shape. This not only reduced the computation cost but also improved the face recognition accuracy. These Haar-like features were tested with convolutional neural networks that perform accurately as well [68].

In the next section, we provide a summary of recent methodologies employed for face recognition. In Section 2.4.2, appearance-based approaches are described, which are divided into linear and non-linear techniques. In Sections 2.4.3 and 2.4.4, model-based approaches, and schemes based on various transforms, are summarised, respectively.

2.4.2 Appearance-Based Methods

In appearance-based approaches, images are considered as high-dimensional vectors, or points in a high-dimensional vector space. Statistical techniques are used to analyse the distribution of points in this vector space, and derive an efficient and descriptive representation of the data. Early appearance-based approaches were based on correlation [69] and template matching [70]. Broadly speaking, these methods can be either linear or non-linear.

2.4.2.1 Linear

Linear face recognition approaches include PCA [68], Independent Component Analysis (ICA) [71] and Linear Discriminant Analysis (LDA) [72], which are all based on linear transformations from the original images to feature vectors obtained using different projection schemes.

Principal Component Analysis: PCA [68] is a dimensionality reduction technique, based on the Karhunen-Loeve (KL) transform. In order to classify a given face image into the registered speakers' faces, the system is required to extract facial feature vectors to mark the identity of different speakers. This algorithm was first conceived as eigenface [73]. It operates by using features to extract principal components from the test faces, taking their mean face image

and later classifying the new faces based on their distances from the mean face image.

A 2D matrix of grey scale values is a flat representation of a face image. This grey scale face portion is also pre-processed to extract unique features called eigenfaces or eigenvalues. Eigenfaces or eigenvalues have been extremely useful for extracting unique features for image recognition. A mean image is formed by adding all given images (key frames) of a speaker from its video files with the formula,

$$MeanImage = \phi = \frac{1}{n}\sum_{k=0}^{n} x^k. \tag{2.7}$$

Following the mean image, each individual image is subtracted from the mean image to compute its distance from the mean image,

$$DifferenceImage = \Phi_i = (X_i - \phi)\,. \tag{2.8}$$

Next is the computation of M orthonormal vectors U_n such that

$$\gamma_k = \frac{1}{M}\sum_{n=1}^{M} \left(U_k^T \Phi^n\right)^2. \tag{2.9}$$

In this equation, U_k are the eigenvectors and γ_k are the eigenvalues, respectively.

Independent Component Analysis: ICA [71; 74; 75] is a computational method for separating a multivariate signal into additive subcomponents, which are all, statistically, assumed to be independent non-Gaussian signals. The independence among the estimated components is boosted to the maximum possible levels, and a special type of blind source separation is performed in finding the independent components. We may choose one of many ways to define independence, and this choice governs the form of the ICA algorithm. In [71], two architectures for face recognition using ICA are provided: statistically independent basis images and a factorial code representation.

Linear Discriminant Analysis: LDA [72; 76] is somewhat similar to PCA in the sense that they both attempt to find linear combinations of the associated

variables that would most closely fit the data. LDA explicitly attempts to model the difference between the classes of data. PCA on the other hand does not take into account any difference in class, and factor analysis builds the feature combinations based on differences rather than similarities. LDA seeks to reduce dimensionality while preserving as much of the class discriminatory information as possible.

Boosting: Boosting [77] is a machine learning meta-algorithm for reducing bias in supervised learning. Boosting is based on combining an ensemble of weak classifiers to produce one strong classifier. Most boosting algorithms consist of iteratively learning from weak classifiers with respect to a distribution and adding them to a final strong classifier. The addition is usually weighted according to the accuracy of the weak learners. Figure 2.9 gives an illustration of how an ensemble of weak classifiers can combine to form a strong classifier.

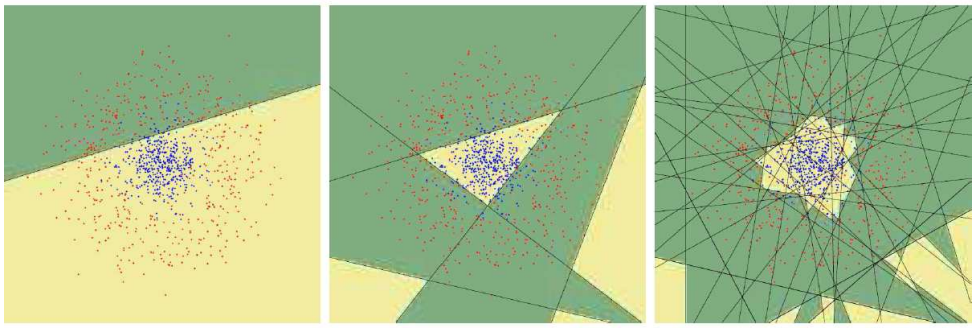

Figure 2.9: Illustration of classification produced by boosting algorithm AdaBoost after 1, 5 and 40 iterations (left to right) [3].

Different boosting algorithms vary in the ways they weigh the training data points. They also have different formulations for the underlying hypotheses. AdaBoost [78] is very popular and perhaps the most significant historically as it was the first algorithm that could adapt to the weak learners. There are several other boosting algorithms that have lately been developed, such as LPBoost [79], BrownBoost [80], etc. Boosting has been initially applied to face detection in [66] and to recognition in [81], where a recognition rate of 86% is achieved on a mixed dataset.

2.4.2.2 Non-Linear

Face images often contain complicated nonlinear variations, which is why various non-linear techniques have also been explored and applied to face recognition. The goal is to transform the face features from a space where the classes are not linearly separable into a higher-dimensional space where they are.

Support Vector Machines (SVM) [82; 83] are supervised learning models with associated learning algorithms that analyze data and recognise patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Although the SVM is linear in nature, it is usually combined with kernel functions to produce a non-linear classifier, such as Kernel Principal Component Analysis (KPCA) [84]. This approach involves using a kernel function to map face images to a higher dimensional space, followed by a PCA step. Kernel Linear Discriminant Analysis (KLDA) [85] involves using a kernel function to map the face images to a higher dimensional space, followed by an LDA step. The technique is applied to face recognition in [86], which employs a Nearest Feature Line (NFL) classifier, instead of SVM. The Kernel Independent Component Analysis (KICA) [87] has also evolved with various approaches. An entire function space of candidate non-linearities has been used in [88] instead of a single kernel function. The work in [89] adopted a different approach using a polynomial kernel to project the input image data into a high-dimensional feature space.

2.4.3 Model-Based Methods

Model-based methods represent faces [90] as parameters of a model instead of feature vectors, and makes it easy to adapt to different sources of variation like illumination, expression, pose, etc. Few such methods are discussed below.

2.4.3.1 Neural Networks

The role of neural networks in speaker identification based on voice has been discussed in Section 2.3.2.4, further details will be laid out in Section 3.3. NN has been widely applied to the field of face recognition as well. The auto-associative

back propagation feature extraction system employed in [91] used a back propagation network for feature extraction, and a second NN as a classifier. Tested on a small database, faces were recognised with high accuracy. In [92], a neural network committee machine obtained its classification result upon combining responses from a number of different NNs, each trained separately by different image blocks and features, and yielded a 95.7% recognition accuracy. In [93], a Self-organising Map (SOM) NN was used to reduce dimensionality and provide invariance to changes in the image sample. In addition, a convolutional NN was used to provide partial invariance to translation, rotation, scaling, etc. This system, tested on the AT&T database, reached an accuracy of 96.2%. The face recognition using Discrete Cosine Transform (DCT), LDA and RBF-NN presented in [94] was reported to have 97.5% accuracy based on the AT&T database, and was computationally efficient.

2.4.3.2 Elastic Bunch Graph Matching

Elastic Graph Matching (EGM) is an object recognition algorithm widely used in computer vision. This neural-inspired algorithm uses visual features that are based on Gabor wavelets. The visual processing in a brain is often modelled using these wavelets. The matching algorithm is a derivative of dynamic link matching, which models the brain's object recognition process. For object types that have a mutual layout, an enhancement of EGM has been developed which uses the same kind of graph in representing occurrences of the same kind. This matching method is known as the Elastic Bunch Graph Matching (EBGM) [4; 95]. Labeled graphs are used to characterise visual objects. The nodes of these graphs correspond to local textures (derived from Gabor wavelets), while distances between the respective nodes' positions in an image are denoted by the edges.

2.4.3.3 Active Appearance Model

The Active Appearance Model (AAM) is a generalisation of the widely used active shape model approach [5], but uses all the information in the image region covered by the target object, rather than just that near the edges. The algorithm uses the difference between the current estimate of appearance and the target image to drive an optimisation process. By taking advantage of the least

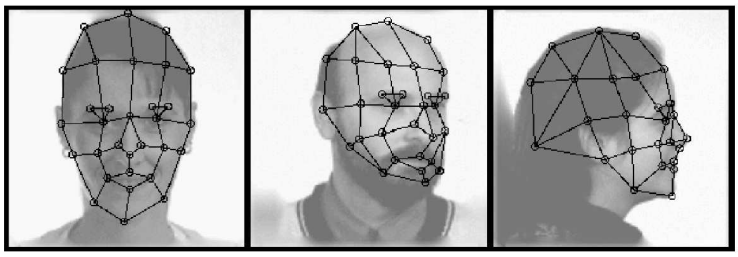

Figure 2.10: Fiducial points automatically located using elastic bunch graph matching [4].

squares techniques, it can match to new images very swiftly. Figure 2.11 shows the extracted shape for a face image, along with the appearance information, which has been warped to conform to the mean image shape to form a shape-free patch. The application of AAM in face recognition [96] has shown an 88% recognition rate based on a dataset containing 10 training images and 10 test images for 20 individuals. Being a rather complicated technique, it has limited use in face recognition applications.

Figure 2.11: A labeled training image provides a shape free patch and a set of points [5].

2.4.3.4 Hidden Markov Model

HMM [97] has effectively been used in speech recognition, as detailed in Section 2.3.2.3. It found a vast application in the area of face recognition in recent times as well. The values of the pixels from a face image can be utilised to form a top-down model as shown in [98]. This approach has been modified using DCT

coefficients to form observation vectors in [99]. In [100], DWT [101] coefficients were taken from overlapping sub-windows of the entire face image. A raster-scan of an entire image can also produce these DWT coefficients as shown in [102].

The structure of face images is two-dimensional (2D), whereas the HMM is one-dimensional (1D). In [103], the modelling of a face image was done using two standard HMMs, one each for observations in the vertical and horizontal direction. This technique has achieved results better than PCA and LDA [103]. The column sequences in the image can also be modelled as super states, which is known as the Pseudo 2D HMM (P2D-HMM) [104]. If the states are modelled using a rectangular constellation, the complexity can substantially be reduced compared to the P2D-HMM; this method is known as the Low-Complexity 2D HMM (LC 2D-HMM) [105]. The Hierarchical Hidden Markov Model (HHMM) introduced in [106] is capable of modelling the complex multi-scale structure in natural sequences. However, it is computationally complex and hence, impractical for face recognition.

2.4.3.5 Other Model-Based Methods

In reality, the structure of a face is three-dimensional (3D), and several methods have been developed to accurately represent its features [107]. In [108], a 3D model is created using textured scans of heads, and computer graphics are used to estimate the 3D shape and texture. In [109] a method is proposed that uses range and texture information to create a canonical surface for each face image. The major advantage of the canonical surface approach lies in its insensitivity to variation in head orientation and facial expression. However, the approach's practicality is reduced by the requirement of having range images for all subjects in the training database. Recognition rates achieved by 3D approaches are mostly great but they are usually computationally expensive.

2.4.4 Transform-Based Methods

It is often helpful to transform an image from its original spatial domain into another one to facilitate feature extraction. Such transformations do not alter the information content of the image, but changes the representation in a way

such that classification is easier. A few such methods deployed in face recognition are discussed below.

2.4.4.1 Radon and Trace Transforms

The Radon transform [110] is an integral transform whose inverse is used to re-construct images from medical CT scans [111]. If the projection is taken for a 2D function $f(x, y)$, it will comprise a set of line integrals. These line integrals are computed by the radon function in a given direction. Multiple sources are considered in parallel beams that have 1 pixel unit separation. The radon function takes several parallel-beam projections of the image from different angles by rotating the source around the center of the image. The trace transform [112; 113] is a generalisation of the Radon transform. It traces an image and calculates a specific functional along different lines. The different functionals that are used are often invariant to different transformations of the image. It scans along the lines of an image using an integral function, T. The trace transform offers rotational invariance. In [113], the Shape Trace Transform (STT) is applied to face recognition, which is performed by obtaining the trace transform on an image, binarising it and locating the edges. Other related techniques have also been developed [114; 115; 116] which perform reasonably well.

2.4.4.2 DFT and DCT

Although DFT is used in numerous pattern recognition problems [117], and has been applied to face recognition as well [118], its use in this area is generally limited due to its inability to localise image features by both frequency and location. DCT has been applied to face recognition by many researchers. As the JPEG image compression standard [119] employs DCT, it is possible to extract features and perform recognition without decompressing an image. This approach is adopted in [120] who use a P2D-HMM for classification, and a recognition rate of 99.5% has been achieved. The recognition time per image was 1.5 seconds. Another approach [121] computes the DCT for the entire face images, as opposed to image blocks. Classification is performed using a Euclidean distance measure. Test results have shown [122] that the best recognition rates can

be achieved with only 64 DCT coefficients. Other DCT-based approaches have involved using SVM [123] and RBF-NN with LDA [94].

2.4.4.3 Multi-Resolution Analysis

Multi-resolution analysis provides a useful alternative that allows a signal to be decomposed by frequency, time and location. The main categories of multi-resolution analysis for image processing are Gabor filtering and either DWT or WPD, which were discussed in Section 2.3.

The high frequency wavelet coefficients are used in [124] to find face features, and the algorithm locates features based on thresholded DWT responses. In [125], the wavelet face approach is used upon decomposing a face image using different decomposition levels and applying LDA to further reduce the feature vector's dimensionality. DWT can be utilised for image normalisation as well. In [71] face images are decomposed using DWT and histogram equalisation is applied to the low-frequency coefficients to offset the effect of illumination. A two-level WPD has also been applied to face recognition. A two level WPD decomposition is performed in [41] on face images. The statistical measures are calculated, including the mean and variance of the coefficient values within the sub-image. A recognition rate of 80.5% can be achieved in this manner.

Gabor filtering has been initially applied to face recognition in [126]. A Gabor wavelet [75] is formed from the multiplication of two components, a complex sinusoidal carrier and a Gaussian envelope. In Gabor filtering, a filter bank is constructed containing Gabor filters at a variety of scales and rotational orientations. In [4], the approach is extended by using the phase information of the complex Gabor wavelet and placing the Gabor jets at facial fiducial points. In [76], Gabor filtering is combined with an LDA-based classification technique, and the Gabor feature extraction is followed by PCA and ICA. Further studies have been performed in [127; 128] with Gabor filtering for face recognition.

2.4.4.4 Other Transform-Based Techniques

A few more tools based on various transform-based techniques, such as the ridgelet [129], curvelet [130] and contourlet [77], etc. have generated plenty of

interest because of their ability to represent higher dimensional features. In some applications, their performances have also been proven superior to wavelets.

2.5 Fusion of Face and Voice

The motivation behind identification using multiple biometric modalities may be related to our natural human perception. Human perception and recognition of an object starts with using multiple senses like touch, use of both eyes, seeing and touching at the same time, or seeing or hearing somebody talking (which improves the perception quality in noisy environments).

In the context of speaker identification, the term information fusion implies utilising the combination of different sources of information, either to generate one unified data model or to reach a more informed decision. This includes fusion of multimodal data, combination of multiple experts/classifiers and multiple sensors. Many research papers were published back in the early 1980s [131]. When considering decision-making applications, there are several motivations for using information fusion:

1. Utilising complementary information (e.g. audio and video) can reduce error rates.
2. Using multiple sensors (i.e., redundancy) can increase reliability.
3. Reducing cost of implementation by using several cheap sensors rather than an expensive one.
4. Separating physically sensors to allow the acquisition of information from different points of view.

This information fusion can be carried out at the following levels:

1. Sensor-level;
2. Feature-level;
3. Score-level; and

4. Decision-level.

Although sensor-level fusion is of prime importance, the current research is focused on the feature-level, score-level and decision-level fusion of face and voice modalities.

2.5.1 Architecture of Fusion-based Systems

The basic architecture of a system based on fusion of face and voice modalities is shown in Figures 2.12 and 2.13.

Feature-level fusion occurs just before the data model is made. Features from both face and voice are either concatenated or combined through a weighted sum of features. Decision-level fusion relies on the opinions of a face recognition system and speaker identification system together forming a better informed decision for classifying the current instance. Decision-level fusion takes place using one of the following strategies:

1. Majority Voting;
2. OR; and
3. AND

2.5.2 Levels of Fusion

The fusion of different modalities is generally performed at three levels: early fusion (feature-level), late fusion (decision-level), and hybrid fusion (score-level) [132; 133; 134]. The detailed mechanisms used in these different methods along with their utilities and disadvantages are depicted below.

2.5.2.1 Feature-level Multimodal Fusion

The features of a media or data stream comprises of some discernible traits. In the early fusion (feature-level) approach, the input raw data is examined to extract its characteristic features, which is passed on to the analysis unit. For the purpose of face detection, as an example, a fusion unit may extract the motion information

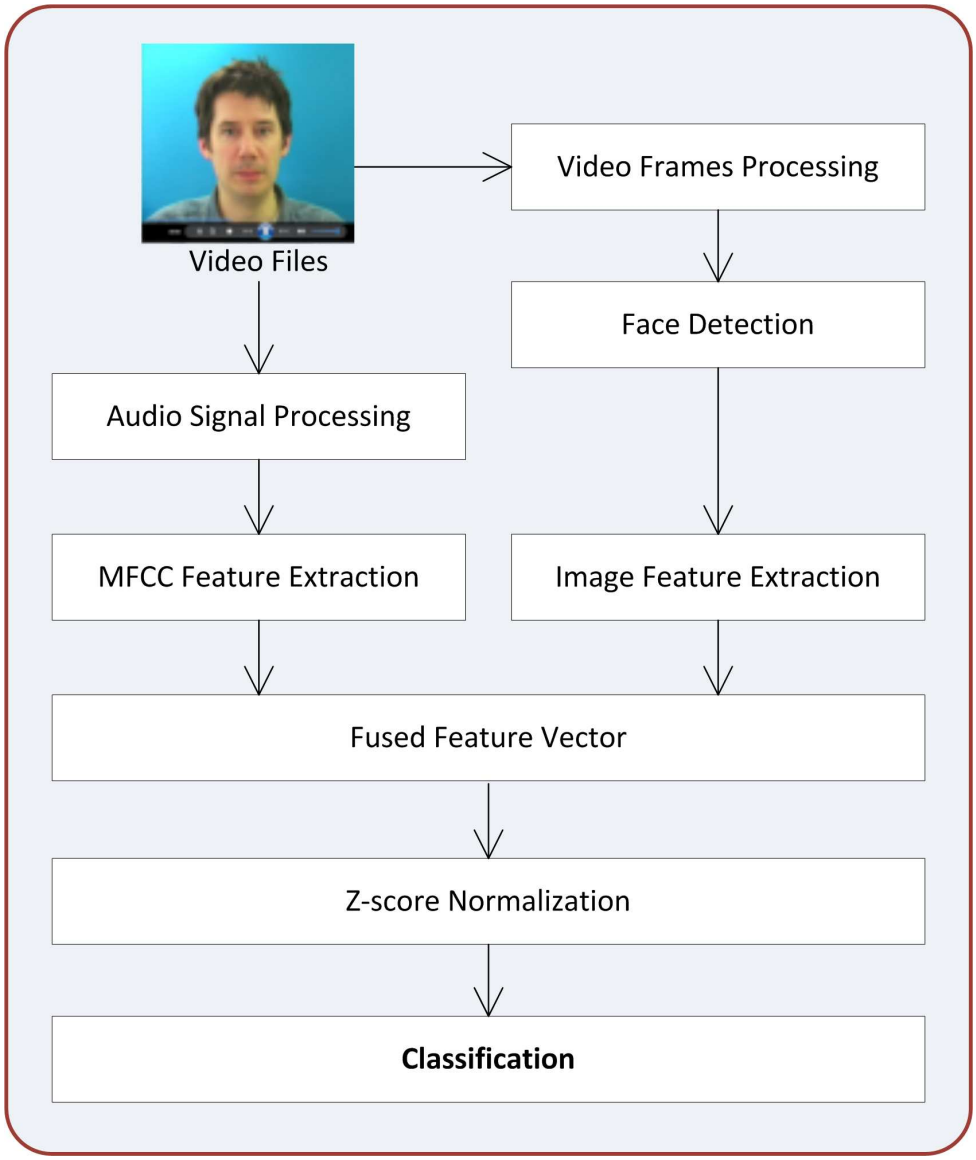

Figure 2.12: Architecture of a speaker identification system based on feature-level fusion of face and voice.

and colour, and save them into a feature vector for further processing. There can be a large number of modalities that outputs different types of features [135; 136]. These features may include but are not limited to the following:

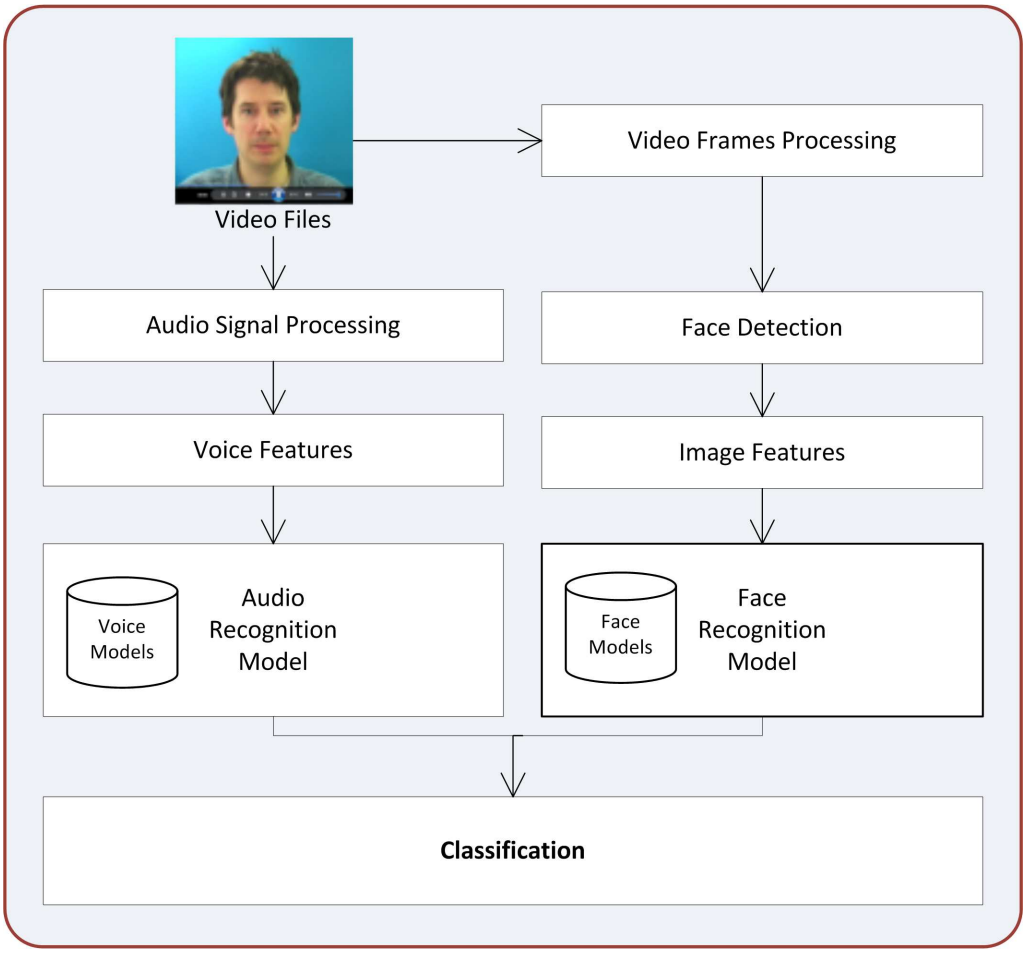

Figure 2.13: Architecture of a speaker identification system based on decision-level fusion of face and voice.

1. Audio features: The audio signal is one of the most important and reliable source of unique features specific to a user. Many different schemes like FFT, MFCC, Zero Crossing Rate (ZCR), or LPC, etc. are employed in retrieving the audio features.

2. Visual features: Specific traits like colour and texture, as well as size and shape can be extracted from an input image. Instead of an entire image, sometimes partial blocks are used, or feature points are detected automat-

ically.

3. Text features: A speech recogniser or optical character reader can be used to obtain the relevant textual features from the metadata.

4. Motion features: The pixel variations in a video stream can be measured using optical flows and magnitude histogram, etc. to find the direction and pattern of motion, and transformed to a representative kinetic energy.

5. Metadata: The supplementary information (e.g. the, source, location, duration, or time stamp of a video or image) contained in the metadata (recorded during production) can often provide useful information in facilitating the quantification and calibration of different audio-visual features.

The biggest and most unique advantage of feature-level fusion is its capability to utilise any correlation that may exist among features extracted from dissimilar modalities. This can be done at an early stage when all the raw data are available. Furthermore, since a combined feature vector is created, a single learning phase is sufficient [134]. But modalities that are coupled closely are often obtained at different times and are subject to some temporal variations. Characterising the proper time synchronisation and compensating for its effects is necessary for early fusion, which often turns out to be a demanding task. Also, when the number of modalities start going high, it is very hard to estimate and keep track of the cross-correlation that may exist among the heterogeneous traits. Nevertheless, depending on the application at hand, the above mentioned pros of early fusion can substantially outweigh its cons and prove as a handy multimedia analysis tool. Many researchers, like the authors of [137], therefore, tend to lean towards the early fusion approach for audio-visual speech recognition. Chetty et al. proposed a speaker verification system based on audio-visual hybrid fusion from a set of features that are cross modal [138]. For a personnel authentication system based on face and voice, they also developed a feature-level fusion [139] to check the liveness, and presented test results performed on the VidTIMIT and UCBN databases.

2.5.2.2 Decision-level Multimodal Fusion

When there are multiple modules present in a biometric system, each of them process their own feature sets to provide a local decision. In the fusion schemes that work at the decision-level, a decision fusion unit takes these local decisions, integrates then into a fused decision vector, and analyzes it at the semantic level to conclude a final decision about the hypothesis. Decision-level fusion has certain advantages compared to feature fusion. It can overcome many of the inadequacies of feature-level fusion. Earlier we discussed the problems of incompatibility of features and dimensionality that early fusion faces. Since decision-level fusion deals only with the local decisions obtained from the individual multimodal units, regardless of how different their respective feature representations are, the fusion becomes much more straightforward than feature-level. For the same reason, a large number of multimodal units can be easily managed with this strategy and makes the system highly scalable. This type of scalability is very hard to maintain in early fusion schemes [140]. In addition, since compatibility and dimensionality do not pose any restrictions on the local analysis methods, each of these units have more flexibility in selecting suitable feature extraction and analysis techniques.

However, the major drawback of late fusion methods is their inability to make any use of the correlation that exists among the different modalities in feature-level. The learning process is also cumbersome and lengthy since a wide variety of classifiers are locally implemented. Nevertheless, many biometric systems have been designed based on decision-level fusion because of its scalability and the other merits mentioned above. For instance, a decision-level fusion using a linear weighted sum was implemented in [141] for a face detection module in conjunction with a speech recognisation module.

2.5.2.3 Score-level Multimodal Fusion

As laid out in the previous two sections, the feature and decision-level fusions both have some unique advantages and drawbacks. Some researchers have naturally searched for a middle ground so that the pros of these strategies can be combined while keeping the disadvantages under tolerable limits. This has led to the advent of hybrid fusion strategies. Both feature and decision-level techniques are

combined in an efficient manner with this goal. Similar to feature-level schemes, the hybrid fusion scheme also analyzes the feature vectors from all units and concludes its decisions. The individual features are also analyzed independently to yield autonomous decisions which are fused similar to late fusion schemes. Finally, a score-level fusion approach is taken to further process and merge decisions obtained at different stages to acquire the ultimate decision. Many multimedia related problems have been addresses and solved [142; 143] effectively using such techniques.

2.5.3 Methods for Multimodal Fusion

A detailed overview of the existing fusion methods of all classes are presented in this section. Fusion schemes can broadly be categorised into rule-based, classification-based and estimation-based schemes. Such classifications are results of the inherent characteristics of these schemes with respect to their problem space.

2.5.3.1 Rule-based Fusion Methods

There is a wide diversity of fundamental rules that can combine the multimodal data in biometric systems, such as sum and product, majority voting, AND, OR, MAX, MIN, etc. The basic building blocks and operations of these rules can be found in [144]. A plenty of other custom-defined rules have also been established for particular applications. When the different modalities involved are well aligned temporally, these systems can yield acceptable levels of performance. Some of the existing rule-based methods are outlined below.

In the linear weighted fusion scheme, one of the most popular and easy to implement methods, the data acquired from the different modules of dissimilar modalities are linearly integrated either using semantic decisions [145] or low-level features [146]. If the weights for the modalities are normalised and then combined, the performance can be fine-tuned as needed. If it is possible to compute the matching scores of the modalities, the weights are normalised using decimal scaling, minmax or z-score techniques. However, since these methods can sometimes have different degrees of sensitivity to outliers, people have attempted

linear feature-level fusion strategies for multimedia investigations [146; 147; 148]. Many researchers have also used decision-level fusion strategy for speaker and speech recognition, or for the detection of monologues, etc [149].

Majority voting can be thought of as a unique case of weighted combination, considering that all the weights are equal. If most of the classifiers deduce the same decision, or decisions that are very close, then it is accepted to be the final decision. The speaker identification system in [150] treated the raw samples of speech data as features, and identified some patterns to recognise the speaker. This type of pattern usually consists of utterances of vowels. A novel decision-level fusion scheme, based on production rules, was proposed in [151] to combine speech with the inputs from a pen. Under this scheme, the processing state of a certain recogniser is tracked by applying some synchronisation rules. The input occurrences that will be integrated are then identified using event interpretation techniques. If the recognisers are unable to yield any useful result, unimodal interpretations are then used.

2.5.3.2 Classification-based Fusion Methods

Classification-based fusion methods are comprised of various classification techniques aimed at categorising multimodal traits into some classes that are predefined. These schemes include neural networks, Bayesian and dynamic Bayesian networks, the maximum entropy model and SVM, etc. For data classification purposes, especially in multimedia fields, SVMs are widely popular for face detection, modality fusion, feature and text categorisation, etc. SVM is a type of supervised learning method which partitions sets of data vectors to attribute them to different learned classes. It is used to solve many problems on pattern classification, and has also been extended to non-linear classifiers by means of the kernel concept.

The authors of [152] developed a hybrid fusion approach combining early fusion upon normalisation and late fusion with semantic indexing. The entries of the concatenated vectors were fused after the normalisation. In this work, a kernel-based fusion scheme was built upon SVMs, and the modalities decided the choice of the kernel functions. Bayesian inference methods have also been effectively implemented in multimodal fusion at different levels. It has been used

in the feature-level in [153], in the decision-level in [154] and at a hybrid level in [155] for multimedia surveillance.

2.5.3.3 Estimation-based Fusion Methods

The class of estimation-based fusion involves schemes that rely on Kalman, extended Kalman and particle filters, etc. The position and direction of moving objects can be adequately estimated using these filtering methods, especially for multimodal data. If an object needs to be tracked, its instantaneous locations can be estimated from compound modalities comprising audio and video data that are fused appropriately. The Kalman filter (KF) [156], although developed about half a century ago, still remains as one of the most prolific and widely deployed data fusion algorithms available. Often referred to as the linear quadratic estimator, it can perform well under the presence of noise and many other obstacles by making a series of temporal measurements. In real-time, it can process low-level fusion data and extract significant state estimates of a system. The feature-level fusion scheme developed in [157] attempted to estimate a speaker's translational motion (velocity, acceleration, etc.) from different audio-visual features. In addition to the authors of [157], many other researchers have used KF for tracking objects or localising sources. The extended Kalman filter, a non-linear version of the KF, has been proven to perform even better than KF when the systems are non-linear. The particle filter method has also shown good robustness for non-Gaussian and non-linear models.

2.5.4 Acoustic Visual Speaker Models

Multimodal systems can be divided into classifiers where the different features are fused at the feature-level by concatenating both feature vectors or by the combination of both modalities [158]. The choice of the fusion method is mainly dependent on the assumption of conditional independence of the two modalities and of the availability of synchronised features. If conditional independence is assumed, two separate models might be constructed. This might also be necessary if different frame rates would complicate the feature fusion at the frame level. Fusion at the feature-level is the more general case which avoids making assump-

tions about conditional independence. Composite feature models assume that both modalities are conditionally dependent. The motivation and drawbacks for this integration method are similar to those for acoustic visual speech recognition. For this method, composite feature vectors are constructed by concatenating both feature vectors at the frame level. The training and recognition procedures are performed the same way as for visual features. Acoustic features are represented as MFCC coefficients which are extracted at the same frame rate as the visual features to facilitate their combination. The composite feature vectors are used for both the HMM and the GMM speaker models [158].

Parallel feature model is another important class of modelling. A different method for audio-visual speaker modelling is proposed in [159] based on separate models for visual and acoustic features. The motivations behind this approach are as follows. The reliability of the information of both modalities is different and should therefore be weighted accordingly. The quasi-stationary events of the acoustic and visual modalities are different and should therefore be modelled by individual HMMs with different topologies. Parallel models facilitate the use of different sampling rates for different modalities. A detailed description of the system can be found in [158; 159]. The acoustic training and test sequences are segmented using forced alignment based on a word sequence that is known. The acoustic segmentation then segments the visual speech sequences. Separate HMMs are trained on the acoustic and the visual feature vectors. The models can therefore have different HMM structures. For example, since the frame rate of the visual signal is often smaller than that for the acoustic signal, a smaller number of states might be chosen for the visual models. The acoustic and visual streams of an acoustic visual model are constrained to be time synchronous at the beginning and end of the model but can be asynchronous within the model. The likelihood of each modality is estimated for the whole test phrase and acoustic visual classification is performed on the weighted sum of these likelihoods. The weights for the modalities were estimated on a separate validation set.

2.5.5 Additional Existing Approaches

In this section, we briefly discuss some additional existing approaches such as mosaic transform with score-level and feature-level fusion, coupled HMM, NN, SVM and EM decision-level fusion, coupled HMM, PCA and GMM score-level fusion, etc. and provide a qualitative comparison between the schemes.

2.5.5.1 Mosaic Transform with Score-level Fusion

Combining multiple modalities like face and voice to enhance speaker identification accuracy has been observed in [160]. In that work, the mosaic transform was used to perform fusion at the score-level for both face and voice. Using the subspace method algorithm, Ariki et al. [160] were able to improve speaker identification results by 15% over the conventional Class-featuring Information Compression (CLAFIC) method.

2.5.5.2 Mosaic Transform with Feature-level Fusion

Later in the decade, the mainstream approaches shifted towards realising speaker identification by combining face and voice through principal component analysis-based approaches. One of these approaches used eigenvoice and eigenface together with Maximum Likelihood Eigen Decomposition (MLED) [161] on isolated sets of words and argued that combining face and voice to cater for speaker variability improves results and that speaker identification researchers should emphasise the use of multimodal identification to improve results.

2.5.5.3 Coupled HMM, NN, SVM and EM Decision-level Fusion

Coupled Hidden Markov Models (HMM) were studied [162] for audio-visual speaker identification with decision-level fusion. MFCC was used for audio features and mouth region features were extracted using a cascaded approach with Neural Network and an SVM. Expectation Maximisation (EM) was used to train a speaker-independent model and a MAP (Maximum a posteriori) hypothesis-based algorithm for a speaker-dependent model coupled with HMM was used. Test results on VidTIMIT [11] were published for different SNR and an overall 95% error rate reduction was reported in the audio-only portion of the system.

2.5.5.4 GMM and KFD Score-level Fusion

GMM applied to voice- and score-level fusion of face and voice were further studied in [163]. In this research, the Kernel Fisherface Discriminants (KFD) were applied for face, GMM for voice and an SVM was used to fuse face and voice at the score-level. Average response time of the presented system was 2.1 seconds per sample. For use of only the voice modality, the EER for the two test sessions were 11.06% and 10.29%, respectively. The fusion of voice and face modalities has achieved 95% and 86% EER reductions compared to speaker verification only. If only the face modality is used, the EER of the two sessions are 9.42% and 16.05%, and the EER reductions are 95% and 91% for fusion. These results were reported on a local corpus.

2.5.5.5 PCA and GMM Score-level Fusion

In recent years, speaker identification researchers experimented with the combination of various behavioural and physical biometrics to improve accuracy. In experiments in [164] [165], voice, face and retina (iris) were fused at the score-level and the results were reported on a local database. PCA was used for face, GMM was applied on voice, and multi-scale edge matching was applied on the iris. The system was convenient for the user if one of the three individual modules accepted the legitimate user.

Another score-level fusion approach using Fisherface (for face) and MFCC with GMM (for voice) proved successful [166]. Scores were normalised using a sigmoid function. When the distance between the camera and the object (a person) is one metre, the performance of the fusion technique is almost same as the Fisherface technique (99.5%). When the distance is 3 metres, the Fusion technique is almost 50% more efficient than Fisherface.

2.5.5.6 PCA, MFCC, VQ and Subspace Method Score-level Fusion

Voice and face were also combined quite recently with signature recognition [167]. PCA and LCA were used for feature extraction for face, MFCC for voice and DCT for features from signature. The subspace method (for face), VQ (for voice) and nearest neighbour algorithm were applied for signature classification. Experimen-

tal results indicate the efficacy of multimodal systems even when the biometric data are affected by noise. In this case, the performance of the uni-modal system is around 60% and yet the multimodal system performance is still around 95%.

2.6 Summary

In Table 2.1, a summary is provided for today's existing feature extraction and modelling techniques as discussed in this chapter. A detail comparison and key features of the existing audio-visual systems (based on fusion of face and voice) are highlighted in Table 2.2.

Table 2.1: Existing feature extraction and modelling techniques.

Source	System	Algorithm	Identification	Subjects
[48]	Robust Speaker Identification using GMMs	MFCC, GMM	96.8%	49 Speakers
[52]	Text-independent Speaker Identification and Speaker-independent Speech Recognition using an Iterative Clustering Approach	PLP, MF-PLP	91%	50 Speakers
[59]	Improved Text-independent Speaker Identification using Fused MFCC and IMFCC Feature Sets based on Gaussian Filters	MFCC, IMFCC, Gaussian Filter	97.42%	131 Speakers
[60]	Signal-to-signal Ratio Independent Speaker Identification for Co-channel Speech Signals	KL-divergence	97%	34 Speakers
[61]	A Text-independent Speaker Identification System using a Novel Parametric Neural Network	Parametric NN	94%	40 Speakers

Table 2.2: Survey of audio-visual systems currently available.

Source	Feature Extraction	Classifier	Fusion	Corpus	Database Size	Performance
[131]	Voice: Mosaic Face: Mosaic Transform to remove changes caused by camera angle and lighting	Subspace Method	Score-Level	Local	Dataset: Face images of 30 subjects. Training dataset: 8. Size of the image: 82x102 with 8-bit grey values.	Face recognition result improved 15% over conventional CLAFIC. speaker identification result improved 16% over conventional CLAFIC. Integration of face and speaker: Multiple Feature Subspace-100% for 22 dimensions in face recognition and 6 dimensions in speaker identification. Weighted Project method - Almost same result as the multiple feature subspace
[160]	Voice: Eigen voice Face: Eigen face	Max. likelihood Eigen decomposition		Isolate	Database [11]: 5 set of 30 speakers (total 150). Training Data size: 120. Test data size:30.	10-span ($k = 10$) Eigen voice $\Rightarrow$ MAP (Maximum A Posteriori) estimation provides the best result

[Continued to next page]

Source	Feature Extraction	Classifier	Fusion	Corpus	Database Size	Performance
[24]	Voice: LPCC Coefficient for text independent speaker verification Face: Gabor Jets: Elastic graph matching using the Gabor filter responses	SVM, MLP on c4.5 decision tree, Fisher linear discriminant and Bayesian classifier for fusion	Matching Score	XM2VTS	4 recordings of 295 subjects	Failure Acceptance (FA) and Failure Rejection (FR) rates promising for both SVM and Bayesian Fusion (1.07, .25 & 1.21, 0.0), respectively.
[161]	Voice: MFCC Face: Visual features obtained from mouth region through cascading algorithm. The feature extraction uses neural network followed by SVM classifier.	Expectation-Maximisation used to train a speaker-independent model and MAP used to train speaker-dependent model; coupled hidden Markov models	Decision-Level	XM2VTS	Four training sequence for each of the 87 speakers for training. Testing: 320 sequences.	Test results were published for speaker identification for various SNR for different values of visual stream exponent. The error rate of the audio-only system at SNR = 0 db is reduced by over 95%

[Continued to next page]

Source	Feature Extraction	Classifier	Fusion	Corpus	Database Size	Performance
[168]	Templates methods of classifier fusion in the context of multimodal personal identify verification.	Baseline classifier	Fusion rules (sum and vote)	XM2VTS	295 subjects. 200 clients, 25 evaluation imposter and 70 test imposter	Trainable fusion strategy do not offer better performance
[162]	Face: Cascade-boosting method to find a rough face region at first and eye region is localised using EigenEye technique and facial image is cropped. Discrete wavelet transform is extract lower-dimensional facial features.	Kernel Fisherface discriminant (KFD) for face. Log-likelihood ratio (LLR) for voice. Support Vector Machine used to fuse voice and face.	Score Level	Local	Dataset: 17 subjects for experiment. 13 subjects for training data set and the remaining for test.	Average response time: 2.1 seconds per sample For only voice modality use, the equal error rate for the two test sessions are 11.06% and 10.29%, respectively. The fusion of voice and face modalities has achieved 95% and 86% equal error rate reductions compared to speaker verification only. If the face modality is used only, the equal error

[Continued to next page]

Source	Feature Extraction	Classifier	Fusion	Corpus	Database Size	Performance
	Voice: Model selection based self-splitting Gaussian mixture learning algorithm					rates of the two sessions are 9.42% and 16.05%, and the equal error rate reductions are 95% and 91% for fusion.
[163]	Voice: Gaussian Mixture model Face: PCA	Eigenface Log-likelihood ratio Multi-scale edging matching for Iris	Matching Score	Local	Dataset: 19 users (14 males & 5 females). 20 triplets of samples (2 sessions and 10 triplets for each session).	Convenient for the user if one of the individual modules accepts the legitimate user.
[165]	Voice: MFCC Face: Fisherface	Fisherface for face GMM for speaker identification	Fusion Technique: Scoring. Score normalised using sigmoid function.	Local		When the distance between the camera and the object (a person) is 1 metre, the performance of fusion technique is almost same as the Fisherface (99.5%). When the distance is 3 metres, the Fusion technique is almost 50% more efficient than Fisherface.

[Continued to next page]

Source	Feature Extraction	Classifier	Fusion	Corpus	Database Size	Performance
[166]	Voice: MFCC Face: PCA, LCA Signature: Discrete Cosine Transform	Subspace Method for face Vector quantization (VQ) for speech Nearest neighbour for the signature fuse voice and face. Min-max normalisation for score normalisation and same rule for biometric data fusion.	Score Level	Local	Image: 1600 images for 10 individuals. Training data set: 40 face image per person Testing data set: 40 face image per person Voice: 1600 speech for 10 individuals. Training data set: 40 speech per person Testing data set: 40 speech per person Dataset: 24 face images each for 30 users. Training data: 16 images each for 30 users. Testing data: 8 images each for 30 users.	Experiment results indicate the efficacy of multimodal system even when the biometric data are affected by noise. When the biometric data are affected by noise, the performance of the uni-modal system is around 60% and still the multimodal system performance is around 95%

Chapter 3

Speaker Identification Using DWT & Multimodal Neural Networks

3.1 Overview

Chapter 3 is devoted entirely to describing a robust model for real-time text-independent speaker identification based on wavelet analysis and bootstrap aggregating (bagging) neural networks.

One of the prime objectives of this research on speaker identification is to improve upon the existing techniques by applying new methods for various sub-problems. After conducting the literature review in Chapter 2, it is necessary to experiment with the most recent approaches and attempt to improve them. There were two main areas for apparent improvement at this stage: feature extraction and classification.

This section enumerates the reasons behind the choice of speaker identification technology and gives a brief overview.

Neither PNN, GRNN, RFB-NN nor wavelet analysis was new to the world of speaker identification at the time this research began. Nevertheless, experimenting with the recent approaches seemed the only way forward to identify their shortcomings and to highlight the improvements we made.

Artificial Neural Networks (ANNs) are powerful classifiers that have been successfully applied to a number of pattern recognition problems. Their dynamic

and adaptive nature attracts researchers from a multitude of domains. Speaker identification has also benefitted from ANN technique because of its ability to handle changing patterns in the training set. In addition, it accordingly adjusts its weight matrix to learn the contribution of each individual feature in the final classification of an instance.

ANNs are designed based on a set of neurons that are connected. There can be multiple layers each containing different numbers of neurons. Each node or perceptron has the capability to assign weights to its contributing factors and produce a partial decision. The final layer is dependent upon the output of the previous layers and the corresponding weight matrices. Back propagation is a classic example of penalising contributing layer nodes/perceptrons by reducing their weight whenever the final classification deviates from the required output. This lets the network propagate the error penalty along the contributing nodes starting from the output nodes to the input layer nodes, successively penalising each layer and the inner neurons. This is what happens in the Back Propagation Neural Network (BPNN), the most popular and commonly used ANN.

A BPNN lets us recognise complex patterns and supports any number of training epochs to produce a learnt classifier for the unseen data. But this procedure has a cost associated with how much learning is required to perform classification within a predictable accuracy on the unseen data. Learning, on the other hand, not only requires computer resources for computation but also hinders system performance. BPNN not only requires long training time but also a huge number of instances/patterns to become fully trained for classification of the unseen data.

There are other types of ANNs which are more robust and therefore do not require huge training time. Examples are PNN, RBF-NN and GRNN. This class of neural networks does not require a huge number of training epochs and feedback loops as they are instantly formed and trained on the basis of the training data.

In order to accomplish the objectives of this doctoral research we started looking at the possible flaws in the existing solutions to speaker identification and address them. Certainly, it is not the case that we have an indefinite set of points to address; however, in general, we can improve upon existing systems in the following areas:

A . Input signal pre-processing (noise reduction, silence removal, etc.);

B . Feature extraction methodology (MFCC or any improved method); and

C . Choice of classifier and strategy of classification (machine learning domain).

However, our goal in this thesis was not to address the improvement of pre-processing algorithms. As a matter of fact, we use readily built databases for which the pre-processing has already been done. The pre-processing tasks are relatively straight-forward and does not relate to the identification algorithms for a certain speaker. Nevertheless, data pre-processing is the first step performed to extract meaningful features for further processing. For example, the raw audio signal is an almost meaningless stream of numbers, and the video frames are merely sets of red, green and blue pixel values. Each of these streams is further processed separately, as explained below.

An audio stream consists of thousands of values in the range [-1, 1] that are sampled at regular intervals. An 8 kHz sampling rate means that 8000 such values vary each second when a speaker's audio is recorded. There are a large number of values left after trimming the silent sections at the start and end of the audio, which do not contain useful information. These raw values only tell us about the amplitude variations in the speech and do not convey any explicit information about the speaker. Because we are using text-independent speaker recognition, we must extract distinguishing speech features that describe a speaker's orientation or, more specifically, the qualities of the speaker's glottal tract, which are independent of the language being used. Therefore, if the same speaker speaks a different set of words next time, our system should recognise the speaker. Therefore, we transform the raw signal into a parametric representation as detailed in Chapter 2. Throughout this thesis, the pre-processing task (as well as the "Pre-processing" block in all the illustrative images) refer to established techniques which we did not attempt to (or even need to) modify.

In developing this Multimodal Neural Network (MNN) speaker identification system, our first aim was to utilise the newer approaches like wavelet analysis, in the feature extraction phase that were made popular in recent years. MFCC has been suggested as the most popular feature extraction methodology by many researchers working on speech recognition systems. However, if using this scheme for speaker identification, it has its own limitations since it was originally bor-

rowed from speech recognition. It has the "curse of dimensionality" in the sense that the feature matrix for a second speech signal consists of 500 x 20 values on average.

This high number of feature values for a single signal places huge pressure on the system accuracy and performance when dealing with 30 users and 100 files (speech signals) for each user during system training. It not only slows down the system but also adds to its response time and a huge demand on memory consumption. Although memory is quite cheap now-a-days, computation time is one thing speaker identification systems cannot afford to compromise with, since it is very crucial to keep it low.

Secondly, instead of the recently popular back propagation-based methods, we were attracted to the more adaptive instantly trained specific class of neural networks (PNN/GRNN/RBF-NN) to improve both the identification time and classification accuracy of the speaker identification system.

Subsequent experimentation resulted in newer observations for the wavelet analysis combined with one instantly trained ANN. These sets of new observations contributed to the utilisation of multiple combined ANN instances of the same class selected through the majority voting scheme. This increased the robustness of the overall system and increased the classification accuracy as well. This scheme of combining multiple classifiers to search for a better hypothesis in the decision space is called bootstrap aggregating or bagging. Bagging is a powerful machine learning strategy to enhance the hypothesis search and convergence of the overall system to a global error minima.

In this chapter, we describe the design and implementation of a text-independent multimodal speaker identification system based on wavelet analysis and neural networks developed during this doctoral research. Wavelet analysis comprises DWT, WPT, WSBC and MFCC. The learning module comprises a GRNN, PNN and RBF-NN, forming decisions through a majority voting scheme. This system has been fully tested on the GRID [2] corpus for validity of the proposed approach and comprehensive results are reported. The system has been compared with BPNN, GMM and PCA-based approaches to speaker identification. The suggested scheme of feature extraction and classification improved the identification rate by 15% compared to the classical MFCC and reduced the identification

time by 40% compared to BPNN, GMM and PCA based approaches.

In Section 3.2, we present an overview of the wavelet analysis algorithms we studied, implemented and experimented with during construction of this system. In Section 3.3, we explain the motivation behind the proposed system while introducing the set of algorithms like PNN, GRN, RBF-NN and wavelet feature extraction strategies, and highlight their architectural aspects. After the prerequisites of wavelet analysis and ANNs are reviewed, we introduce in Section 3.4, the proposed system while showing the various architectural aspects of the overall system through various diagrams. Subsequently, in Section 3.5, we explain the comprehensive evaluation and testing performed while documenting the results and discussions on the proposed system. We also compare the system performance with other relevant systems. Finally, we sum up in Section 3.6 the description of the system and its limitations and suggest new directions for further research.

3.2 Overview of Wavelet Analysis Techniques

Wavelet transforms [169; 170; 171] have been studied comprehensively in the recent times and widely utilised in various areas of science and engineering. Under the class of wavelet analysis, a mother wavelet is processed upon dilation and translation. Many signals of interest can be represented with wavelet decompositions in general. If only the respective wavelet coefficients are adjusted, proper signal processing algorithms can be effectively implemented. A wavelet can have arbitrary real values for its scale parameter [172], or have discrete values belonging to a lattice [173; 174].

Wavelet and WP analysis have been proven as effectual signal processing techniques for a variety of digital signal processing problems. Wavelets have been used in two different methods in feature extraction plans designed for the task of speech/voice identification. In the first method [175], DWT replaces DCT throughout the duration of feature extraction. In the second method, wavelet transform is used directly on the speech/voice signals and either wavelet coefficients containing high energy are extracted as features [176] but suffer from shift variance, or sub-band energies are used instead of the Mel filter-bank sub band

energies as proposed in [177].

A feature extraction scheme derived from the wavelet eigenfunction was proposed in [178]. In [179], a text-independent speaker identification system was proposed based upon an improved wavelet transform, which relies on the kernel canonical correlation analysis to learn the correlation of the expression vector and the wavelet transform. WPT, which is analogous to DWT in some ways, obtains the speech signal using a recursive binary tree and performs a form of recursive decomposition. Instead of performing decomposition only on approximations, it decomposes the details as well. WPT, therefore, has a better feature representation than DWT [178]. The performance enhancement of a speaker identification system requires a careful selection of suitable features from the raw set available, which is usually somewhat redundant [180; 181]. The most relevant and significant information must be chosen from the original feature space using an appropriate feature selection scheme. The perceptual decomposition tree for wavelet packets and the energy indices of WPT for speaker identification were introduced in [27] and [19], respectively. The authors of [182; 183] obtained terminal node signals from DWT and computed the sure entropy for the waveforms.

In this phase, the rough audio signal is pre-processed to extract only the distinguishing features for analysis from the entire signal. The feature extraction techniques used are DWT, WPT, WSBC and MFCC. Chapter 2 reviewed the basic ingredients of wavelet analysis techniques, including DWT, WPT, WSBC and Irregular Decomposition. More details about the wavelet analysis methodology is available in Chapter 2 .

A speech signal contains a huge amount of data. For example, a one-second speech signal consists of approximately 50,000 floating-point values in a single linear vector.

The feature-extraction block of this system consists of the following algorithms: DWT, WPT, SBC and Irregular Decomposition. All feature vectors are linear vectors of length less than or equal to 64, as summarised in Table 3.1.

During the scope of this research, experimentation and testing were performed with all of these approaches, selecting one at a time. In Section 3.3, we present the rich new class of instantly trainable ANNs used during the course of this

Table 3.1: Summary of feature extraction vectors for wavelet analysis.

Input	Feature Extraction Scheme	Output Vector Length
1 second long audio signal (GRID) recorded at 44.1 kHz	Discrete Wavelet Transform	8
	Wavelet Packet Transform	64
	WPT in Mel-scale (WSBC)	6
	Irregular Decomposition	57
	MFCC	20×450

doctoral research, focusing on PNN, GRNN and RBF-ANN.

3.3 Overview of Neural Networks for Speaker Identification

Neural networks are the most common approach to learning non-linear or complex training spaces. NNs are vastly applied in numerous data analysis and speaker identification schemes, and have not been left out from classification tasks either [57; 184]. For an ANN, there is no need to predict the transfer function between the input and output ahead of time, and this is one of its greatest advantages. In addition to the discussion on neural networks provided in Section 2.3.2.4, below we provide a more comprehensive description of different neural networks in context of our proposed speaker identification scheme.

3.3.1 RBF Networks

An RBF neural network is an ANN that uses radial basis functions as activation functions. The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters. These networks have many uses, such as time series prediction, classification, and system control. In the context of speaker identification, the RBF neural network utilises the projection of an eigenface space to compute the neural network input features.

There are two main categories of learning: the supervised learning and the

unsupervised learning. The RBF neural network [57] has both a supervised and unsupervised component to its learning. It consists of three layers of neurons: input, hidden and output. The hidden layer neurons represent a series of centres in the input data space, as shown in Figure 3.1 Each of these centres has a typical Gaussian activation function. The activation depends on the distance between the presented input vector and the centre. The further the vector is from the centre, the lower is the activation and vice versa. The generation of the centres and their widths is done using an unsupervised k-means clustering algorithm. The centres and widths created by this algorithm would then form the weights and biases of the hidden and constant layer. The output layer (which has non-linear activations) is trained by back-propagation.

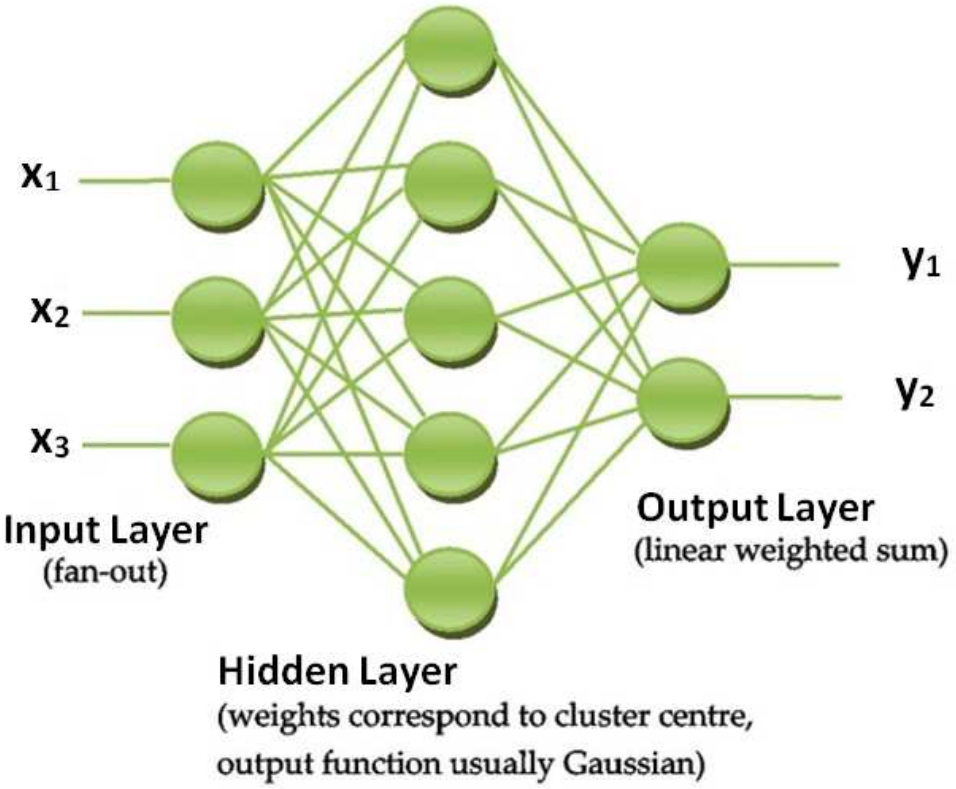

Figure 3.1: Structure of an RBF neural network [6].

RBF neural networks employ radial basis functions directly on each input value without associating a weight line from input to this radial basis function layer, as shown in Figure 3.1. The radial basis function is given by,

$$y_m = f_m(\mathbf{X}) = exp\left[-\frac{|\mathbf{X} - c_m|^2}{2\sigma^2}\right], \tag{3.1}$$

where $|\mathbf{X} - c_m|^2$ is the square of the distance between c_m and the input feature vector $\mathbf{X}$. The network output is a weighted sum of the radial basis functions from the nodes, which is calculated as,

$$z_j = \frac{1}{M}\sum_{m=1}^{M} u_{m,j} y_m. \tag{3.2}$$

3.3.2 PNN Networks

PNN has an input layer where the input vectors, in our case the audio feature vectors, are fed. It also includes one or more hidden layers with multiple neurons connected through weighted paths. Additionally, it includes one or more output neurons depending on the number of different classes to identify through classification. PNN is a statistical classifier network that applies a maximum a posteriori (MAP) hypothesis to classify a test pattern X into Class C if:

$$P(X_i|C_i)P(C_i) \quad \geq \quad P(X_i|C_j)P(C_j) \qquad \forall j, \tag{3.3}$$

where $P(C_i)$ is the prior probability of Speaker i determined from the training feature vectors. $P(X_i|C_i)$ is the conditional probability that this pattern is generated from the class C_i assuming that the training data follows a PDF. Such a PDF is to be estimated for each speaker class. PNN uses Parzen window estimation with a Gaussian windowing function as the PDF estimator. As shown in Figure 3.2, the third layer from the left is the class layer and the last layer is the output, which represents the winning class.

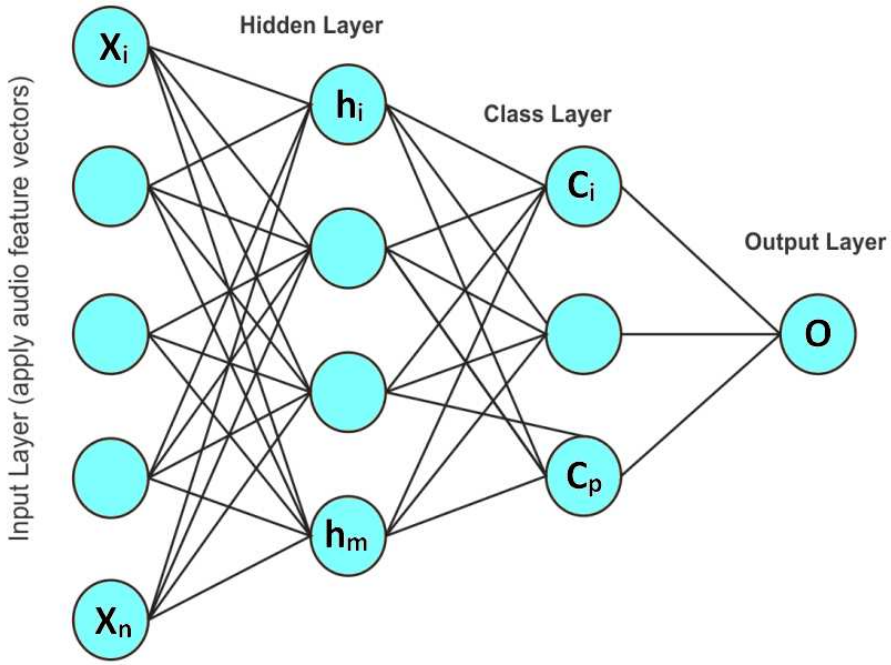

Figure 3.2: Architecture of PNN [6].

3.3.3 GRNN Networks

General regressive neural networks (GRNN), first proposed in 1991 [57], are widely used in many identification tasks. Figure 3.3 shows the block diagram of the GRNN architecture. It is a one-passing learning algorithm. Continuous variables like transient contents of speech signals can be successfully estimated using this network. GRNN has a structure similar to PNN and RBF networks but are based on general regression as proposed by [7]. While architectures like BPNN usually require iterative training methods to achieve convergence towards a preferred solution, GRNN does not need that. It only requires a fraction of the samples that BPNN uses. In contrast to PNNs and RBF-NNs, a GRNN uses a PDF based on a Normal Distribution.

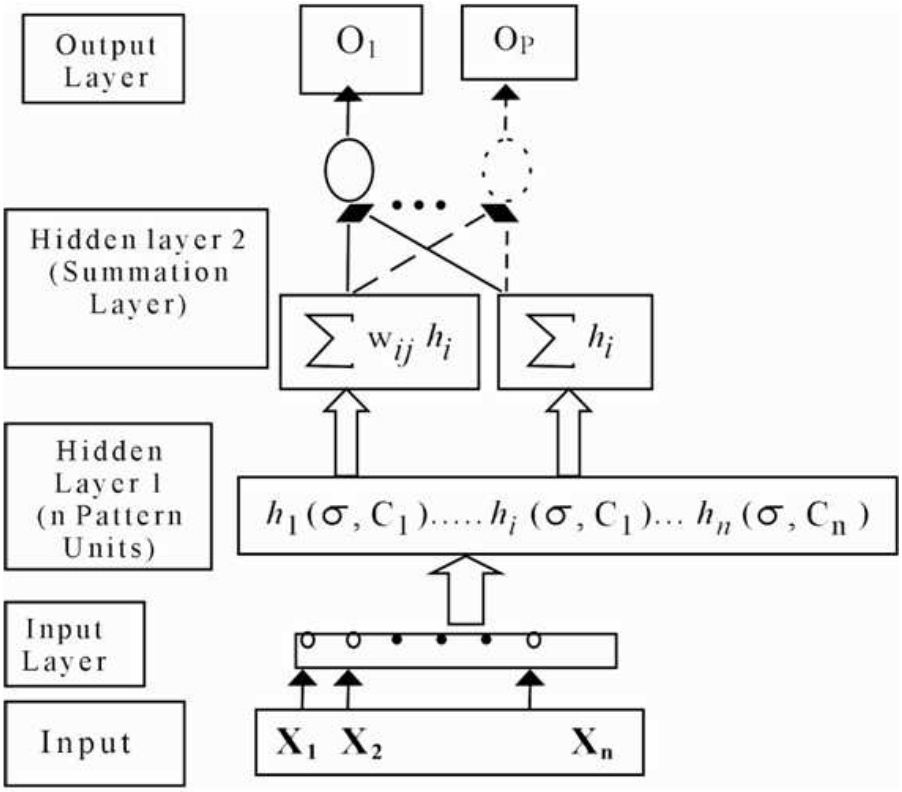

Figure 3.3: GRNN architecture [7] [8].

3.3.4 Comparison of Different Neural Networks

BPNNs, RBFs, GRNNs and PNNs can be easily differentiated from each other on the basis of structure, training strategy, samples requirement, training time, accuracy and suitability for various types of data. Figure 3.4 shows the general structure of a BPNN.

Training Samples: PNNs, RBFs and GRNNs require just a fraction of the samples that are normally required for a BPNN.

Training Time: PNNs, RBFs and GRNNs require no time as they are instantly trained upon initialisation compared to a BPNN which requires thousands

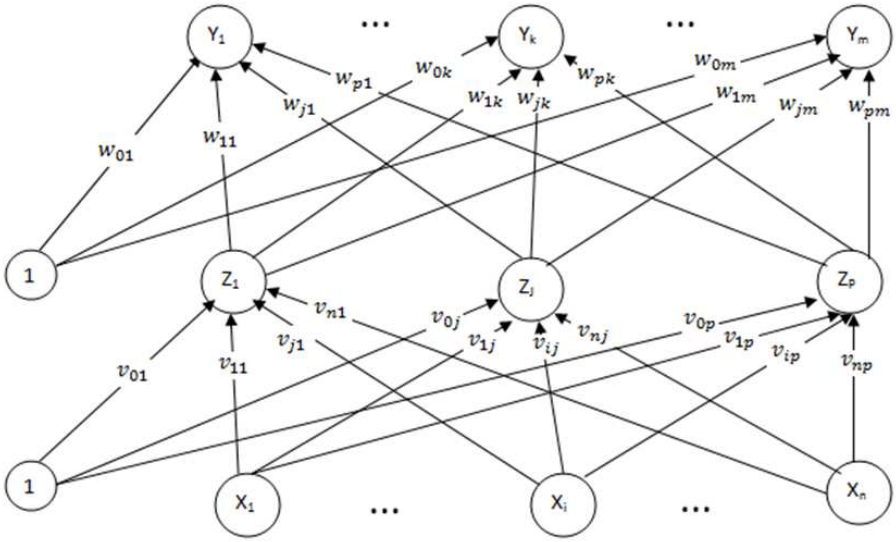

Figure 3.4: Back Propagation Neural Network (BPNN) with one hidden layer [6].

of epochs and passes of huge training data to converge to a suitable decision surface.

Dynamicity: PNNs, RBFs and GRNNs are more adaptive in converging quickly to a decision surface as more neurons can be added at runtime to aid the results compared to BPNNs, which have a fixed number of neurons in the hidden layers.

Suitability for Low Dimensional Data: PNNs, RBFs and GRNNs are more suitable for low-dimensional data like that the different wavelet analysis methods yield through DWT, WPT, WSBC or Irregular Decomposition. On the other hand, BPNNs are more suited to MFCC with multiple dimensions and huge feature matrices.

Best Candidates for Bagging: PNNs, RBFs and GRNNs can be trained much more quickly than BPNNs. Although BPNN-based systems produce relatively more accurate test results than RBFs and GRNNs, BPNNs have the

drawback of requiring huge training times and a large number of instances for training. On the other hand, PNNs, RBFs and GRNNs are the best candidates for bagging as they are simultaneously strong and fast learners.

The next section presents the major highlights of the text-independent speaker identification system based on bootstrap aggregating three equally robust but fast learners: RBF-NN, PNN and GRNN.

3.4 Proposed System with Wavelets and Bagging

This section presents a novel approach for text-independent speaker identification system based on bagging PNN, RBF-NN and GRNN and wavelet analysis techniques. In this system, we use three ANNs (a PNN, a RBF-NN and a GRNN) for classification in speaker identification system using wavelet-based feature extraction methods, namely: DWT, WPT, WSBC and Irregular Decomposition.

The proposed system architecture for speaker identification is illustrated in Figure 3.5. A system with text-independent speaker identification methods was constructed using multimodal neural network with majority voting, including GRNN, PNN and RBF-NN models. All three learners have equal participation or equal weight in contributing towards prediction of the final classification of an instance. This is given as,

$$\begin{aligned} VoteCount(X_i|C_i) &= GRNN\_Output(X_i|C_i) + PNN\_Output(X_i|C_i) \\ &+ RBFNN\_Output(X_i|C_i). \end{aligned} \tag{3.4}$$

Every test instance is passed through each of the three neural networks. If any two of the networks relate the given test instance to the same speaker from the training data, this choice is approved by the system. On the contrary, if any network relates this instance to some other class, then, the system labels it as "not recognisable".

The identification experiment was performed using the GRID [2] speech corpus. GRID is a multi-speaker audio-visual sentence database that supports joint

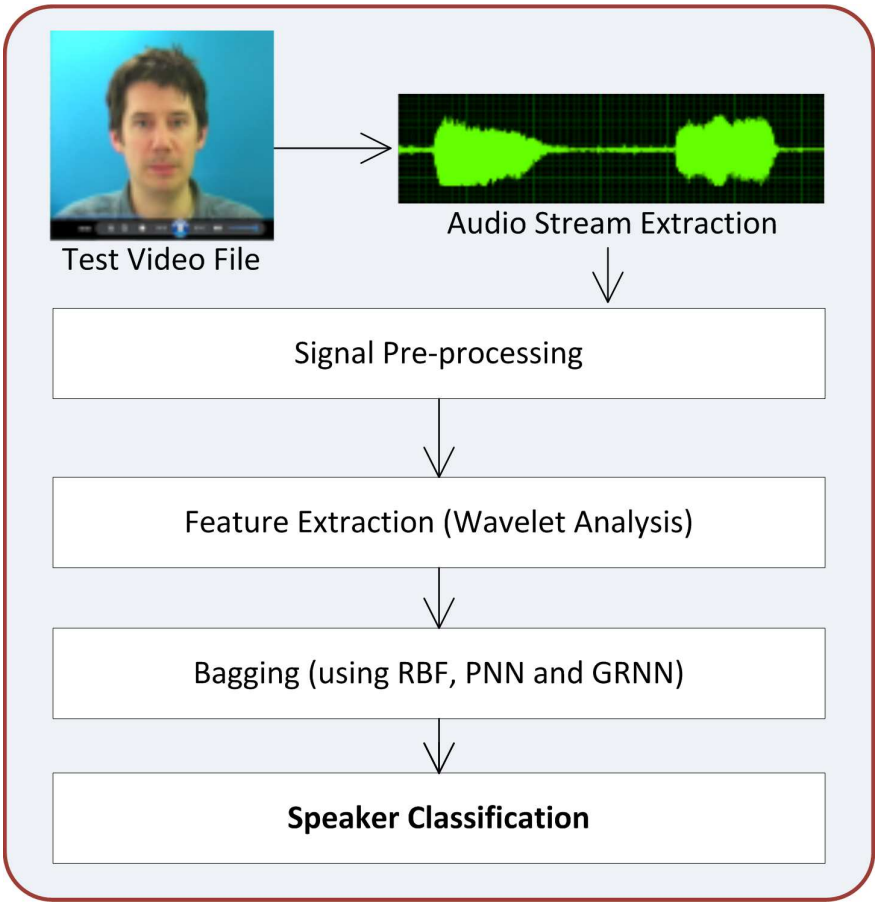

Figure 3.5: High-level system architecture and information flow of the Multimodal speaker identification System.

computational-behavioural studies in speech perception. GRID consists of high-quality audio and video recordings of 1,000 sentences spoken by 18 male and 16 female speakers. It uses simple and syntactically identical sentence structures, such as "put red at G9 now". 10-fold cross validation experiments tested all the 34 speakers from GRID and the results from these 10 experiments have been averaged. The blocks of the system architecture are shown in Figure 3.5.

During the training phase, feature vectors extracted from the training data are fed into each of the networks in parallel. These networks require only one pass

through the data in contrast to the multiple epochs/iterations that are used in BPA. For testing, the extracted feature vectors from the test signal are fed to all the ANNs in parallel and three classification outputs are calculated corresponding to the three classifiers used here. The majority voting scheme is employed for the classification results of the ANNs. The class that obtains two out of three votes is taken to be the final classification result. The training phase is illustrated in Figure 3.6.

During the test phase, the procedure used was almost the same as the one in the training phase. The test speakers file is pre-processed to extract wavelet features. These features are classified individually by the trained PNN, GRNN and RBF-NN. The majority voting scheme ensures equal weight and the final classification is made on the premises. The test phase is illustrated in Figure 3.7.

In contrast to the BPNN and feed-forward networks, none of these networks requires iterative training, which takes a considerable amount of time. Additionally, each of these networks focuses on a different probing level to fit the training data. One of them focuses on the training data completely (over-fitting), the second learns the training data with an error margin (under-fitting) and the third lies between the previous two and thus helps to increase the ability to generalise the overall system for both known and unknown signal instances. Moreover, the combination of these networks with a majority voting scheme helps to overcome the under- and over-fitting problems. This approach improves the classification accuracy of the overall system.

Only the fusion of the PNN, RBF-NN and GRNN networks in the voting scheme is capable of reducing the training time and obtain a higher accuracy than those of the BPNN and feed-forward networks, However, such a method would still be faster than methods that use BPNN and feed-forward networks.

3.5 Testing and Results

Tenfold cross-validation experiments were used to test all 34 speakers in the GRID [2] database using different values of the spread. The spread denotes how closely the neural network should fit the training data. The default value range

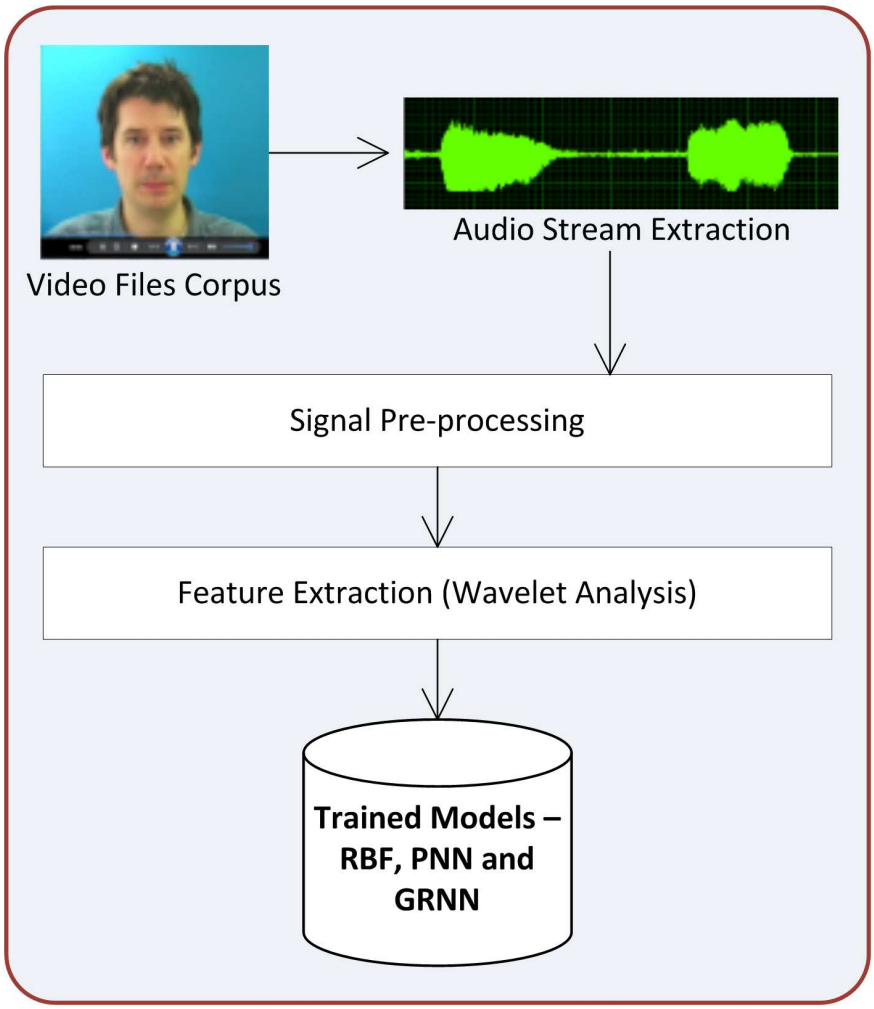

Figure 3.6: Dataflow for the training phase of the Multimodal Neural Network-based speaker identification system.

for spread is between 0 and 1, with 1 being the most generalised fitting to the training data with relatively lower accuracy. A spread of 0 is a complete close fit to the training data and produces maximum accuracy. We can say 1 under-fits the training data whereas 0 over-fits the training data. The spread is also known as the radius of a neuron. With larger spread, neurons at a distance from a point have a greater influence. There is a trade-off in choosing different values of spread

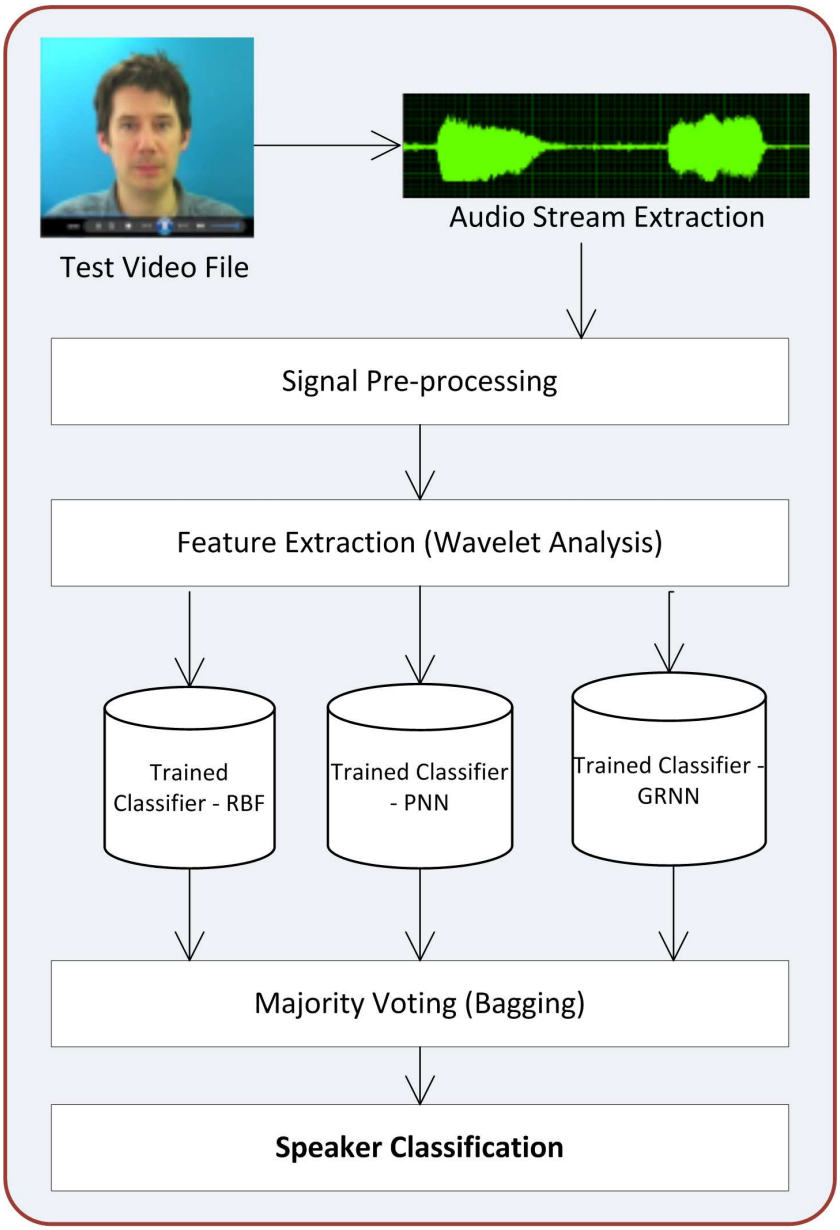

Figure 3.7: Dataflow for the test case of a single speaker.

between 0 and 1. This variable was chosen as the base variable and 30 different values were assigned to it. Therefore, they resulted in 30 different experiments on the same data from the 34 speakers in GRID. The averaged identification results are summarised in Table 3.2.

We also calculated the False Positive Rate (FPR) and True Positive Rate (TPR) for different values of the spread. The fraction of true positives out of the total actual positives is known as the TPR, and the fraction of false positives out of the total actual negatives is called the FPR. The FPR is the same as the complement of specificity (i.e. one minus specificity), also known as the False Reject Rate (FRR). TPR is also known as sensitivity, and is the complement of the False Accept Rate (FAR). The Receiver Operating Characteristic (ROC) curve shows the TPR as a function of FPR. It is a graphical plot that can illustrate a binary classifier's performance with variation of the discrimination threshold. TPR and FPR depend on the size of the enrollment database and the decision threshold for the matching scores and/or number of matched identifiers returned. Therefore, an ROC curve plots the rate of accepted impostor attempts against the corresponding rate of true positives parametrically as a function of the decision threshold. The results can be changed by adjusting this threshold. The 30 experiments we conducted produced different combinations of the FPR and TPR. These two values were plotted using Matlab [185] to generate an ROC curve for DWT, WPT, WSBC, Irregular Decomposition and MFCC for a comparison of accuracy with the proposed Multimodal Neural Network System as shown in Figure 3.8. This curve shows that the ROC curve for WPT lies very close to the upper left boundary and has more area under it compared to DWT, WSBC, MFCC and Irregular Decomposition. The ROC curve for MFCC with the same data lies closest to the diagonal and shows the least effective accuracy as compared to the rest of the pre-processing algorithms.

This ROC curve shows that WPT combined with the proposed fusion system of multimodal neural networks outperforms DWT, WSBC, MFCC and Irregular Decomposition. The main reason why our system works more efficiently is because of the application of the majority voting scheme during the parallel combination of three classifiers. This has not only improved the accuracy for the text-independent speaker identification but also sets new standards for a real-time identification system. The same testing criterion was applied to GMM, BPNN and PCA for comparison with the proposed MNN. Overall, the wavelet packet analysis (8-level) produced an 89.5% identification rate, a huge improvement compared with MFCC (with 20 feature vectors) capable of 77% identification rate.

Table 3.2: Identification rate of the multimodal neural network compared with other algorithms.

	DWT	**WPT**	**MFCC**	**WSBC**	**Irregular Decomposition**
GMM	35.80	38.60	83.30	36.50	33.26
MNN	84.70	89.50	77.50	85.40	80.80
BPNN	40.38	41.47	21.20	34.48	32.07
Parallel BPNN	61.25	65.43	-	58.67	56.85

It is noteworthy that MFCC produces a 2D matrix of size 20×450 (rows $\times$ columns) whereas BPNN is by nature designed to cater for 1D (1 dimensional) input. So there is a mismatch between these algorithms and they can not be used together without losing much of the information in the 2D matrix to make it 1D. Hence, there is no data for MFCC and Parallel BPNN in Table 3.2. Also, it can be observed that MNN outperforms the other classifiers for most feature extraction schemes as expected (in most columns of Table 3.2, MNN results in the highest identification rate). The only exception is MFCC; in this case, GMM gives a better result than MNN, which is due to the following reason. When GMM is fitted to a smoothed spectrum of speech, an alternative set of features can be extorted from the signal. In addition to the standard MFCC parameterisation, complementary information is embedded in these extra features [186]. Combining GMM means with MFCCs by concatenation into a single feature vector can therefore improve identification performance. This is the reason why MFCC performs the best with GMM. However, the best result in this table is achieved using MNN when used with WPT.

Jian-Da Wu and Bing-Fu Lin suggest in [51] that irregular decomposition gives better results than DWT, WPT and WPT in Mel-scale algorithms. On

the contrary, we found that both WPT in Mel-scale and Irregular Decomposition were less accurate than WPT when tested on the [11] speech database. This difference is explained by the authors of [51] when they used a limited, text-dependent set of 5 sentences for each speaker. In contrast, our training set is truly text-independent because the user speaks up to 1000 different sentences.

Performance results summarised in Table 3.2 show that our proposed system gave the best accuracy (89.5%) yet achieved by any text-independent speaker identification. This estimation is based on the comparison with an established set of algorithms, including GMM, PCA, the parallel classifier model [6] and BPNN. It should be noted here that the data used in the parallel classifier experiments were text dependent, whereas in this research and this system we followed a text-independent approach, which is harder to work with.

Table 3.3: Training time (sec) and identification time (sec) for the multimodal neural network compared with other algorithms.

	GMM	**MNN**	**3-BPNN**	**PCA**	**BPNN**
Avg. Training Time	5.8	0.8	120	2.8	90
Avg. Identification Time	2.5	0.05	0.10	1.5	0.8

The proposed system has twofold advantages in terms of accuracy and speed. PCA, due to its dual nature (a classifier and a dimensionality reduction algorithm) is compatible with MFCC as the feature extraction strategy, the system obtained 82.9% identification rate with this combination. Table 3.3 shows the training time and identification time of the state-of-the-art algorithms in speaker identification compared with our multimodal neural network system. The proposed multimodal neural network system takes only 0.05 second on average for the identification phase of a test signal. This is the fastest identification time ever seen in a text-independent speaker identification system. Training is the phase where all the time is spent in traditional systems but the proposed system takes care of it as

it employs the instantaneously adaptable classifiers in parallel with no training time.

The performance results described in Tables 3.2 and 3.3 are fully supported by the maximum area under the curve with WPT for MNN in the ROC curve in Figure 3.8. This comprehensively validates that the proposed system is much more adapted to today's real-time demands, as it outperforms competing approaches both in terms of accuracy and speed. As stated before, the proposed system owes its performance improvement to: (a) bootstrap aggregating of multiple classifiers for a better hypothesis in the decision space; (b) careful selection of multiple combined ANN instances of the same class that complement each other by tackling the under- and over-fitting problems; and (c) selection of the most suitable feature extraction strategy (i.e. WPT). Instead of using other recently popular methods like BPNN methods, we explored the more adaptive instantly trained class of neural networks (PNN/GRNN/RBF-NN), and it substantially improved the classification accuracy as well as reduced the identification time.

3.6 Summary

Conventional approaches to speaker identification with slow identification and poor accuracy are inadequate in a real-world setting. We have been motivated by these shortcomings to conceive and implement in this doctoral research, a novel approach. This approach combines multiple neural networks (an idea from machine learning to improve classification) with wavelet analysis to construct a system that outperforms the classical GMM, BPNN, multiple classifier [6] and principal component analysis techniques, in terms of both accuracy and identification time. In the course of a comprehensive testing on the GRID [2] database, the current system showed an overall accuracy of 89.5%, with an identification time of 20 ms if WPT is used as the feature extraction method. Our real-time approach is directly applicable to industrial devices for security and authentication.

This system has competitive performance compared to GMM, PCA and multiple classifiers but it also has, like any other system, its limitations. For example, the number of speakers (scalability) was an important factor. MNN was quite

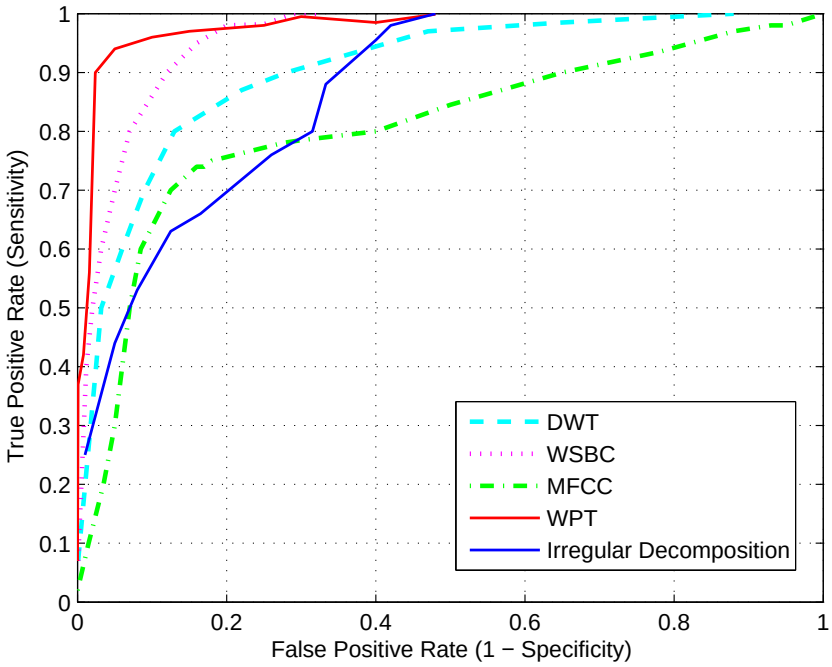

Figure 3.8: ROC curve showing the true positive rate against false positive rate.

accurate up to 20 speakers but it slowly lost accuracy with increases in the number of speakers above that. This fact made the current system more suitable for a small-level deployment.

In this chapter, we described the design and implementation of a text-independent multimodal speaker identification system based on wavelet analysis and neural networks. Wavelet analysis comprises DWT, WPT, WSBC and MFCC. The learning module comprises GRNN, PNN and RBF-NN, forming decisions through a majority voting scheme. The system was found to be competitive and it improved the identification rate by 15% as compared to the classical MFCC. In addition, it reduced the identification time by 40% as compared with BPNN, GMM and PCA based on the performance testing conducted on the GRID [2] corpus.

The main reason why our system works more efficiently is because of the ap-

plication of the majority voting scheme during the parallel combination of three classifiers, which are all fast and robust. These classifiers may suffer from inadequacies such as under- or over-fitting problems when used alone, but mitigate each others shortcomings when combined in the proposed manner.

Our original method lays the foundation for further research in speaker identification for real-time systems. In the future, we plan to further improve the current approach by combining real-time facial recognition with speaker identification to make the system more robust and applicable for industrial use. The idea is to combine audio and visual features in combination with MNNs to further improve accuracy.

Chapter 4

A Highly Scalable Text-independent Speaker Identification System using Vowel Formants

4.1 Introduction

Earlier in this research, in Chapter 3, various wavelets were explored along with bagging neural networks, but the methodology, although very powerful, had some limitations related to the number of speakers. The performance dropped gradually as more and more users were registered with the system. This made the system inapplicable to medium and large scale deployments. The need for a more scalable system emerged as the logical next step towards realisation of the goals of this doctoral research.

The main contribution of this chapter is the design of a scalable text-independent speaker identification system based on vowel formant filters and a scoring scheme for classification of an unseen instance. MFCC and LPC have both been analysed for comparison to extract vowel formants. It is observed that LPC is more efficient in this task. LPC was developed in 1960s [187], but is popular and widely used because LPC coefficients represent a speaker by modelling vocal tract parameters. It is the most powerful and most computationally efficient way of estimating formants. The reasons lie in the close similarity of this strategy with the human

vocal tract. For identification, the proposed score-based strategy has been compared with BPNN and GMM. Our score-based strategy outperforms both BPNN and GMM.

This chapter is organised as follows.

In Section 4.2 we present an overview of the vowel formants and their role in identifying a speaker uniquely. In Section 4.3, we describe the feature extraction process through LPC. In Section 4.4, we explain the vowel formant filtering process for English language instances. In Section 4.5, we highlight the vowel database construction scheme. In Section 4.6, we illustrate the key elements and components of the proposed vowel formant-based scalable text-independent speaker identification system. In Section 4.7, we present the classification method using the proposed max score scheme. In Section 4.8, we include a comprehensive validation and testing of the proposed vowel formants-based scheme. In Section 4.9, we compare the results with the system developed in Chapter 3. Finally, in Section 4.10, we conclude with a discussion of the effectiveness of the proposed scheme and discuss the limitations.

4.2 Overview of Vowel Formants

In 1962, vowel formants and frequencies were exhaustively studied and formulated by J.C. Wells [188]. This was one of the few approaches researchers in the speaker identification field started investigating. An audio formant refers to the recurrent frequency peaks in a speech signal. These recurrent peaks show up with different frequencies in a speech signal. These are also called resonant frequencies. These frequencies resonate according to the vocal tract of the speaker. Vowel formants refer to the recurrent frequencies associated with vowel sounds in a language. Vowel formants have been found to be unique for each speaker and these frequencies lie in a specific range [189].

The current approach evaluated the hypothesis that vowel formants extracted from speech signals of various speakers could be used to distinguish one speaker from another while constructing a scalable speaker identification system. The classical MFCC method is the most popular feature extraction strategy extensively used so far. It extracts the pitch tracks in a user's speech signal. MFCC

was used with formant analysis for comparison with LPC. Vowel formants were extracted using the standard LPC scheme. With LPC, all the recurring formants were extracted, where each formant portrays itself in terms of four resonating frequencies, as illustrated in Figure 4.1.

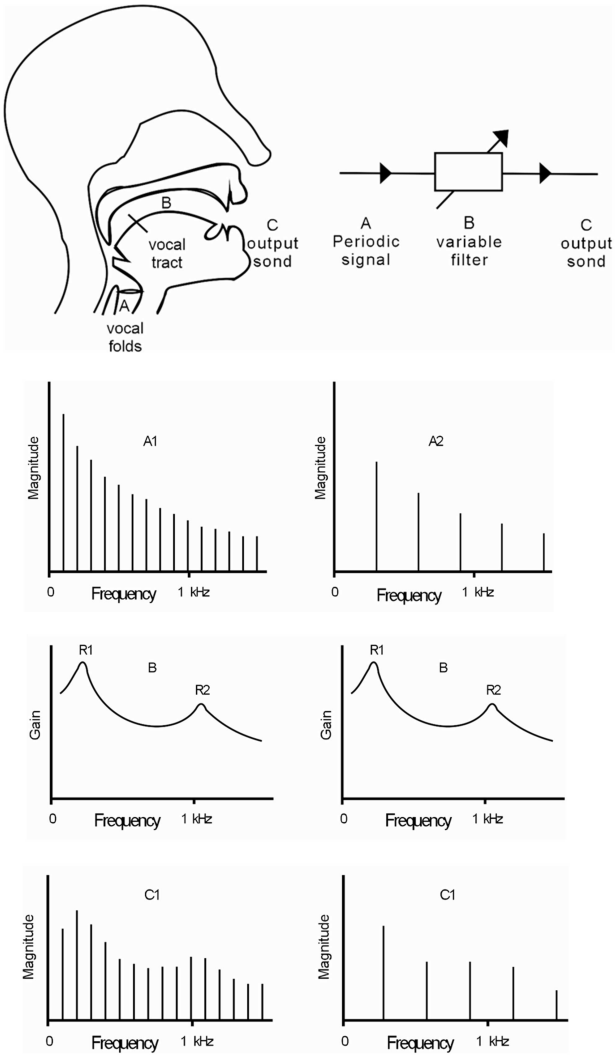

Figure 4.1: Resonating frequencies R1 and R2 for a sample vowel, inspired from [9].

These formants were filtered with a vowel formant filter to separate the vowel

from the consonant formants. During the training phase of the system, a vowel formant database is created after processing the training signals that store unique vowel formants for each speaker. In order to distinguish one speaker from another, vowel formants are tracked in the test file and are compared against the vowel formants database.

A score-based scheme is used to assign the current signal to the speaker with the highest number of matching formants for the current test signal. This scoring scheme also follows a penalty rule that assigns a negative score if a formant does not match with the current vowel in hand from the test file. MFCC and LPC have both been analysed for comparison to extract vowel formants. It is observed that LPC is more efficient in this task. It is the most powerful way of estimating formants, and is also computationally the most efficient. The reasons lie in the close resemblance of this strategy with the functioning of the human vocal tract.

4.3 Formant Extraction through LPC

The fundamental idea behind speech formants is the assumption that an audio signal is produced by a buzzer at the end of a tube which closely resembles the actual way sound is produced by humans. The glottal part, or buzzer, produces the sound with the help of our breath pressure, whereas the human vocal tract combined with mouth is the tube. The audio speech can be fully described in terms of frequency graph and loudness [9]. With this assumption together with the vocal tract and mouth being considered the tube, the human voice is seen as consisting of resonating frequencies called formants [190].

LPC processes a signal in chunks or frames (of 20 ms each) to extract these resonating frequencies or formants from the rest of the noisy signal through inverse filtering [190] [191]. LPC assumes that the next frame frequency can be predicted or expressed in terms of previous observed frequencies. For a given input sample $x[n]$ and an output sample $y[n]$, the next output sample $y'[n]$ can be predicted with the equation,

$$y'[n] = \sum_{k=0}^{q}(a_k x[n-k]) + \sum_{k=1}^{q}(b_k y[n-k]), \tag{4.1}$$

where the coefficients a and b correspond to the formants. The difference between the predicted sample and the actual sample is called the *prediction error* given by,

$$e[n] = y[n] - y'[n]. \tag{4.2}$$

Therefore, we can write,

$$y[n] = e[n] - \sum_{k=1}^{q} b_k y[n-k]. \tag{4.3}$$

The linear predictive coefficients b_k are estimated using an autocorrelation method which minimises the error using least square error reduction [192].

4.4 Vowel Formant Filtering

Formants are defined as "the spectral peaks of the sound spectrum of the voice" [193]. In the field of speech science, a formant also implies acoustic resonances of the human vocal tract [194]. It can be quantified as amplitude peaks in the sound spectrum using spectrograms. In acoustics, it usually means a peak in the envelope of a sound or a resonance occurring in a sound source. Formants are the distinguishing or meaningful frequency components of human speech. Some investigators have suggested that vowel formants exclusively provided the essential acoustic information for vowel identity, while other investigators have emphasised the importance of overall spectral shape on vowel identification [194].

In human speech there are consonant formants, vowel formants and noise reverberations. Out of all these we are interested in only the vowel formants. There are twelve vowel formant sounds in the English language, as concluded by a study at the Dept. of Phonetics and Linguistics, University College London [188]. These vowel formants, with their first, second and third formant frequency ranges, are listed in Table 4.1.

There are several methods of filtering vowel formants. It can be done by passing it through a series of bandpass filters in the audio frequency domain and systematically varying the filter width and slope [195; 196], by low-pass or high-pass filtering in the temporal modulation domain [197; 198], or by varying the

Table 4.1: Vowel formant frequencies in English language [10].

Vowel	Formant	Mean Frequency (Hz)	Std. Dev.
/i/	1	285	46
	2	2373	166
	3	3088	217
/I/	1	356	54
	2	2098	111
	3	2696	132
/E/	1	569	48
	2	1965	124
	3	2636	139
/æ/	1	748	101
	2	1746	103
	3	2460	123
/A/	1	677	95
	2	1083	118
	3	2340	187
/Q/	1	599	67
	2	891	159
	3	2605	219
/O/	1	449	66
	2	737	85
	3	2635	183
/U/	1	376	62
	2	950	109
	3	2440	144
/u/	1	309	37
	2	939	142
	3	2320	141
/V/	1	722	105
	2	1236	70
	3	2537	176
/3/	1	581	46
	2	1381	76
	3	2436	231

number of audio-frequency channels in the context of cochlear implant simulations [199; 200]. In [198], it has been demonstrated that both low-pass and high-pass filtering in the temporal modulation domain were analogous to a uniform reduction in the spectral modulation domain.

4.5 Vowel Database Construction

Recent trends in constructing large speech database files for research purposes are now worldwide. Databases containing varied linguistic features can be built by condensing large corporations. Speech researchers' activities are dependent on the scale and quality of the speech database. In the US, many speech databases collected within DARPA family became available to the general public, and in France, there is a large speech database project through the GRECO group. The work in [201] deals with the general framework of the construction of databases presenting varied linguistic features. We need a continuous speech database made of a large number of phonetic units, which is small at the same time.

Each vowel formant lie in specific frequency ranges but every speaker has a unique vocal tract and produces vowel formants which are unique. During the training phase, the system is presented with speech files for different speakers. These speech files are pre-processed with LPC to create formants that are filtered to extract only vowel formants. These vowel formants are stored with each speaker's name in a database. This database is a Matlab [185] file to be used during the testing phase of the system.

4.6 Proposed System Architecture

The proposed system comprises a video file processing unit which extracts the audio stream from the video file. This component was implemented in Matlab [185] along the other components of the system. This was followed by a formant extraction component that used LPC to find the first 3 fundamental frequencies of resonating formants. The formants at this stage were not exactly the vowel formants but a mixture of consonants and vowels that produced resonating frequencies.

The next component was a frequency filter that separated English language vowels from the rest of the formants found in the signal in the previous step. This process was done in bulk for each speaker's video files and vowel formant frequencies were collected to form the speaker's model and stored in the database for future reference. The training phase of the proposed system is shown in Figure

4.2.

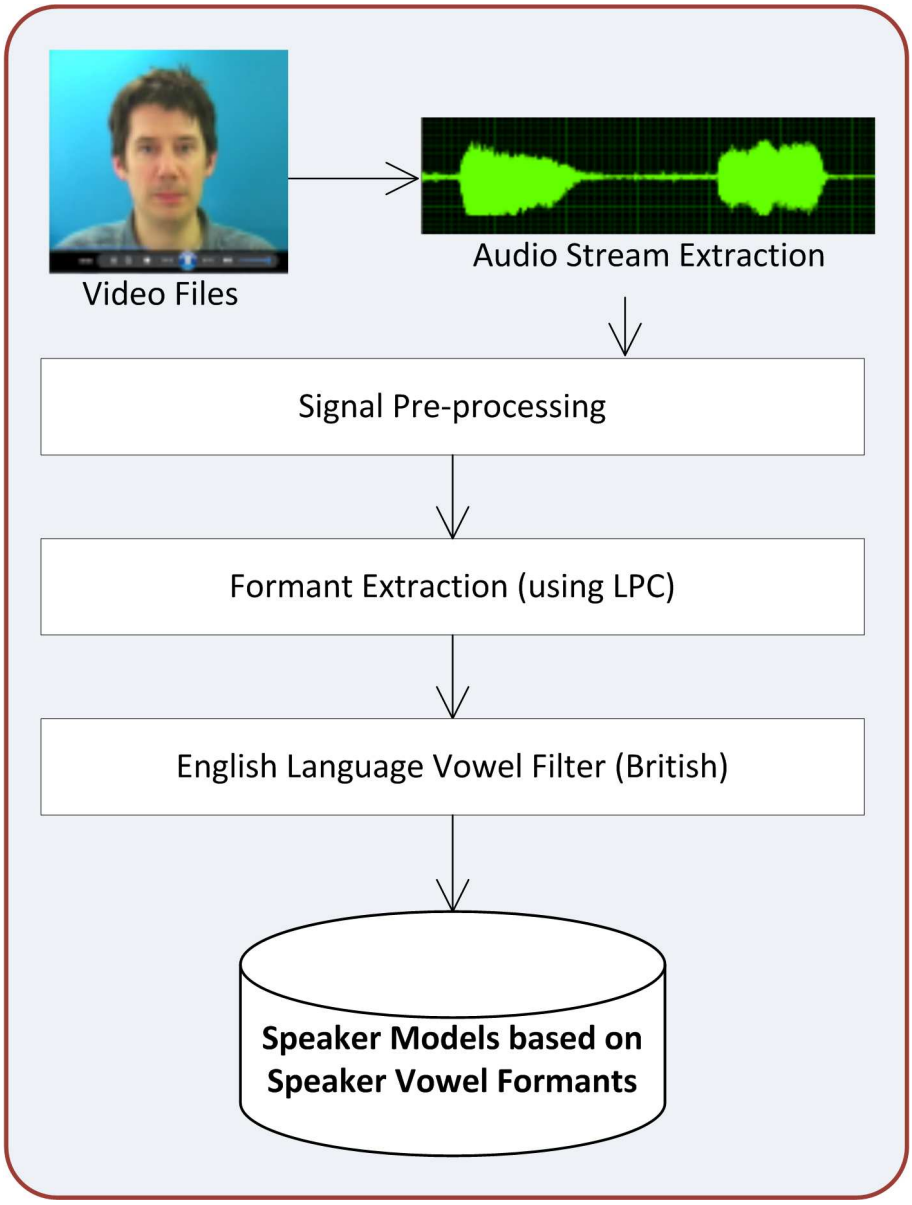

Figure 4.2: Proposed system architecture for the training phase.

The training phase depicted in Figure 4.2 has similar information as the testing phase seen in Figure 4.3 with the only difference being the last processing

steps. In the testing phase, the system compares the vowel formants extracted from the test file with the stored speaker models for a close match. The speaker's model that obtained highest matching score for the current test formants wins the classification. However, if the score falls below a threshold value, the speaker is rejected as an "imposter". This behaviour of the system is demonstrated in Figure 4.3.

The next section weighs the pros and cons of the Max score scheme used for classification in this system.

4.7 Classification with the Max Score Scheme

The testing phase of the system requires the test signal to be pre-processed with LPC and that vowel formant filtering be used to extract the unique vowel formants, along with their first, second and third formant frequencies present in the sample. The vowel formants are then compared to the vowel formants in the database constructed during the training phase.

As a preliminary effort, different strategies for comparing the test vowel formants with the known vowel formants in the database were tested. These strategies included the following combinations:

1. Both the first and second formants;
2. The first, second and third formants;
3. Both the first and third formants;
4. Both the second and third formants; and
5. Averaging and comparison with least distance.

Extensive testing of these enumerated schemes against known results revealed that these strategies are not powerful enough to yield high accuracy as vowel formants often overlap for the same vowel among different speakers. Sometimes it is only one of the three formant values which overlaps, and sometimes it is two of the three formant values that overlap with only the difference of the third formant frequency. This challenging complexity is attributed to the text-independent

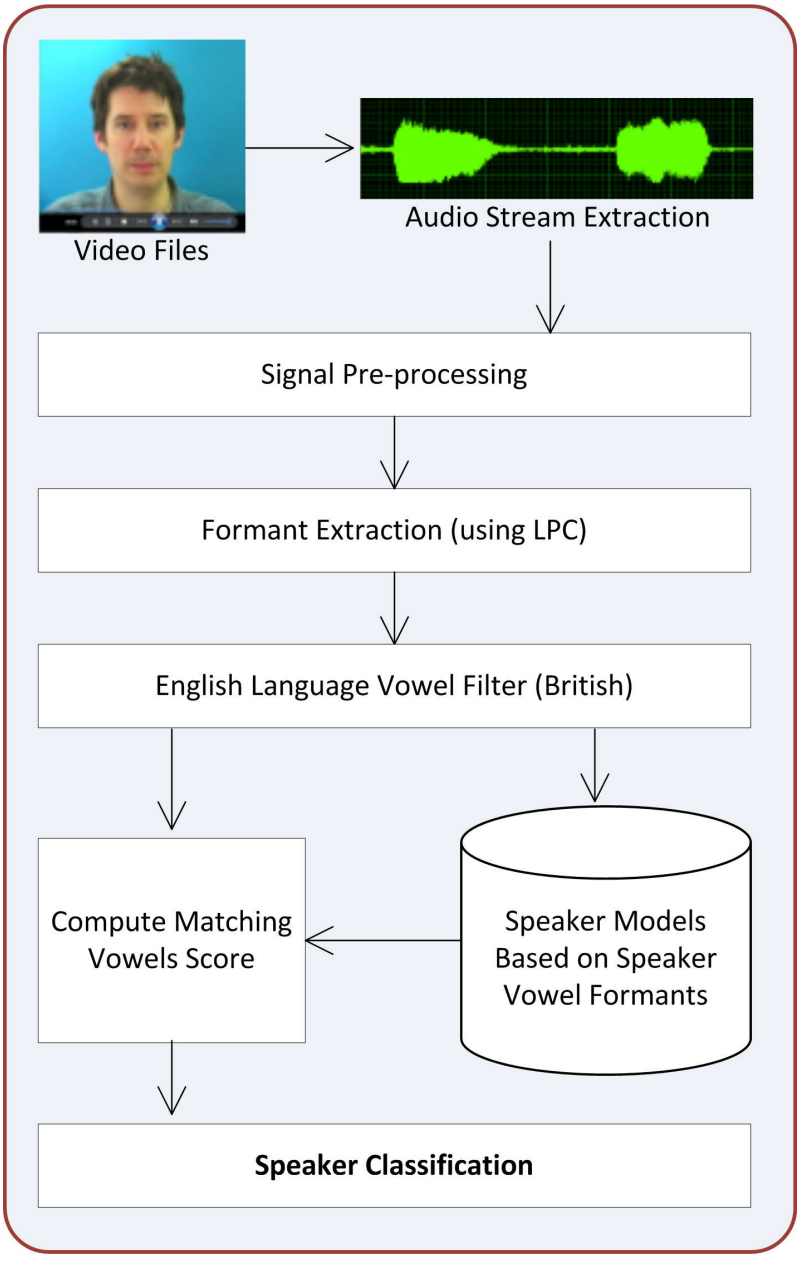

Figure 4.3: Proposed system architecture for the test phase.

nature of the system, where we have a speaker uttering the same vowel but in a different word with a slightly different formant track.

To handle this kind of situation, a score-based scheme was conceived that

awards a positive score if all three formants are matched and penalises a speaker with a negative score otherwise. For a given speech signal, the three test formants are compared against the vowel formant stored in the database for each speaker. For example, if $D_{x,k,1}$ is the first formant frequency of vowel formant k stored in the database for speaker x and $T_{k,1}$ is the first formant frequency of k in the test speech, we conclude that they are "matched" when the difference between the two, i.e. $T_{k,1} - D_{x,k,1}$ is below a certain threshold $\varepsilon_{k,1}$. In this case, we say that the difference, diff$(T_{k,1} - D_{x,k,1})$ is zero. This quantity is computed for all three frequencies $D_{x,k,1}$, $D_{x,k,2}$, $D_{x,k,3}$, and the test score is calculated as,

$$\sum_{m=1}^{3} diff\left(T_{k,m} - D_{x,k,m}\right) \quad = \quad 0 \longrightarrow Score(S_{k,x}) = 1, \tag{4.4}$$

$$\sum_{m=1}^{3} diff\left(T_{k,m} - D_{x,k,m}\right) \quad > \quad 0 \longrightarrow Score(S_{k,x}) = -1, \tag{4.5}$$

The thresholds $\varepsilon_{k,m}$ are selected experimentally. These two scores aid in calculating the net score of each speaker x, against the test vowel as,

$$Identification(k) = arg.Max(Score(S_{k,x})) \quad for \quad k = 1, 2...n. \tag{4.6}$$

4.8 Results and Analysis

We used GRID test sets (described in Chapter 3) to evaluate BPNN and GMM against the score-based approach proposed in this chapter. Table 4.2 compares the performance results of the proposed scheme with those of other state-of-the-art approaches for speaker identification. The same vowel formants were experimented with using BPNN [10] for identification against the same training files and their extracted formants, for comparison with the proposed score-based scheme. The same training vowel formants were supplied as inputs to the Gaussian Mixture Model (GMM) [202] and the test sets were tested against these mixtures. The experiments revealed that the vowel formants with the score-based strategy are not only more accurate in identification but also that this is a more scalable model as it gives almost a linear accuracy when the number of speakers is increased incrementally from 10 to 30, in increments of 5 and 10 as highlighted in

Table 4.2.

An identification algorithm is critically evaluated for its accuracy against the test data for a number of speakers. The context of evaluation gets more critical if the algorithm aims to be applicable for industry devices for biometric security and identity management [203]. Therefore, as the number of speakers increases traditional algorithms start becoming less accurate. It has been an important consideration while designing and testing the current system to ensure that it is a scalable model.

Table 4.2: Performance comparison (accuracy in %) with BPNN and GMM algorithms.

	Score-based Scheme	**Formants with BPNN**	**Formants with GMM**
10 speakers	97.12	58.24	52.33
15 speakers	96.70	53.39	32.86
20 speakers	95.25	44.70	23.20
34 speakers	95.12	40.00	20.90
Avgerage Accuracy (%)	96.05	49.08	47.82

The performance graph depicted in Figure 4.4 shows the performance statistics as the number of speakers is gradually increased from 10 to 30. It is to be noted that both GMM and BPNN start losing accuracy as the number of speakers increases while the proposed score-based scheme shows a linear accuracy which is not affected by the number of speakers.

During these tests, the identification time for the score-based strategy was also observed as shown in Table 4.3. Although BPNN takes less identification

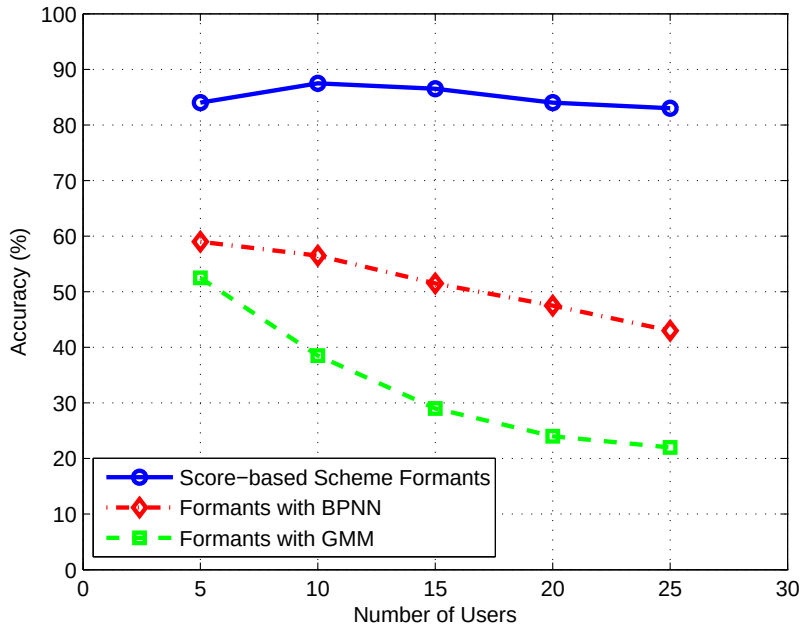

Figure 4.4: Performance statistics of the three algorithms with varying number of speakers (percentage accuracy vs. number of users).

time than the score-based scheme, it has a major drawback when it comes to the accuracy of its results. GMM for formants is not only less accurate but also slower than the proposed score-based scheme.

Table 4.3: Comparison of the average training time and identification time (sec).

	Score-based Scheme	**BPNN**	**GMM**
Training Time	2.5	80	150
Identification Time	0.11	0.01	4.3

It is to be observed that the score-based scheme does not require any training other than saving the filtered vowel formants in the database, which in this case is a Matlab [185] file. Please also note that the identification time does not include the pre-processing time with LPC and vowel formant filtering as that was considered to be common for all the algorithms tested in this case.

4.9 Performance Comparison

This section compares and discusses the performance benchmarks of other speaker identification systems with the proposed vowel formants-based scheme.

Figure 4.5 shows the ROC curves for the proposed scheme as well as for the BPNN and GMM algorithms. The proposed formants with Max score scheme has the maximum area under the curve, while BPNN and GMM covers only just above 50% of the area in the graph. As can be seen, the proposed scheme can achieve a very high level of True Positive Rate (TPR) while restraining the False Positive Rate (FPR) to a low level. For both BPNN and GMM, the FPR would severely have to be sacrificed in order to gain a reasonable TPR. Figure 4.5, along with the results presented in Section 4.8, clearly shows that the proposed Max score scheme outperforms the BPNN and GMM classifiers. Both these classifiers are widely used and regarded as efficient schemes in many speaker identification implementations. However, in a vowel formant based scheme, they fail to perform at a desired level due to the following reasons.

BPNN has the problem of entrapment in local minima, and the network should be trained with different initial values until the best result is achieved. The number of hidden layers and neurons in each layer are required to be determined. If the number of layers or neurons is inadequate, the network may not converge during the training; if the number of the layers or neurons is chosen to be too high, this will diminish the effectiveness of the network operation. This often causes this algorithm to be biased by a specific resonant frequency while completely disregarding the others. The proposed score-based scheme tackles this problem better by exploiting information received from all frequencies. A score-based scheme allows the speaker with the highest matching formants to own the current signal. Furthermore, we choose LPC as the accompanying feature extraction

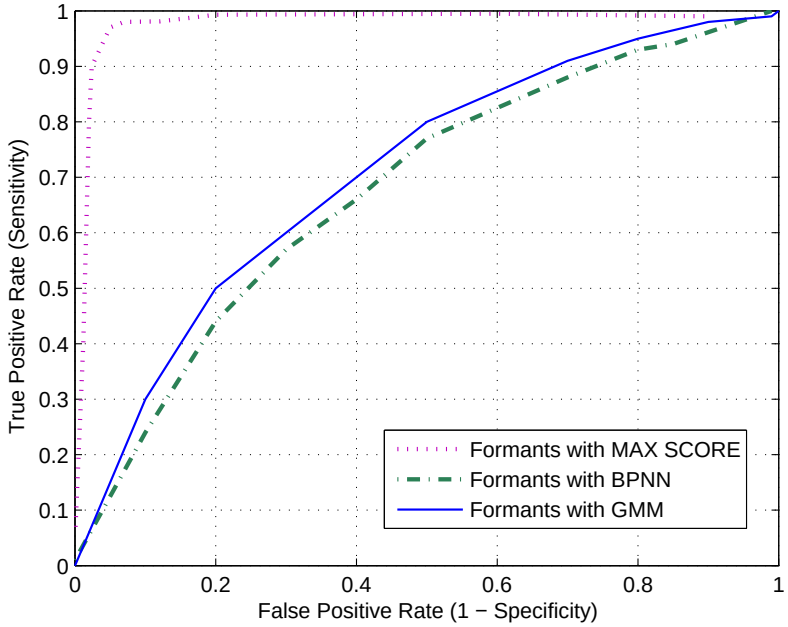

Figure 4.5: ROC curve showing the maximum area under the curve with the proposed scheme.

strategy of our novel scheme, which is the best strategy due to its resemblance with the functioning of the human vocal tract.

Another difficulty of BPNN lies in its use of the back propagation algorithm that is too slow for practical applications, especially if many hidden layers are employed. The appropriate selection of training parameters in the BP algorithm is sometimes difficult. As for the GMM algorithm, its main limitation is that, it can fail to work if the dimensionality of the problem is too high. This causes the GMM to suffer badly when the number of speakers increases. Another disadvantage of the GMM algorithm is that the user must set the number of mixture models that the algorithm will try and fit to the training dataset. In many instances the user will not know how many mixture models should be used and may have to experiment with a number of different mixture models in order to find the most suitable number of models that works for their classification problem.

Finally, we compare the proposed vowel formant-based approach with our

previously proposed system based on multimodal NN and wavelets in Chapter 3 in Table 4.4. We also compare it to another state-of-the-art speaker identification system available in the literature. Our earlier proposed method in Chapter 3, although very powerful, had some limitations related to the number of speakers. The performance dropped gradually with an increasing number of users. This fact led us to design the highly scalable vowel formant based scheme that yields a very good level of accuracy. According to many investigators, vowel formants can exclusively provide the essential acoustic information for speaker identity; and our results corroborate this claim. On top of that, it has a shorter identification time and requires less than 100 bytes of data to be saved for each speaker to be identified. This makes it suitable for large scale implementation with a very high number of users. Therefore, the vowel formant based scheme is comparable to or better than most other speaker identification systems in terms of recognition accuracy yet a better choice in terms of scalability.

4.10 Summary

This chapter explored the combination of LPC-based vowel formants coupled with a score-based identification strategy. For comparison, two other combinations of LPC-based vowel formant methods have been tested with BPNN and GMM. A comprehensive testing on the GRID [2] corpus revealed that our proposed scheme outperforms BPNN and GMM based schemes. On average with 10-fold cross validation, BPNN is 49.08%, GMM is 47.82% and our scheme is 85.05% accurate when LPC is used to extract the vowel formants. We observed that our scheme does not require any training time other than creating a small database of vowel formants. It is therefore faster as well. The results also show that increasing the number of speakers makes it difficult for BPNN and GMM to sustain their accuracy. Both models start losing accuracy whereas the proposed score-based methodology stays almost linear. Our scheme outperforms the multimodal NN and wavelets-based text-independent speaker identification system proposed earlier in Chapter 3. The vowel formants-based approach with the max score scheme has competitive accuracy when compared with the state-of-the-art speaker identification systems in the text-independent domain.

Table 4.4: Performance comparison with state-of-the-art speaker identification approaches.

Reference	**Approach: Algorithm, Database**	**Performance Accuracy (%)**
[48]	GMM-UBM (independent of pre-processing algorithm), TIMIT	96.8
Multimodal NN and Wavelets (Chapter 3)	Wavelets, Multimodal Neural Network 34 speakers (GRID) 43 speakers (VidTIMIT)	89.12
Proposed Vowel Formants with Max score	LPC, Formants, 34 speakers (GRID)	96.05

One of the limitations of the system was with files with no vowels or where the volume was too low to allow proper vowel detection by our algorithm. This limited the system to those domains in which the signal was long enough to contain vowel sounds (at least 5 seconds) to be classified.

Chapter 5

Audio-Visual Speaker Identification

5.1 Introduction

Chapter 5 describes the architecture, implementation and performance benchmarks of a novel method for audio-visual text-independent speaker identification based on fusion of face and voice. This is the main contribution of this chapter.

This chapter describes both the face (denoted face recognition) and the voice (denoted speaker identification) parts of the proposed system. The background on fusion techniques are discussed in Section 5.2. Section 5.3 describes the subprocesses, architecture and implementation of the fusion of face and speaker identifications. Section 5.4 contains a comprehensive evaluation of the fusion-based scheme on the GRID [2] and VidTIMIT [11] corpora. Section 5.5 compares the proposed scheme with the state-of-the-art approaches and the approaches developed earlier in the current research. Section 5.6 concludes the chapter with a summary of the overall progress in this chapter along with a set of limitations of the proposed approach.

Chapter 5 is focused on fusing two different modalities of face and voice at the feature-level and decision-level using GMM and PCA. It proposes a robust, fast, text-independent audio-visual speaker identification system based on score-level fusion of GMM (voice) and PCA (face) which outperforms the state-of-the-art approaches. It also describes the minor performance tweaks done for both GMM and PCA to improve the overall performance and computation time. Perfor-

mance testing and benchmarks conducted on the GRID [2] and VidTIMIT [11] audio-visual corpora are reported in a publishable paper. It evaluates GMM and PCA and the hybrid of both approaches for feature-level and decision-level fusion. Next, it describes the architecture of the proposed scheme and comprehensively validates the test results in comparison with the research done in previous chapters. The chapter concludes with a description of limitations and a discussion of future directions.

5.2 Background on Fusion Techniques

Many tools and approaches have been reported to fuse image and voice signals. In [204], a bimodal identification design is realised using image and speech information. Fusion weights of image and voice signal are determined by means of distances between sampled data (i.e. image and voice data), and the standard deviation and hyperplane of the training data. In [166], the scheme computes scores of each feature's recognition result for the multimodal biometrics features. Ban et al. [165] also proposed a fusion recognition method by using face features and speech features, but their design worked in a normal environment. Zhang et al. proposed to combine face and ear features to recognise a subject's identity and enhance identification rate [205], but their method uses only image prediction. Features in too bright or dark environments are not well extracted. In this work, face recognition is performed based on principal component analysis (PCA), which is often denoted as the eigenface [206; 207]. It is a highly effective method and does not incur much computational burden [208].

The benefits of multimodal biometrics fusion have been corroborated in different studies found in the recent literature. A detailed study of signature detection and palm veins was explored and presented in [209]. Morphological operations, along with the Scale Invariant Feature Transform (SIFT) algorithm, were used to extract features relating to both modalities for the purpose of comparison. A simple sum-rule facilitated the feature-level fusion of the two modalities, and DCT was employed to reduce the dimensionalities of the fused feature vectors. The classification of different speakers was done using linear vector quantisation upon adjusting the parameters.

In [210], a novel identifier was proposed called the Finger-Knuckle-Print (FKP) and a multi-instance fusion method was performed in feature-level. This scheme was shown to achieve better results than any method employing a single instance. The effect of various fusion rules were studied in this paper. The Median or the Min-Max schemes did not perform as well as the fusion of two instances, and the performance was even worse when Z-score was used for estimation in the single instance scenario. In [211], several other multimodal tactics in biometrics were studied and compared. The merits of the biometric identifier sets were examined in the context of the individual sensing methods and the level of their mutual interaction in various steps of processing. Almost every research report in multimodal biometrics considers several possible ways of fusing the results. It seems that no fusion approach has emerged that generally achieves a statistically significant improvement over a simple sum of scores but in this they found that min-fusion performed significantly better than sum.

An interesting scheme was introduced in [212] for voice and face fusion. The hyperbolic tangent is a useful function because of its asymptotic properties. It was exploited to normalise and weigh the feature and fusion parameters, and obtain their geometric average in [212]. The combination scheme this multimodal speaker identification system utilised was of hierarchical nature. Kittler et al. [213] concluded that the sum rule, among the many other face-voice fusion techniques they tested, is the least affected by estimation errors. It outperformed most other rules that were devised based on minimum, maximum, median, and product, rather than the sum.

The authors of [214] pruned the database by means of face matching, and then applied fingerprint matching for the identification system they designed with the face and fingerprint biometrics. Several fusion strategies for face and voice biometrics like SVM and tree classifiers were considered in [24], along with the multi-layer perceptron and the Bayes classifier. The latter turned out to be the most efficient method. In [215], three identifiers were combined. In addition to face and fingerprint, they included hand geometry in their system. They employed methods based on linear discriminants and decision trees, but a system based on the sum rule was shown to be the best.

The work reported in [147] attempted to fuse an exceptional piece of data re-

garding the objects, which is derived from their trajectory. They used a weighted sum method that was linear. A sensor network was employed, which was distributed in nature, comprising sensors that captured video data of moving objects (known as the blob). Each of the sensors provides separate blob locations, and in turn, trajectories, which need to be matched up in an orderly fashion so that the probability of correct estimation is high. The authors of [147] adopted an unequal weighting scheme for this coordination; however, no concrete scheme was developed for the valuation of the weighing coefficients. A somewhat similar linear scheme regarding the objects' location information, although with equal weighting, has been proposed in [148] for the purpose of fusion. Feature-level weighted aggregation based scheme was developed in [146], which was applied to human tracking. A number of spatial items, such as motion and colour, and even texture, were incorporated in this research. Although it was suggested that these items need careful consideration for weight allocations, the authors did not develop an appropriate method to accomplish the allocation task. The authors, however, enhanced their contribution by addressing the detection of face, speech and traffic using a sigmoid function for modality normalisation.

The audio-visual system in [149] extracted the speaker recognition decisions and detected the speech events independently, and applied a weighted sum scheme for the fusion of the two decisions. The weights were tuned based on the training data by examining the dependability of the data sets pertaining to the modalities. In a somewhat comparable manner, the authors of [141] obtained a synchrony score to correlate the face and speech data for their fusion. Both weighted sum and weighted product methods were employed linearly at the decision-level to detect monologues. The audio features and the video features were locally assumed to be Gaussian distributed when computing the correlation between face and speech data.

The attention properties of humans, which are markers of their psychological beahaviour, were utilised in [216] for devising a fusion scheme at the decision-level. These behaviours can determine how strong a sound is, or how fast a motion is, etc. The histogram and moment of the colour, or the block wavelet, etc. are measured to deduce a function known as the attention fusion function. It is different from the previously used weighted sum methods, and exploits the

difference between multiple decisions.

The authors of [217] chose the text (from audio) and the texture and colour (from video) as their modalities for the purpose of retrieving video. They combined the visual and text results, upon normalising them (by means of the max-min strategy), and ranked the retrieval indices. A weighted fusion strategy was used linearly at the decision-level, and was found to perform well when the various weights were properly tuned. Another research effort in [218] devised a combination scheme for the retrieval scores in the same way that [217] did, but their goal was to rerank video segments. As before, the scores were obtained from motion along with the audio and video. All these research efforts imply a common suggestion. While weighted fusion can really be helpful if linearly capturing the effects of dissimilar modalities, the assignment of the proper weights turned out to be the most challenging task. This is an underexplored area of research that requires more attention and refinement.

Gesture detections have also been attempted to be fused with speech in some recent works. For example, for the purpose of interacting with a robot designed for household assistance, both speech and gestures can be used. A multimodal integration method has been developed for this cause in [154] at the decision-level. The authors utilised the temporal correlation that exists between gesture and speech. For different event parsers, they generated individual lists recording the relevant speech and gestures, based on which they interpreted their decisions. While this work dealt with 3D gestures, a system addressing 2D gestures along with speech was proposed in [219] as a rule-based fusion system. The target application was interaction between humans and computers. In another work [134], the authors analysed the semantic content of videos and compared the early and the late fusion schemes. For an early fusion scheme, the visual and text vectors are joined and normalised first. The result is fed to an SVM so that the semantic content can be extracted. A probabilistic aggregation scheme was employed for late fusion schemes. It turned out to be the most effective concept because of its increasingly evolutionary learning capacity. However, there are some cases where early fusion outperforms the late fusion by a substantial margin.

5.3 Proposed System Model

Feature extraction is the first and the most important stage of any classification system. The goodness of the extracted features highly affects the performance of the overall system. Audio and visual data, though correlated, are in completely different forms and are sensed differently by humans. Thus, the features used for both are also different. The fields of audio and face recognition are highly developed and many different ways of capturing features are available in the literature. Our approaches to feature extraction are first described in the next two sections.

5.3.1 Audio Feature Extraction

Speaker identification is an expert system based on the single biometric of voice data. It extracts the audio from the provided video file, computes MFCC feature vectors for a speaker-specific information mapping, and creates a speaker-specific single-state HMM-GMM.

An audio stream consists of thousands of values in the range [-1, 1] that are sampled at a regular interval T. An 8 kHz sampling rate means that 8000 such values vary each second when a speaker's audio is recorded. The video files used in this study are approximately 3 seconds long. Thus, there are approximately 10,000 to 30,000 values after trimming the silent sections at the start and end of the audio, which do not contain useful information. These raw values only tell us about the amplitude variations in the speech and do not convey any explicit information about the speaker. Since we are using text-independent speaker identification, we must extract distinguishing speech features that describe a speaker's orientation or, more specifically, the qualities of the speaker's glottal tract which are independent of the language being used. Therefore, if the same speaker speaks a different set of words next time, our system should identify the speaker. Therefore we must transform the raw signal into a parametric representation.

Usually, short-time spectral analysis techniques, such as LPC and MFCC, are used to transform the raw signal into a parametric representation containing the most important characteristics of the signal [183]. MFCC, originally developed for speech recognition systems, employs both logarithmically spaced filters and

Mel-scale filters, which are less susceptible to noise and variations in the physical conditions of the speaker. Herein, we have used MFCC to capture the most phonetically important characteristics for speaker identification from the audio signal. We select the widely accepted MFCC features for our research due to their demonstrated superior performance. As a pre-processing step on the audio signal, we perform pre-emphasis to compensate for the high frequency falloff. We then use the short-term analysis technique using windowing as stated below.

An audio signal is usually segmented into frames of 10 to 30 millisecond (ms) with some overlap [220]. Each frame has to be multiplied with a Hamming window in order to keep the continuity of the first and the last points in the frame. Discontinuities at the edge of the window can cause problems for the FFT. The Hamming window smoothens out the frame edges. Overlapping windows allow analysis centered at a frame point, while using more information. The overlap percentage is usually held at 67% [220], 60% [221], 50% [222], etc. In our case, the audio signal is divided into 15 ms frames using Hamming windows with a 10 ms overlap (66%) to smooth out the frequencies at the edges of each frame or window. These frames are further processed using logarithmically spaced filters and Mel-scale filters, Fourier transforms and cosine transforms to produce an MFCC vector for each frame.

GMM is extensively used for classification tasks in speaker identification [51; 184; 185]. GMM is a parametric learning model that assumes that the process being modelled has the characteristics of a Gaussian process. A Gaussian process assumes that the parameters do not change over time. This assumption is valid because a signal can be assumed to be stationary during a Hamming window. This assumption is also valid for the case of a face image, where the image features, such as the distance between the eyes, the shape of the nose and the area in the forehead, do not change in a single frame.

GMM tries to capture the underlying probability distribution governing the instances presented during the training phase. Given a test instance, the GMM tries to estimate the maximum likelihood that the test instance has been generated from a specific speaker's GMM. The GMM with the maximum value of likelihood owns the test instance and is declared to belong to the respective speaker. In this study, we employ GMM for the classification task. The GMMs

for each speaker (both voice and face) are developed separately.

5.3.2 Face Recognition

Many different kinds of features can be used for face recognition as described in Chapter 2. The most widely used ones include eigenfaces, DCT and Gabor wavelets. Eigenfaces are well suited for face recognition. In this technique, the features independent of the person's facial expression (principal components) are preserved while dynamically changing features are discarded. We adopt this in our studies. The face recognition system is one part of the overall proposed scheme based on fusion of face and voice. It comprises a video stream extraction component, a key frame selector component, the face detector, post processor, feature extractor and classifier. The output of the face recognition is a classification result which contributes towards the overall result of the proposed fusion based system. These components are described below.

5.3.2.1 Video Stream Extraction

The given video file has both video and audio streams encoded in it. During this step, only the video stream is extracted; reading of the audio stream is disabled completely. This is done in Matlab [185] using the multimedia reader object. A video file comprises images or frames sequenced in a timely fashion. One must have seen the making of cartoon films in which individual images, varying slightly from each other in order to mimic the motion of an object, are sequenced in such a way that the human eye identifies the frames as video. The same is the case with the current video stream being discussed. The GRID [2] corpus contains high quality video files which are 3 seconds long and use the standard MPG video format. However, our program is designed to support all major multimedia formats including AVI, MPEG-1, WMV, ASF and ASX without any code change. GRID database video files were captured at 25 frames/second with a total of 75 frames in each file. A frame size of 398 x 288 pixels was fixed for all video files. The proposed system used the built-in Matlab [185] multimedia object called mm reader to read the entire video file and separate the audio signal from this video file as a wave file. See the Appendix for the detailed implementation of this

system.

5.3.2.2 Key Frame Selection

Video files consist of a series of still images. Many applications extract one or more of these still images, termed key frames, for various purposes. A single key frame is often a convenient representation of a video segment. Both GRID [2] and VidTIMIT [11] have a single speaker showing up in each video file. The speaker has an illuminated background and we select only one frame from the video file as our key frame for further processing. After trimming the silent portions from the beginning and ending of an audio signal, the silent zones within the signal were determined. The largest of these portions has a middle frame in which the user is silent with closed lips. Although many advanced algorithms for key frame selection have been developed [223; 224; 225], this simple idea of key frame selection was sufficient for our purposes.

5.3.2.3 Face Detection

The selected key frame is an RGB image with red, blue and green components for each pixel in the image. The key frame also contains the upper body, face including the hair portion, shoulders and neck together with the illuminated background of the speaker. From this entire key frame we need only the face portion, including the hair, forehead, eyes, nose, cheeks and chin. For this purpose the face portion detection algorithm was adapted from [66]. It uses a cascade-based learning scheme which starts by marking the entire image with rectangular areas and then trying to fit these rectangles each around the eyes, nose, cheeks, chin and forehead portions. It uses the distance between these rectangles to converge and detect the entire face portion from the given image.

5.3.2.4 Grey-scale Normalisation

Once the face portion is detected with this algorithm, the next step is the reduction of the data's dimensionality. An image is a three-dimensional matrix containing RGB values for each of the pixels in the 2D image. The image is transformed to a grey-scale 2D matrix using the following formula on the RGB

pixels,

$$GreyImage = R(0.299) + G(0.587) + B(0.114). \tag{5.1}$$

This process is shown with the actual images from the GRID [2] database in Figure 5.1.

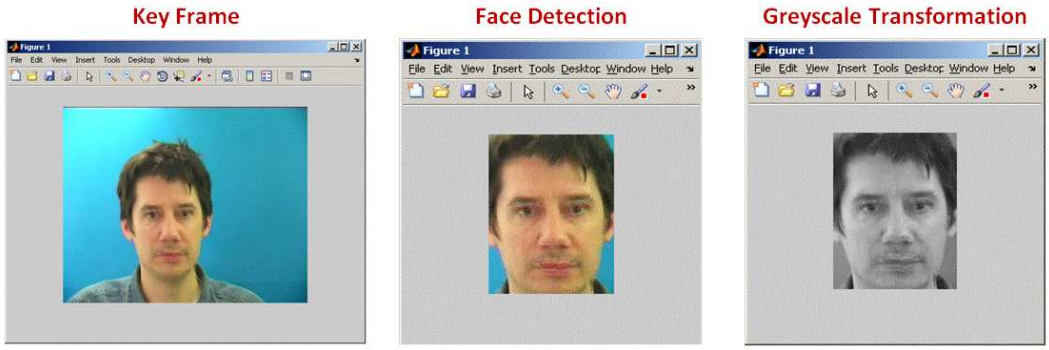

Figure 5.1: Key frame, face detection and grey scale transform on a sample image.

5.3.2.5 Architecture of the Face Recognition System

The enhancements for the face recognition system described in the earlier sections were transformed into a complete system implemented in Matlab [185]. The developed system was thoroughly tested to see the improvements.

The current section presents an overview of the overall system architecture from training as well as testing perspectives.

The system took video files as input and produced a classification result as output and therefore contained a number of components, including a video stream extractor, a key frame selector, a face detector, a PCA feature extractor and, finally, a mean face finder. These components functioned in the same order they are mentioned above and are shown in Figure 5.2.

Training and test data were split using 10-fold cross validation and the results of the 10 experiments were averaged. During the training phase, 90% of the video files (in any given experiment) were selected for training and the rest for

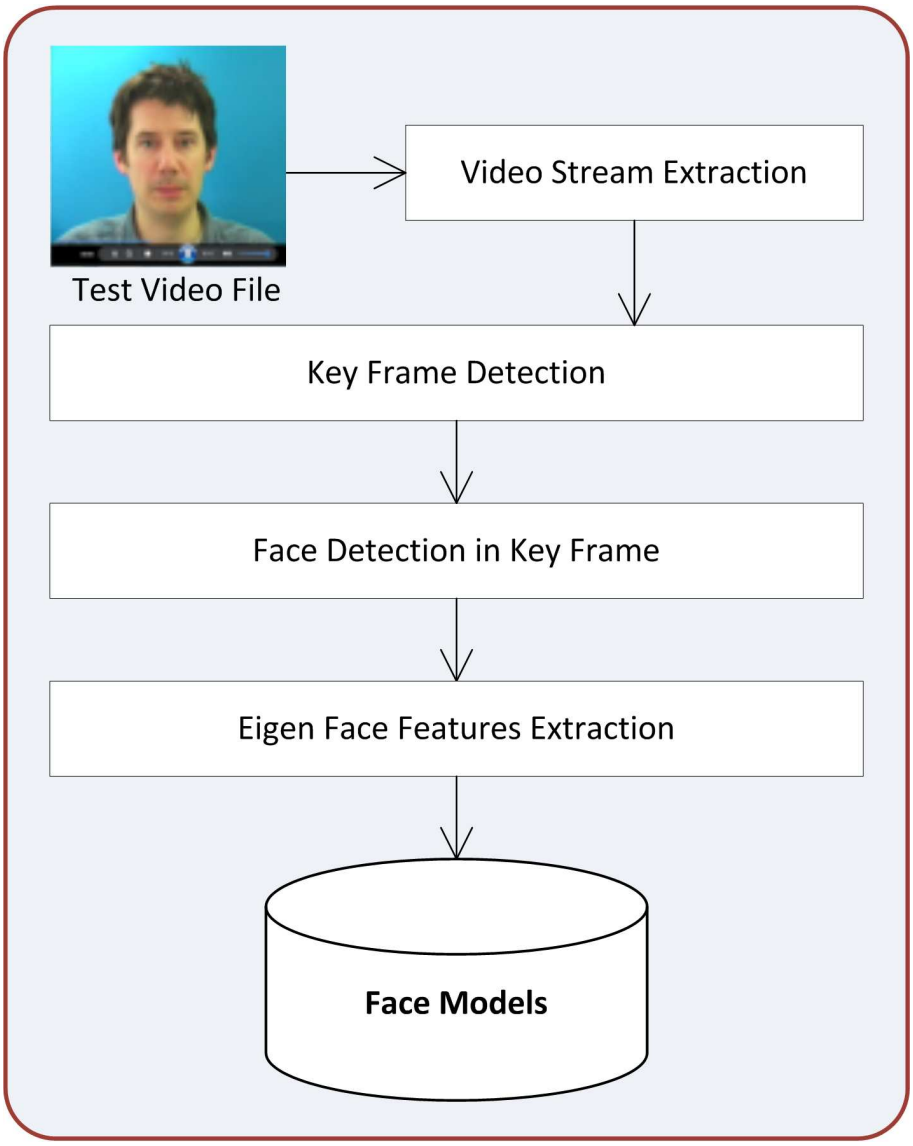

Figure 5.2: Architecture of the face recognition system.

testing. On the training video files the system read in each video file one by one, speaker by speaker. It then detected the key frame in each of the video frames based on the largest silence zone within the video files (excluding the boundaries). A human face was detected from the key frame in the very next

step and a rectangular portion (face portion) was cropped automatically and saved to disk for later processing. Eigenface features were computed for each such face image stored on the disk. A mean profile image was calculated as in PCA. This process is highlighted in Figure 5.3.

Testing of the above system differed only by 10% as compared to the procedure of the training phase. Test video files passed through the key frame selector, face detector and the eigenface feature extractor but just before the final classification the test features were compared with that of the mean profile face. The classification was done based on the basis of the distance of the test features from the mean profile face developed during the training phase. This process is highlighted in Figure 5.4 displayed later in Section 5.3.

5.3.3 Fusion of Face and Speaker Identifications

Biometric systems make use of different traits for identification purposes, like fingerprints, face, voice, iris, palm-print, etc. Systems using a single trait (i.e. unimodal biometric systems) are often affected by various challenges such as the noise in sensor outputs, high error rates, temporal and spatial short-term variations, etc. A multimodal biometric system can often overcome these problems. For example, in an image-based identification systems, the lighting and posture angle can cause uncertainties in face recognition. In a voice-based identification system, speech signals can be affected by environmental noise. Multimodal biometric systems combine the traits obtained from dissimilar biometric sources and hence, provide a more reliable identification performance, in general, compared to the unimodal ones.

A biometric system usually comprises of four independent modules, which sometimes work interactively and iteratively depending on the application sphere. The raw biometric data pertaining to the end-user is first obtained using the sensor module. This data is processed by the feature extraction module so that a feature set can be constructed to summarise the representation. The previously stored user templates are then compared to these feature sets and matching scores are generated as indicators of the level of similarity. This module is called the matching module. Finally, these scores are assessed to verify or identify a certain

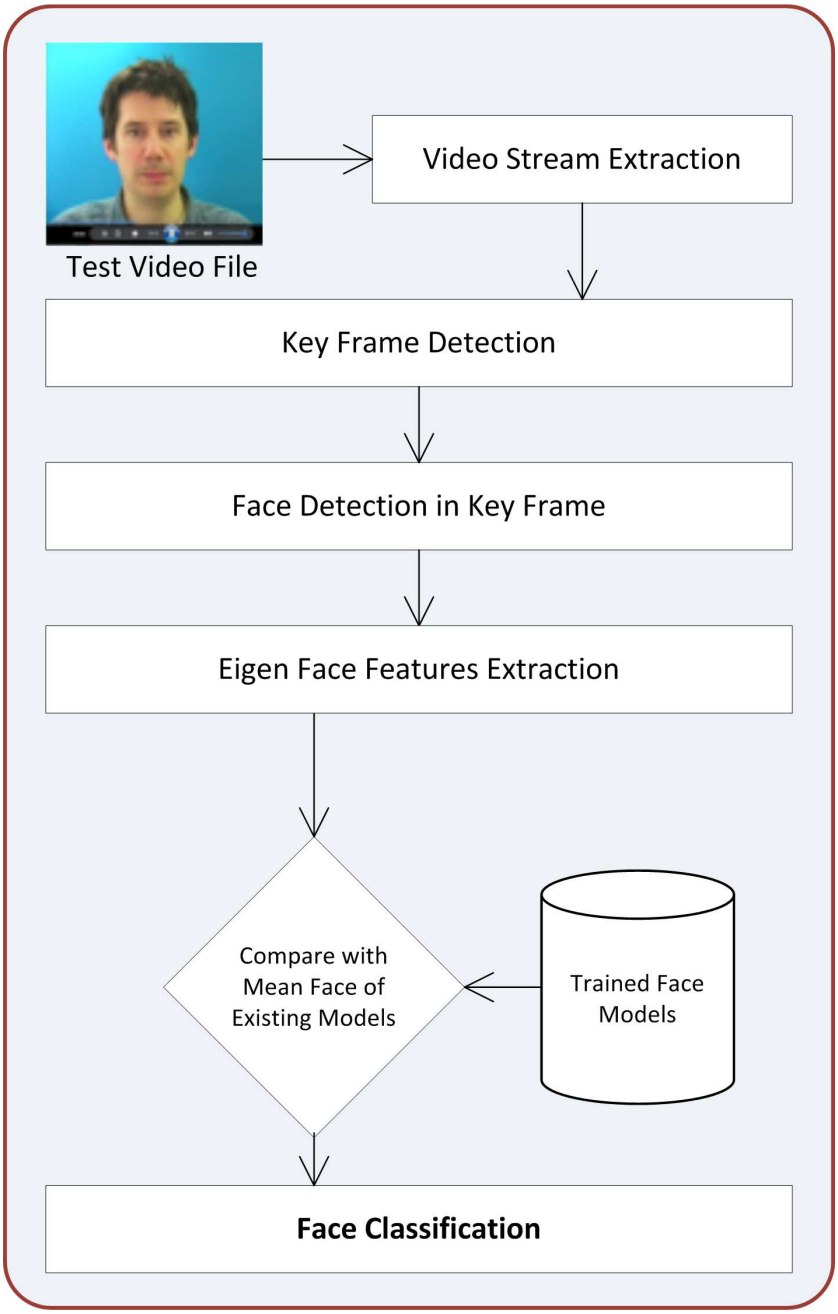

Figure 5.3: High-level information flow of face recognition during the test phase.

user, and a concluding decision is made in the decision module

In biometric systems, the fusion of information can be classified as either

pre-classification or post-classification [226]. The former is the fusion performed before any of the matching algorithms or classifiers are applied. If the information is merged after the multimodal classifiers deduce their decisions, then it is known as the post-classification fusion.

A biometric system with multiple modalities needs to combine the information it receives from the different data sources. This assimilation has to be coherent such that it becomes useful in the subsequent stages. In some cases, the integration can be performed before the matching module begins its operation. For example, if a system relies on multiple cameras capturing pictures of a face, then the pixels captured by these cameras can complement each other and increase the information content and reliability if they are merged appropriately. Since the data from the different sensors (in this example, multiple cameras) are first integrated, this is called sensor-level fusion. The data collected from each sensor can be considered as independent feature vectors, which are then merged to create a single feature vector.

For data that are homogeneous (represents the same physical property and have the same dimensions), the feature vectors from all the sensors can be simply averaged to find the merged vector; however, assigning proper weights during averaging has proven to be the preferable option. For non-homogeneous data, the merged vector can be created upon concatenation of the source vectors. Sometimes this type of fusion may apply even to a single sensor when different algorithms are run on the same sensor output yielding correlated but different feature vectors. If these vectors are integrated into one vector, it also falls in the class of sensor-level fusion. Compatibility of the data obtained from the different sensors is a prerequisite to performing sensor-level fusion.

Information integration prior to matching or classification can generally be more powerful and effective than that done afterwards. Once the classification is done, much of the raw data is discarded and only a derivative decision is used thereafter. Since the raw data has the highest amount of original content, they can contribute the most in making decisions. Therefore, integrating them before matching or classifying would ideally yield recognition results that are the most reliable. But there are several limiting factors that often prohibit systems from doing this in a practical implementation. First of all, there is an issue of

compatibility as mentioned above. Secondly, for the case of concatenation, the integrated feature vector may have very large dimensions and make it hard to process because of the associated cost and time. This is especially burdensome in the field of biometrics since almost every sensor involved results in large output data sets. In addition to that, the system needs to have knowledge of how the feature spaces of different sensors or component systems are related. If the component sensors provide data that have some portions with high correlation, it may transform into an unwanted and misleading redundancy, and therefore, needs to be trimmed out appropriately. When the system starts this sort of processing before integration, it inevitably starts losing its pre-classification traits and leaning towards a post-classification system. Finally, if one chooses to employ an off-the-shelf commercial biometric component, they may not have access to the raw data or feature vectors that these products obtain and use. Therefore, most researchers tend to study and implement post-classification fusion schemes rather than that at the feature-level.

There are many different methods for performing post-classification fusion, which can be classified into multiple categories. Fusion can be performed at the abstract, rank or matching score-levels, or it can be based on dynamic classifier selection. For a particular input stream, a classifier can provide multiple results. For each input data set, we can only choose the result that has the highest likelihood of yielding the right decision. The dynamic classifier selection works based on such classifiers. The biometric matcher may also decide on the best match based on the individual inputs it receives and perform integration at either the abstract or the decision-level. In making the final decision, it can employ something as simple as the majority voting or a modified voting scheme with unequal weights assigned to the identifiers. More complex methods (e.g. behaviour knowledge space) also exist.

For fusion at the rank level, the probable biometric matches from each matcher are first arranged with the most confident matches appearing at the top. The combined ranks are then utilised in reaching a final decision. If there is a tie, a random selection is performed so that the ranking order is stringent. The ranks can also be simply summed up (known as Borda count) or weighted and added (logistic regression). If the confidence levels of the probable biometric matches

are quantified and integrated, it is known as fusion at the matching score-level or confidence level. Since obtaining and merging these scores is a straightforward task, this is a widely popular method.

Broadly speaking, there are two distinct ways of fusion. The first approach is the classification problem. It exploits the outputs from the matching scores, constructs a feature vector, and then classifies into either "Accept" or "Reject". The second approach is the combination approach, where the individual matching scores are combined to generate a single scalar score used to make the final decision. Multimodal fusion is at the heart of any system which uses more than one modality. The choice of a fusion strategy is highly dependent on the modalities being used. Fusion techniques can also be broadly divided into the categories: early integration, intermediate integration and late integration [227; 228]. Late integration techniques use different classifiers for both modalities and combine their decisions. This combination can be decision-level fusion (AND, OR, etc.) or opinion (score-level) fusion (weighted summation, weighted product, etc.). The inherent assumption in using such techniques is that the modalities used are independent of each other. This is not the case when audio-visual modalities of speech communication are used. A person's face deforms differently depending on what is being spoken and the underlying speaking style variations. Also, such systems require separate classifiers for each modality which may complicate system design. Intermediate integration techniques use multi-stream HMMs. Although superior to late integration techniques, the inherent drawback in this technique is that it again assumes independence between the modalities used. This assumption enables it to handle audio and video streams asynchronously but some useful information correlating the two modalities are lost.

Early integration offers a natural way of integration for our problem. Feature-level fusion is a type of early integration technique. Here, we process the different modalities separately and extract appropriate features and merge them by either concatenating or by weighted summation, etc. This enables the use of a single classifier which simplifies system design. It also takes into account the correlation between two modalities inherently. A drawback of this technique is that it needs data in time synchronism. On the other hand, the hybrid technique is more suitable for the recognition task which has to be done in real-time and all possible

redundancies need to be removed. Audio and video modalities have complementary as well as redundant information. The complementary information in these modalities is usually independent and provides extra information which increases the system accuracy as well as its robustness. This redundancy can be advantageously utilised to give a high degree of robustness against many different kinds of replay attacks.

Many different classifiers have been used for audio and visual recognition over the years (including DTW, GMM, HMM, SVM and NN) and significant literature is available on them [226; 227; 228]. PCA and SVM has also been used successfully for recognising the gender of a person [229]. HMMs are widely used for speech recognition and they give high accuracy, flexibility and robustness. They can be used for speaker recognition with the same efficacy. Since our task is text-independent, we do not need to capture phone specific information. The GMMs (single state HMM) exploit this. They give a similar performance as compared to HMMs and computationally are more efficient than the HMMs. Other advantages of GMMs include low memory requirement (only means and variances need to be stored), flexibility (well suited for text-dependent as well as text-independent applications), high accuracy and robustness. Due to these reasons, we use GMMs for our classification task.

Fusion of the face and speaker identification systems is performed at the decision-level through AND voting of the opinions of both experts. Figure 5.4 illustrates the various components involved in this process flow.

5.4 Performance Testing and Results

Extensive testing and analysis has been performed on the GRID [2] and VidTIMIT [11] audio-visual corpora before reporting the results in this section. The test results presented in this section were collected on a computer with a 2.8 GHz Intel Core 2 Duo processor and 4GB of memory.

5.4.1 Tests with GRID

At first, in this section, we describe the outcomes of the comprehensive testing on the GRID corpus of video files.

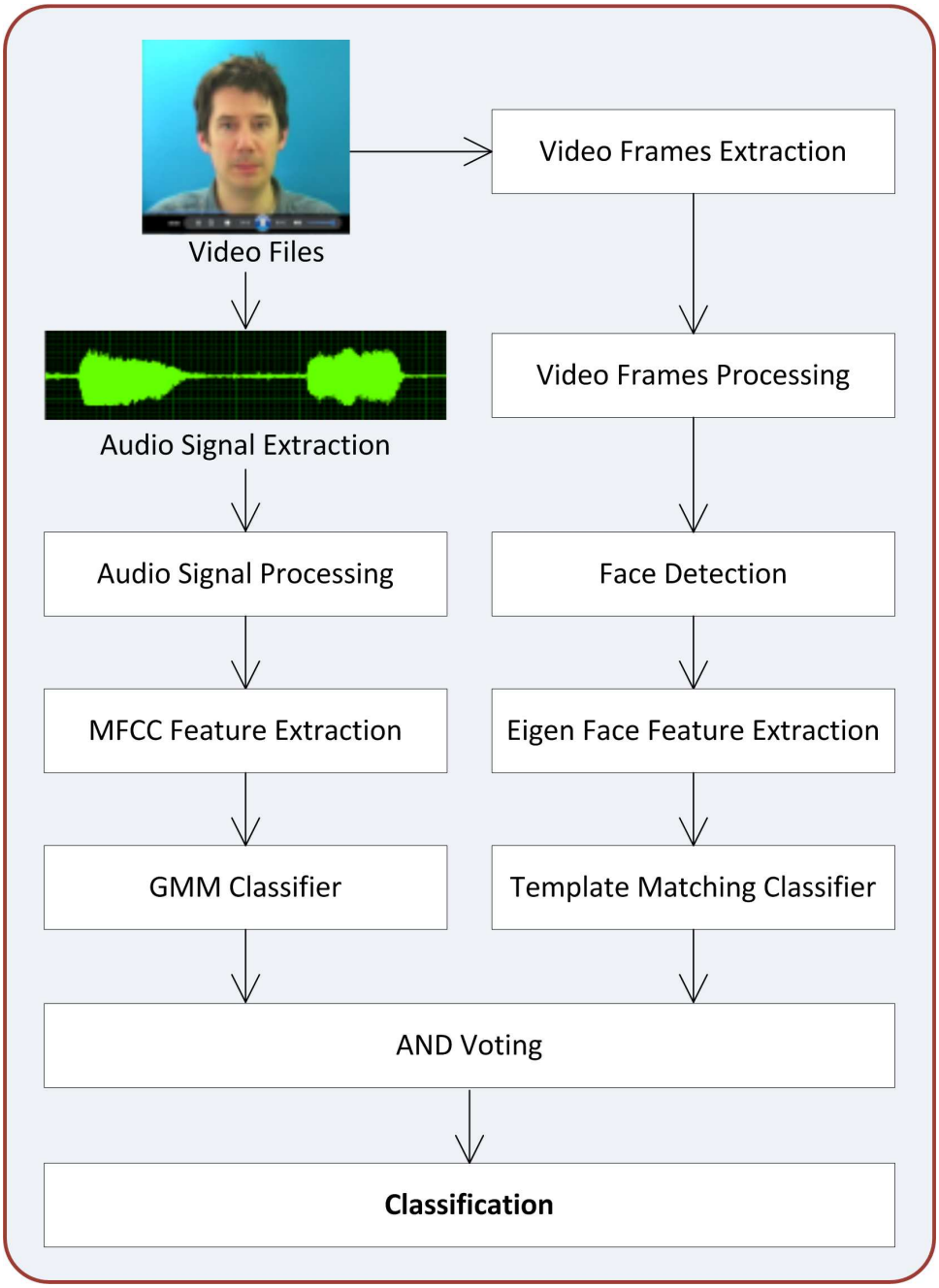

Figure 5.4: The proposed system for fusion of face and voice.

Tests were conducted on the GRID corpus video files. Files were available in low, medium and high quality and these results were calculated on the low quality video files. The data of 30 speakers in GRID were split during training and testing using 10-fold cross validation and the results were averaged. Tests were conducted with the following closed sets for comprehensive evaluation when users were increased gradually from 10 to 34.

Group A : Total Speakers = 10, training files per speaker used = 100.

Group B : Total Speakers = 10, training files per speaker used = 150.

Group C : Total Speakers = 20, training files per speaker used = 300.

Group D : Total Speakers = 34, training files per speaker used = 500.

Table 5.1 summarises the average accuracy (%) of the speaker identification experiments (face only) on the GRID [2] corpus. It can be easily inferred from the above test results that reducing the image size does not reduce the accuracy substantially. A small image size was desired to minimise redundant information and reduce the calculations to facilitate real-time implementation.

Table 5.1: Accuracy of the proposed scheme with images of different sizes.

Face Size (pixels)	**Performance Accuracy (%)**
150 x 150	98.0
50 x 50	97.9
25 x 25	97.3

Our first goal was to explore the accuracy of the individual face and voice models with the proposed GMM scheme. After separating the data into voice

and face streams, in the voice section, we tried to establish the optimum number of MFCC features for our case. Table 5.2 summarises the results obtained with the various numbers of MFCC features used. These results were collected for 20 files per speaker using 10-fold cross validation.

Table 5.2: Accuracy (%) of speaker identification using MFCC features and GMM.

Performance Accuracy (%)	**GRID**
10 MFCC features	84.0
14 MFCC features	90.08
18 MFCC features	97.8
20 MFCC features	99.3
40 MFCC features	99.34

The above experiments indicate that the number of files used for training does make a difference in the overall performance of the system but varying the number of MFCC features gradually from 10 to 20 has a minor impact on the overall accuracy of the system. Moreover, the number of MFCC features is directly proportional to computation time. 20 MFCC features were found to be the best tradeoff value between the computation time and the overall accuracy.

In the next experiments, we incorporated the above findings and used 20 MFCC features for the voice and a 25 x 25 pixel input image for the Eigenface Approach (EFA). These experiments were conducted using different numbers of speakers (S) and files per speaker (f/S) using 10-fold cross validation. The strategies included individual classifiers and their combinations using decision-level, feature-level and score-level fusion techniques. The results of these experiments

were averaged and are summarised in Table 5.3.

Table 5.3: Performance of different fusion schemes.

Scheme Used	**Performance Accuracy (%)**
GMM (voice)	98.3
PCA (face)	97.2
GMM(voice) and PCA(face) - decision-level fusion with AND Voting	97.2
GMM(voice) and PCA(face) - decision-level fusion with OR Voting	98.3
GMM(voice) and PCA(face) - weighted score-level fusion	98.7
GMM(voice) and PCA(face) - feature-level fusion	86.6
GMM(voice) and GMM(Face) - feature-level fusion	56.3
GMM (face)	56.3
PCA (voice)	69.5

The results of these experiments clearly indicate that weighted score-level fusion of PCA for face recognition and GMM with MFCC features for voice

outperforms all other approaches, as highlighted in Table 5.3. We have shown the results of GMM (voice) and PCA (face) with decision-level fusion for both AND voting and OR voting, and that with the weighted score-level fusion. The latter two approaches outperformed the AND voting and individual classifiers. Fusion of the two classifiers with OR voting gives 98.3% accurate classification and that with AND voting gives 97.2% accurate classification, both of which are comparable with the individual classifiers alone. However, weighted score-level fusion gives 98.7% accurate results (better than the individual classifiers used alone) proving it to be the most robust fusion strategy among all. This result has a logical explanation as well, since either of the two classifiers is fully capable to classify the current utterance. This combination lets the outcome be accurate even if one classifier fails to recognise the utterance successfully. At times when GMM (voice) fails, e.g. due to noise or similarities in the voices of two speakers, their faces enable our PCA (face) classifier to detect the speaker from its face alone and vice versa. This is the reason this strategy outperforms others.

Since the voice and face are two distinct modalities, combining the two at the feature-level does not give any promising results. No matter what scheme we employ, the fused feature model of the two modalities stayed disparate and inaccurate. The accuracy of the decision-level fusion with AND voting is clearly not better than the accuracy of the individual GMM (voice) or PCA (face) models as per above table. This result has a valid justification as well. AND voting limits the strong classifier and forces the classification outcome to be wrong due to AND logic with the decision with the weak classifier. So this does not improve, strictly speaking, the fusion accuracy more than the individual models. The reason behind employing GMM for voice and PCA for face for most of the schemes in Table 5.3 is evident from the last three rows of the table where we tried changing the schemes. GMM proved to be more accurate for modelling voice features as compared to modelling the PCA data, as showed in our earlier work [230]. The opposite is true for PCA which models the face features efficiently but not the voice features. Using GMM for face and PCA for voice yields only 56.3% and 69.5% performance accuracies, respectively. Using GMM for both face and voice with feature-level fusion also results in a 56.3% accuracy. This clearly justifies the choice of classifiers for voice and face.

For the weighted score-level fusion, both individual models GMM (voice) and PCA (face) were redesigned to output the classification result with a confidence interval between 0 and 1. Here, 0 means the classifier failed to verify the current utterance based on who it claimed to be. Filenames were used to claim which speaker they belonged to. Each classifier is given a weight (m) based on the its individual classification capability. In this case, GMM (voice) was 98.3% accurate individually (denoted as m_1) where PCA (face) was 97.2% accurate individually (denoted as m_2). Therefore, we used these weights to arrive at the final classification result (F),

$$F = \frac{m_1[output\ of\ GMM\ (face)] + m_2[output\ of\ PCA\ (face)]}{m_1 + m_2}. \tag{5.2}$$

A threshold of 76.2 for F was determined through trial and error, yielding the maximum accuracy of the fused model. If F falls below 76.2, the current utterance is said to be imposter/not recognisable. However, anything above $F = 76.2$ ensures that the speaker is exactly he who he claimed to be (encoded in the filename).

The Receiver Operating Characteristic (ROC) curve plotted with the results of our tests is shown in Figure 5.5. As elaborated in Section 3.5, the ROC curve shows the True Positive Rate (TPR) as a function of the False Positive Rate (FPR). This graphical plot illustrates the performance of different systems as their discrimination thresholds are varied. It clearly shows that the maximum area is covered under the curve for score-level fusion as compared to the individual classifiers on face and voice. The performance accuracies shown in Table 5.3 had demonstrated that PCA (face), GMM (voice) and the weighted score-level fusion of these two modalities are 97.2%, 98.3% and 98.7%, respectively. The same order can be observed in Figure 5.5, with PCA (face) covering the minimum area and the weighted score-level fusion covering the maximum. From a rough look at the figure, the three curves may seem insignificantly distinct. But if inspected carefully, we can see that the superiority of the score-level fusion is pronounced. A speech identification system would be far from usable if the TPR is too low or the FPR is too high. The goal is to operate with low values of FPR and high values of TPR; therefore, the upper-left portion of the figure is the practical

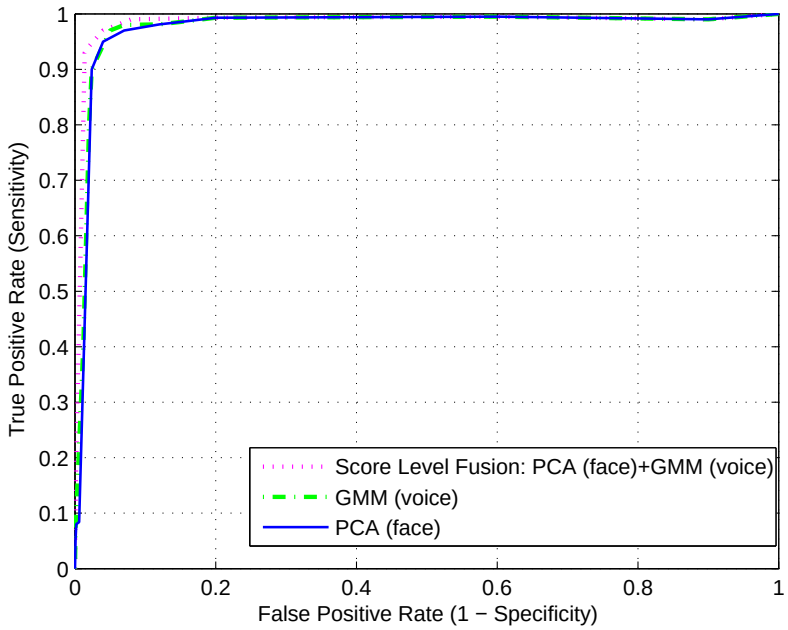

Figure 5.5: ROC curve showing the maximum area under the curve with the proposed scheme.

region of operability. It can be seen from Figure 5.5 that a 90% TPR can be achieved with the score-level fusion with only a 1.3% FPR; whereas both the individual modalities would yield about a 2.5% FPR. To achieve a 99% TPR, both PCA (face) and GMM (voice) would have a 15% FPR, but the score-level fusion would only have it 6.6% (less than half). Conversely, if the FPR was restricted to 2%, the TPR for PCA (face) or GMM (voice) would go down to 70%, while the score-level fusion would still have a 94% TPR. Therefore, this performance improvement is quite significant.

We also investigated the applicability of our proposed system, simulating a real-life scenario by conducting the same experiments with an emphasis on the recognition time required for voice and face. Herein, we experimented further to improve the accuracy of the model with respect to the number of Gaussian mixtures allowed per GMM by varying this number gradually from 1 to 16 to find the most appropriate value with minimal computation time. Only 100 files

per speaker with a total of 10 speakers were used for these experiments. Table 5.4 summarises the results of these experiments.

Table 5.4: Summary of the performance accuracy of GMM with varying number of Gaussian mixtures.

Gaussian Mixtures	**Accuracy (%) of GMM (voice)**	**Model Training Time (sec)**	**Avg. Identification Time (sec)**
1	99.3	5	1
2	100	5.2	1
4	98.2	6	1
6	98	6.9	1.2
8	94	8	1.4
16	99	10	2

From the data in Table 5.4, we conclude that increasing the number of Gaussian mixtures gradually from 1 to 16 does not necessarily increase the system accuracy. Upon further investigation, we found that many of the mixtures reduced to single points, as they did not have enough values to carry on further computation. However, the above experiment shows that 1 and 2 Gaussian mixtures provide the optimum accuracy for voice.

From the above experiments, we also made a note of the training time, feature extraction and identification time for GMM and EFA for both face and voice to see how efficient decision-level fusion is with respect to computation time. These findings are summarised in Table 5.5.

Table 5.5: Average identification time for GMM and EFA for face and voice.

	GMM (voice)	**EFA (face)**
Training Time (sec)	10	0
Identification Time (sec)	1	1
Feature Extraction Time (sec)	0.6	0.9

5.4.2 Tests with VidTIMIT

The VidTIMIT [11] database comprises audio and video recordings of 43 speakers (19 female, 24 male), reciting short sentences. The database was recorded in a noisy office environment in three different sessions using a broadcast quality video camera. There are 10 sentences recorded per person and the mean duration of the audio recording is 4.25 seconds. The video is recorded in the form of a sequence of JPEG images with a resolution of 384 x 512 pixels. Each audio is recorded as a mono, 16-bit, 32 kHz WAV file.

10-fold cross validation tests were used in the following experiments and the classification accuracy was averaged for all the 10 tests. A detected face image of 25 x 25 pixels was used while varying the number of MFCC features (audio) gradually from 10 to 44.

These tests were conducted on a Core i5, IBM Lenovo machine with 4GB of RAM, running only the Matlab [185] process on top of Windows 7. The classification accuracy is shown in Table 5.6, and the average identification times and average training times for these algorithms are shown in Table 5.7.

The ROC curves for the VidTIMIT database are showed in Figure 5.6. Similar to the GRID database, the maximum area is covered under the curve for score-level fusion as compared to the individual classifiers on face and voice. It can be seen from Figure 5.6 that a 90% TPR can be achieved with the score-level fusion

Table 5.6: Classification accuracy on VidTIMIT [11] with various features.

Scheme Used	**VidTIMIT**
PCA (face)	91.34
GMM (voice)	94.8
Decision-Level Fusion: PCA (face) and GMM(voice) with AND voting	82.8
Decision-Level Fusion: PCA (face) and GMM(voice) with OR voting	94.8
Score-Level Fusion: Weighted Average- PCA (face) and GMM(voice)	93.8

with only a 2.4% FPR; whereas the individual modalities would yield about a 4.8% (GMM) or 6.2% (PCA) FPR. Conversely, if the FPR is restricted to a certain value, the TPR for PCA (face) or GMM (voice) would go much lower than the score-level fusion. Therefore, this performance improvement is quite significant.

Table 5.7: Average training and identification time on VidTIMIT [11] for PCA and GMM.

	GMM (voice)	**EFA (face)**
Training Time (sec)	13	0
Identification Time (sec)	1.2	0.94
Feature Extraction Time (sec)	1.4	0.9

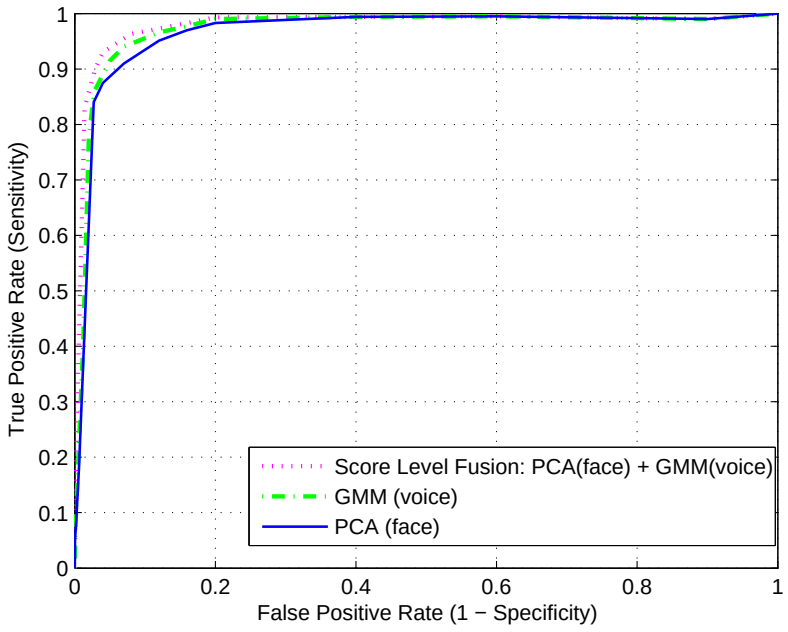

Figure 5.6: ROC curve showing the maximum area under the curve with the proposed scheme for the VidTIMIT database.

5.5 Performance Comparison with Other Systems

The recent development in human identification in other areas of the world has been inspiring and competitive. For any novel method to succeed, the comparative analysis of performance with the state-of-the-art methods is deemed necessary. In Section 5.4, we showed the comparison of different fusion systems and showed the superiority of our proposed method over other fusion methods. Table 5.8 shows the reported performance of the current methods (beyond face-voice fusion) in human identification in line with the performance accuracy of the system proposed in this research. The data in Table 5.8 show that the proposed system outperforms three of the published systems as well as the novel methods we proposed in Chapters 3 and 4.

The three published systems we compare our work to achieve some of the best identification rates available in the literature. The system in [48] is mainly based on the verification using the likelihood ratio test. The likelihood functions used some effective GMMs that are relatively simple and easy to implement. For speaker representation, it employed the Universal Background Model (UBM), from which speaker models were derived using Bayesian adaptation. The verification accomplishment was further enhanced using score normalisation. Their performance was successfully tested in several NIST speaker recognition evaluations. In [52], a robust perceptual features and iterative clustering approach is proposed for isolated digits and continuous speech recognition and speaker identification, and its evaluation is performed on clean test speeches. The training vectors are clustered into a set of book vectors using the K-means clustering algorithm [47].

A new speaker identification technique based on a modified NN was proposed in [61], which is an enhanced version of the multi-dimension SOM. This technique, namely the Multiple Parametric Self-Organising Map (M-PSOM), attempted to reduce the acceptance of impostors while maintaining a high accuracy for identification. Most of the prior systems would rely on a single NN for an entire speaker identification system, but the M-PSOM utilises parametric NNs for the individual speakers to record and depict their distinctive acoustic signa-

tures. Every speaker's voice is recorded digitally and then fragmented into 30 ms speech signals. Acoustic vectors are extracted from each of these frames after passing through an MFCC feature extraction processor. The PSOM model can, therefore, be trained using the speakers' respective pool of vectors. This paper demonstrated that this method outperforms many other competitive methods like Wavelets, GMM, HMM and VQ, in terms of the level of accuracy. Our proposed approach outperforms all three of these published systems in terms of accuracy, and therefore, proves itself as one of the best candidates for speaker identification.

The use of multiple modalities, face and voice, has resulted in the performance improvement achieved in this chapter. Combining facial-based decisions with audio-based ones allows a speaker identification system to leverage two distinct sets of features and make more robust and fault-tolerant system. This type of fusion system is good at minimising false accept rates since highly stringent criteria are used. But a fusion-based system can suffer from an increased false reject rate, and hence lose its overall accuracy, if not designed properly. In this chapter, we have carefully selected our fusion schemes and tweaked their parameters to achieve a 15% improvement in accuracy over the contemporary methods in the same field. We designed a novel scheme towards feature-level and decision-level fusion of audio and face data to construct hybrid GMM and PCA models for speaker identification. However, a fusion-based system requires the availability of video data along with audio. Many practical implementations of speaker identification systems do not have video capturing capabilities. The schemes we proposed in Chapters 3 and 4 would be useful for such cases.

5.6 Conclusion

This study utilised the feature-level, score-level and decision-level fusions of a face recognition system and a speaker identification system. It showed that both the score-level and decision-level (with OR voting) fusions can outperform the feature-level fusion in terms of accuracy and error resilience. The result is in line with the distinct nature of the two modalities which lose themselves when combined at the feature-level. We made use of the most common eigenface techniques together with face detection, grey-scale transformation and a reduced image size

Table 5.8: Performance comparison with state-of-the-art speaker identification approaches.

Reference	**Approach: Algorithm, Database**	**Performance Accuracy (%)**
[48]	GMM-UBM (independent of pre-processing algorithm), TIMIT	96.8
[52]	LPC, K-Means 8 speakers	96.37
[61]	MFCC, Parametric Neural Network, 32 speakers	90.6
Multimodal NN and Wavelets (Chapter 3)	Wavelets, Multimodal Neural Network 34 speakers (GRID) 43 speakers (VidTIMIT)	89.12
Proposed Vowel Formants with Max score	LPC, Formants, 34 speakers (GRID)	96.05
Proposed Fusion of Face and Voice	MFCC + GMM (voice), PCA (face) 34 speakers (GRID)	98.8

to successfully improve face recognition time without sacrificing accuracy. It also tweaks the GMM with MFCC for the speaker identification so much so that the overall identification time of the system stays under 1 second. The suggested scheme of improvements was successful with competitive accuracy when compared

with the state-of-the-art approaches to speaker identification as well as those that were proposed earlier in this research. The proposed scheme was found to be one of the best candidates for the fusion of face with voice due to its low computational time and high identification accuracy.

Chapter 6

Conclusions and Future Work

6.1 Introduction

The use of human identification technology will increase tremendously, keeping in perspective the current momentum of progress in hardware and related areas. This has already reached such a level that identification systems are being installed in every shop, airport, store, clinic, software house and cinema, to mention just a few. This research has investigated, proposed, implemented and tested a robust, novel algorithm for text-independent speaker identification based on a multi-resolution technique (DWT) and GRNN, PNN and RBF-NN with bagging. We have investigated a scalable speaker identification system using vowel formants and maximum score strategy. As a continuum of the findings in Chapters 3 and 4, we have proposed a robust audio-visual speaker identification system fusing face and voice modalities using GMM (voice) and EFA (face) algorithms. This research has successfully led to the submission of three international-level journal papers on the novel approaches. These approaches have been shown, after comprehensive testing, to outperform the state-of-the-art approaches for speaker identification.

This thesis has also investigated the limitations of the current methods in speaker identification and has shown ways to solve them. This final chapter summarises the major findings and the results. Section 6.2 analyses and discusses how well the overall goals, as specified in Chapter 1, have been met. Section 6.3 discusses the limitations of the research carried out, and Section 6.4 suggests

future directions for investigation in this field.

6.2 Goals Reached

The main objectives of this thesis are discussed in Chapter 1. They are derived from the limitations in the contemporary speaker identification systems as analysed in Chapter 2. Below we summarise the contributions of this research.

6.2.1 Investigation of Novel Hybrid Intelligent Methods using DWT and Neural Networks

This research has thoroughly investigated DWT with multiple neural networks including BPNN, GRNN, PNN, and RBF-NN, proposing a novel method with a bagging technique using PNN, RBF-NN, and GRNN in parallel. This technique improves speaker classification accuracy by almost 15% and identifies a speaker within one second. These feature extraction strategies were tested with quick learners with the bagging technique, utilising BPNN, GRNN, and PNN. This work has been documented in Chapter 3, with comprehensive testing on the novel idea of real-time speaker identification using multi-resolution techniques and BPNN, GRNN, and PNN. A publishable paper (submitted) has been produced out of this investigation.

6.2.2 Development of a Robust Real-time Speaker Identification System using Vowel Formants

During the course of this research, a novel method based on formant analysis applicable for highly scalable speaker identification systems with minimised database storage of features for each speaker was implemented. Tests conducted with the proposed score-based scheme confirmed a net reduction of 50% in storage data of the speakers without compromising the accuracy of speaker identification. Chapter 4 describes the design and implementation of this novel speaker identification system based on formant analysis. It also investigated various approaches to vowels detection, including LDA, PCA, and MFCC. LDA was found to be relatively more accurate on vowel detection, thus improving the overall accu-

racy. This work also included proposing a score-based algorithm for vowel-based speaker identification. Performance comparison with state-of-the-art technologies in speaker identification has shown that the proposed method has competitive accuracy and low storage requirements. A publishable paper (submitted) has been produced as a result of this investigation.

6.2.3 Investigation of Different Feature Extraction, Classification and Fusion Techniques for Audio-visual Speaker Identification

This research has thoroughly investigated individual classifiers and their combinations using decision-level, feature-level and score-level fusion techniques to construct hybrid GMM and PCA models for audio-visual speaker identification. Chapter 5 utilises the research and investigation done in previous chapters to propose this novel approach for speaker identification by fusing two different modalities of face and voice at the feature-level and decision-level. The results of these experiments clearly indicate that weighted score-level fusion of PCA for face recognition and GMM with MFCC features for voice outperforms all other state-of-the-art approaches in speaker identification. Performance testing and benchmarks conducted on the GRID [2] and VidTIMIT [11] audio-visual corpora are reported in a publishable paper (under review).

6.3 Limitations

The objectives of this research mentioned in Chapter 1 have been successfully fulfilled. However, a number of limitations and constraints have been identified during this research without which the findings and results of this thesis would be incomplete.

These limitations are reviewed in the following subsections one by one.

6.3.1 Processes Related to Face Recognition

The research carried out in this thesis has largely ignored processes related to face recognition such as face localisation, tracking, intelligent key frame detection and

segmentation. However, this research has taken care of the major processes in face recognition including face detection, key frame detection, face normalisation and grey-scale transformation in order to improve the classification accuracy. The exclusion of the former algorithms has not made much difference in the overall accuracy and applicability of the systems but for production-ready systems these algorithms may be required in order to compete with the contemporary products available in the market.

6.3.2 Database

This research has utilised the GRID [2] and VidTIMIT [11] audio-visual corpora to conduct all the tests. The GRID [2] database contains 34 speakers with 1000 video files with per speaker with British dialect. Please note that the vowel-based approach identified and investigated in Chapter 4 has a direct dependency on the dialect and related formant frequencies associated with English language vowels. For a production ready system, the developers may have to account for the local dialect of the region for which they are targeting their product. Although the rest of the research has no dependency on the dialect, ignoring this fact would produce less efficient results for vowel-based approach as documented in Chapter 4. More comprehensive databases like VidTIMIT were spotted quite late in this research; hence, they are used only in the audio-visual-based final contribution chapter for testing and result compilation. The VidTIMIT database comprises audio and video recordings of 43 speakers (19 female, 24 male), reciting short sentences. The database was recorded in a noisy office environment in three different sessions using broadcast-quality video camera. There are 10 sentences recorded per person and the mean duration of the audio recordings is 4.25 seconds. The video is recorded in the form of a sequence of JPEG images with a resolution of 384 x 512 pixels. Each audio is recorded as a mono, 16-bit, 32 kHz WAV file.

Both these corpora have less than 50 speakers. Due to time and budgetary constraints the proposed methods could not be tested on a set of commercially available large corpora. For production-ready systems, the developers may have to test the models on a number of commercially available audio-visual corpora in order to beat the market competition and improve the algorithms further.

6.3.3 Speaker Verification vs. Identification

Speaker identification has two specialised branches, namely: speaker identification and speaker verification. The algorithms developed for speaker identification can be adapted to perform speaker verification with minor code changes. This research has, in essence, focused on the speaker identification task while slightly touching on the imposter detection algorithm (part of speaker verification). Although, programmatically speaking, all the systems developed in this research have been designed for speaker identification; they can be adapted to perform speaker verification with minimal code changes (less than 2%). The algorithms designed in this research have been implemented and tested for speaker identification. The readers are suggested to adapt the algorithms with minor code changes to perform speaker verification according to their requirements.

6.3.4 Audio File Length

Chapter 4 has investigated a novel technique for speaker identification based on vowel formant analysis. The developed algorithm required 3 to 4 seconds of audio, with sufficient vowels, in order to be recognised fully. This approach may lose accuracy if tested with audio files that do not contain at least 3 vowel sounds. This research has to combine multiple one-second files for a single speaker (GRID [2] database) to produce a 4-second file in order for the algorithm to effectively detect more than 3 vowels and thus successfully classify the underlying speaker. This limitation may become a bottleneck for production-level systems in which the audio sound length is less than 4 seconds, but it can be resolved upon further research depending on the requirements.

6.3.5 Partial Images

This research has utilised partial face video recordings from the GRID [2] and VidTIMIT [11] corpora (GRID does not have pose variation datasets). This research had to ignore a large research area in face recognition focusing on pose variation. For robust face recognition which accounts for pose variation, the proposed algorithms would require further improvements since all the proposed novel methods in this research are based on partial face images/video recordings.

6.4 Future Directions

No matter what the current state-of-the-art technology becomes, there will always be room for discovering new avenues to solve the same problem or for identifying a problem that was not a known problem previously. This is a continuous quest towards evolution as progress in technology gains more momentum every day.

The following are suggestions for future research which build upon the ideas, concepts, limitations and constraints conceived during the current research.

6.4.1 Improving Face Recognition

This research has left much margin for improvement of face recognition using structural HMMs, segmentation and other multi-resolution techniques not covered in this research. For example, Curvelet and Ridgelet techniques (which were experimented with in this research but not documented due to their low accuracy) may be used in conjunction with structural HMMs and segmentation.

6.4.2 Other Measures of Performance

The current research has focused on speaker identification and has presented results in terms of percentage accuracy, identification time, training time and feature extraction time. This work has barely scratched the surface of imposter detection (part of the speaker verification task) which is more applicable to production-ready systems. However, more useful performance measures like FAR and FRR may be integrated to test the accuracy of the proposed approaches.

6.4.3 Moving to Smart Devices

Smart devices such as iPhone, Android and Blackberry, etc. are increasingly becoming more popular, and are now powerful enough to complete all the tasks that only a computer could do 10 years ago. There is a dire need for speaker identification systems to be adapted for these devices. The user should be able to identify him/her using his/her handset with a mobile application connected securely to an organisation's server systems. This would eliminate the need for extra hardware for audio and video recording at the organisation's doorstep.

6.4.4 Fusion of Other Biometrics

This research has focused on fusing face and voice for human identification. There can be a plethora of options for fusing other biometrics with voiceprint, including gait, retina and fingerprint scan. This can lead towards more robust and real time identification systems that leverage emerging technologies. However, some classifiers that are good for face may not be as good for a different biometric. For example, the features and uniqueness of people's faces are much different than their fingerprints. If voice and fingerprints are to be used for identification, there may be other classifiers that are more suitable than those used with face and voice. Moreover, using more than two features for biometrics can make things more complicated and appropriate classifiers need to be designed/identified for optimum performance. Our future research task includes a thorough investigation into such adaptabilities.

6.4.5 Fusing Gesture Signature

The fierce growth in smart devices technology has given birth to gestures like waving one's hand in the air to give input to the system to perform some action. It is possible to use gesture as an integral part of personal identification and authentication system [231; 232]. Just like every person has a different calligraphic identity for handwriting, there also exists for each person a calligraphy of gestures that is unique and personal. Gestures can be very powerful when fused with identification systems based on the audio-visual paradigm, as in the current research. By performing a simple gesture, a person can prove his/her identity to a device or authentication system. Humans can each have a specialised gesture to identify themselves along with the text-independent audio and video (image). One may coin the term Video Signatures for this research idea.

Appendix A

Pseudo Codes for Feature Extraction and Speaker Classification Algorithms

In this section, we present the feature extraction and speaker classification algorithms used in this thesis in a pseudo code fashion. These algorithms have been employed in the research and implemented in Matlab during the course of this research.

A.1 Mel Frequency Cepstral Coefficients (MFCC)

```
01. //input s = read signal file with frequency fs
02. // n is the hamming window frame size, m is the overlap between two
      hamming frames /windows
03. m ← 100
04. n ← 256
05. l ← length(s)
06. nbFrame = floor((l - n) / m) + 1
07. FOR i = 1 to n DO
08.     FOR j = 1 to nbFrame DO
09.         M(i, j) = s(((j - 1) * m) + i)
10.     END
11. END
```

```
12. //Create hamming window
13. h = hamming(n)
14. M2 = diag(h) * M
15. //compute fast fourier transform (FFT)
16. FOR i = 1 to nbFrame DO
17.       frame(:,i) = fft(M2(:, i))
18. END
19. t = n / 2
20. tmax = l / fs
21. //compute mel frequency bank filters
22. m = melfb(20, n, fs)
23. n2 = 1 + floor(n / 2)
24. z = m * abs(frame(1:n2, :)).^2
25. //take discrete cosine transform
26. r ← dct(log(z))
```

A.2 Gaussian Mixture Models (GMM)

Gaussian Mixture Models (GMM) is the state-of-the-art approach for pattern recognition. It is very popular in the speaker identification and speech recognition domain. GMM finds the probability density function based on observed feature vectors (usually supplied during training phase) and constructs a distribution to classify the incoming speech. Gaussian mixtures represent the feature space as a set of Gaussian states, each with three parameters, namely: mean, covariance and weight matrix.

Given a set of feature vectors X (computed using various feature extraction techniques, e.g. MFCC, LPC, wavelets), for a test signal, the classification problem reduces to finding the ratio of the probability that X was generated with Model Ac to the probability that X was NOT generated by the Model Yc governing all speakers. The log likelihood ratio is of particular interest to compare with tuned threshold to accept or reject a speaker model.

Here is the pseudo code for the GMM algorithm using estimated means to initialize the mean, covariance and weight matrix and MFCC to compute feature vectors from training data set of speaker files.

(i) Pseudo Code for Training: GMM for speaker identification - Generate speaker models from training data

```
01. // Load training data files for all speakers
02. FOR x=1 to Speaker m
03.     FOR y=1 to training file n
04.         Input(x,y) = Load( audio filexy )
05.         Preprocessed(x,y) = RemoveNoiseSilence(Input(x,y))
06.         MFCCFeatures(x,y) = computeMFCC(Preprocessed(x,y))
07.     END
08. END
09. // Construct GMM Speaker Models from the training data set
10. Initialize Mixtures to Mix = 16
```

11. Initialize GMM Model to an array model = [ ]
12. Initialize concatenated Feature vector to MFCC_ceps = [ ]
13. For x=1 to Speaker m
14. FOR file y = 1 to j
15. MFCC_ceps(x)= concatenate MFCC_ceps with tmp
16. END
17. Covariance_Matrix ← random diagonal covariance matrix
18. Weight_Matrix ← initialize using K-means
19. Model(x) = EMGMM (MFCC_ceps(speaker x , Covariance_Matrix, Weight_Matrix, Mix)
20. END

(ii) Pseudo Code for Testing: Classification of a speaker file whether it was generated by a specific speaker GMM

01. // Load test file
02. Test(x, frequency) ← Load(audio filex)
03. Preprocessed(x) ← RemoveNoiseSilence(Input(x))
04. MFCCFeatures(x) ← computeMFCC(Preprocessed(x,y)
05. // evaluate PDF for each mixture component
06. FOR j = 1 to speakers m
07. FOR i=1 to Mix (number of mixtures in GMM)
08. dist ← distance(X,Mean(:,i),Cov(:,:,i))
09. y(i,:) ← exp(-0.5*dist)/sqrt((2*pi) ∧ dim*det(Cov(:,:,i)))
10. END
11. y2 (m)= model(m).prior dot product with matrix y
12. //Compute Log likelihood
13. y3 (m) = Log(y2)
14. Distance(m) =-sum(y3(m))
15. END
16. Classification ← model(m) with Min Distance(m)

A.3 Principal Component Analysis (PCA) for Face Recognition

01. // M training images, sized N pixels wide by N pixels tall, c recognition images, also sized N by N pixels
02. Mp ← desired number of principal components
03. // Feature Extraction:
04. // merge column vector for each training face
05. X = [x1 x2 ... xm]
06. // compute the average face
07. me = mean(X,2)
08. A = X - [me me ... me]
09. // avoids N∧2 by N∧2 matrix computation of [V,D]=eig($A * A'$)
10. // only computes M columns of U: $A = U * E * V'$
11. [U,E,V] = svd(A,0)
12. eigVals = diag(E)
13. lmda = eigVals(1:Mp)
14. // pick face-space principal components (eigenfaces)
15. P = U(:,1:Mp)
16. // store weights of training data projected into eigenspace
17. train_wt = $P' * A$

01. Nearest-Neighbour Classification:
02. // A2 created from the recog data (in similar manner to A)
03. recog_wt = $P' * A2$
04. // euclidean distance for i-th recog face, j-th train face
05. euDis(i,j) = sqrt((recog_wt(:,j)-train_wt(:,i)).∧2)
06. // Classification
07. Classification ← Min (euDis(i,j))

A.4 Probabilistic Neural Network (PNN)

```
01. X[N] ← Training set of N vectors.
02. P[N] ← Partition of training set, described as N pointers from X to C.
03. C[M] ← Codebook of M vectors.
    a. C[j].vector = Code vector
    b. C[j].size = Cluster size
04. PerformPNN(C,P)
05. {
06. Q[BookSize(C)]: Nearest Neighbour Pointers;
    a. Q[j].nearest = Index of the nearest neighbour cluster.
    b. Q[j].distance = Cost of merging the two clusters.
    c. Q[j].recalculate = Flag indicating if the pointer must be updated.
07. FOR j=1 TO BookSize(C) DO
08.     {
09.     Q[j] ← FindNearestNeighbour(C,j);
10.     }
11. WHILE BookSize(C) > M DO
12.     {
13.     a ← FindMinimumDistance(C,Q);
14.     b ← Q[a].nearest;
15.     MergeVectors(C,P,Q,a,b);
16.     UpdatePointers(C,Q);
17.     }
18. }

19. FindNearestNeighbour(C,a)
20. {
21. q: Nearest Neighbour Pointer structure;
22. q.nearest ← 1;
23. q.distance ← Infinite;
24. q.recalculate ← NO;
25. FOR j=1 TO BookSize(C) DO
```

```
26.     {
27.     d ← MergeDistortion(C[a],C[j]);
28.     IF a<>j AND d<q.distance THEN
a.        {
b.        q.nearest ← j;
c.        q.distance ← d;
d.        }
29.     }
30. return q;
31. }

32. FindMinimumDistance(C,Q)
33. {
34. MinDist ← Infinite;
35. MinIndex ← 1;
36. FOR j=1 TO BookSize(C) DO
37.     {
38.     IF Q[j].distance < MinDist THEN
a.        {
b.        MinIndex ← j;
c.        MinDist ← Q[j].distance;
d.        }
39.     }
40. return MinIndex;
41. }

42. MergeVectors(C,P,Q,a,b)
43. {
44. IF a > b THEN Swap(a,b); /* So that a is smaller index. */
45. last = BookSize(C);
46. MarkClustersForRecalculation(C,Q,a,b);
47. C[a].vector ← Centroid(C[a].vector,C[b].vector);
48. JoinPartitions(P,C,a,b);
```

```
49. FillEmptyPosition(C,Q,b,last);
50. DecreaseBookSize(C);
51. }

52. MarkClustersForRecalculation(C,Q,a,b)
53. {
54. FOR j=1 TO j < BookSize(C) DO
55.     {
56.     IF Q[i].nearest=a OR Q[i].nearest=b THEN Q[i].Recalculate = YES;
a.      ELSE Q[i].Recalculate = NO;
57.     }
58. Q[a].Recalculate = YES;
59. }

60. JoinPartitions(P,C,a,b)
61. /* The partitions a and b are joined so that all vectors will be
62.    in a and cluster b will be empty. Pointers are updated. */
63. {
64. FOR i=1 TO N DO
65.     {
66.     IF P[i]=b THEN P[i] ← a;
67.     }
68. C[a].size ← C[a].size + C[b].size;
69. }

70. FillEmptyPosition(C,Q,b,last)
71. /* Merging two vectors create empty place in the codebook. In this
72.    routine the empty place is filled by the last entry in the codebook. */
73. {
74. IF b<>last THEN
75.     {
76.     C[b] ← C[last];
77.     Q[b] ← Q[last];
```

```
78.      FOR j=1 TO j < BookSize(C) DO
a.         {
b.         IF Q[j].nearest=last THEN
c.            {
d.            Q[j].nearest ← b;
d.            }
e.         }
f.     }
79. }

80. UpdatePointers(C,Q)
81. {
82. FOR j=1 TO j<BookSize(C) DO
83.      {
84.      IF Q[j].recalculate=YES THEN
a.         {
b.         Q[j] ← FindNearestNeighbour(C,j);
c.         Q[j].recalculate ← NO;
d.         }
85.      }
86. }

87. MergeDistortion(c1,c2):
88. Calculates the merge cost of the cluster using equation (3) in the paper.

89. Centroid(c1,c2):
90. Calculates the weighted average of the two input vectors.
91. BookSize(C):
92. Returns the number of vectors in the (current) code book C.
93. DecreaseBookSize(C):
94. Decreases the counter (the size of the codebook) by one.
```

A.5 Back Propagation Neural Network

-inputs:
X (inputs), Y (outputs), η (learning rate), T (iterations)

-output: ANN

01. Begin
02. FOR t = 1 to T DO
03. FOR i=1 to N DO
04. FOR each output Unit k
05. Δ k = (yik - Outputx)
06. END
07. FOR each hidden Unit h
08. Δ h = Outputh(1 - Outputh) $\Sigma_{k \epsilon DownStream} \Delta$ k W_{kh}
09. END
10. Update every $W_{ji} = W_{ji} + \eta \Delta_j X_{ji}$
11. END
12. END
13. RETURN ANN

A.6 Speaker Identification with Vowel Formants using LPC

```
01. // Load training data files for all speakers
02. Initialize English language Vowel Formants frequency filter in F
03. For x=1 to Speaker m
04.     FOR y=1 to training file n
05.         Input(x,y) = Load( audio filexy )
06.         Preprocessed(x,y) = RemoveNoiseSilence(Input(x,y))
07.         FormantsMap(x,y) = LPCFormants(Preprocessed(x,y))
08.         EnglishVowels(x,y) = FilterEnglishVowels(FormantsMap(x,y),F)
09.     END
10. END
11. // Test a current utterance
12. S ← loadfile(x)
13. Preprocessed(X)= RemoveNoiseSilence(S)
14. FormantsMAP(X)= LPCFormants(Preprocessed(X))
15. For each vowel k in FormantsMap(X)
16.     FOR each vowel p in EnglishVowels(speaker X)
17.         If k==p
18.             Model(speaker X).vote++
19.         END
20.     END
21. END
22. Classification ← SPEAKER with MAX (Model.Vote)
```

References

[1] H. S. Jayanna and S. R. M. Prasanna, "An experimental comparison of modelling techniques for speaker recognition under degraded condition," *Springer*, vol. 34, no. 5, pp. 717–728, 2009. xvii, 21, 22, 23, 24, 25

[2] M. Cooke, J. Baker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *Acoustical Society of America*, vol. 120, no. 3, pp. 2421–2424, 2006, gRID. xvii, 4, 7, 9, 10, 23, 33, 66, 75, 77, 83, 84, 101, 103, 104, 110, 111, 112, 119, 121, 137, 138, 139

[3] J. Matas and J. Sochman, "Adaboost," *Computer Vision Lectures, Oxford University Robotics Research Group.*, 2004. xvii, 37

[4] L. Wiskott, J. Fellous, N. Kuiger, and C. Von Der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775–779, 1997. xvii, 39, 40, 43

[5] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001. xvii, 39, 40

[6] J. Moody and C. J. Darken, "Fast learning in networks of locally-tuned processing units," *Neural Computation*, vol. 1, no. 2, pp. 281–294, 1989. xvii, xviii, 29, 70, 72, 74, 82, 83

[7] A. Amrouche and J. Rouvaen, "Efficient system for speech recognition us-

ing general regression neural network," *International Journal of Intelligent Technology*, vol. 1, no. 2, pp. 183–189, 2006. xvii, 72, 73

[8] Mathwork. (2013, *http://www.mathworks.com/help/techdoc/ref/mmreaderclass.html*, September) *Mathworks*. xvii, 73

[9] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transaction on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994. xviii, 88, 89

[10] U. C. L. (UCL). (2013, *http://www.phon.ucl.ac.uk/home/wells/formants/table-1-uni.htm*, September) Dept. of Phonetics and Linguistics, UK. xix, 91, 96

[11] S. Conrad, "The VidTIMIT database. IDIAP Communication," 2002. xix, xx, 10, 33, 54, 58, 82, 103, 104, 111, 119, 128, 129, 130, 137, 138, 139

[12] A. N. Hoshyar and R. Sulaiman, "Review on finger vein authentication system by applying neural network," in *Information Technology (ITSim), 2010 International Symposium*, vol. 2. IEEE, 2010, pp. 1020–1023. 1

[13] F. Besbes, H. Trichili, and B. Solaiman, "Multimodal biometric system based on fingerprint identification and iris recognition," in *Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference.* IEEE, 2008, pp. 1–5. 1

[14] D. Park, S. Kim, T. H. Shin, and J. Sou, "A security framework in RFID multi-domain system. ," in *The Second International Conference Availability, Reliability and Security.* IEEE, 2007, pp. 1227–1234. 2

[15] B. Debnath, R. Ranjan, F. Alisherov, and M. Choi, "Biometric Authentication: A Review," *International Journal of u- and e- Service, Science and Technology*, vol. 2, no. 3, pp. 13–28, September 2009. 2

[16] T. Kinnunen, E. Karpov, and P. Franti, "Real-time speaker identification and verification," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, pp. 277–288, September 2006. 2, 3, 5

[17] A. Shafik, S. M. Elhalafawy, S. M. Diab, B. M. Sallam, and F. E. Abd El-samie, "A wavelet based approach for speaker," *International Journal of Communication Networks and Information Security (IJCNIS)*, vol. 1, no. 3, pp. 52–58, 2009. 3

[18] J. Wu and S. Ye, "Driver identification based on voice signal using continuous wavelet transform and artificial neural network techniques," *Expert Systems with Applications*, vol. 36, no. 2, pp. 1061–1069, 2009. 4

[19] J. Wu and B. Lin, "Speaker identification using discrete wavelet packet transform technique with irregular decomposition expert system," *Expert Systems with Applications: An International Journal*, vol. 36, no. 2, pp. 3136–3143, 2009. 4, 16, 68

[20] E. Avci and Z. H. Akpolat, "Speech recognition using a wavelet packet adaptive network based fuzzy inference system," *Expert Systems with Applications*, vol. 31, no. 3, pp. 495–503, 2006. 4

[21] J. Martinez, H. Perez, E. Escamilla, and M. Suzuki, "Speaker recognition using Mel frequency cepstral coefficients (MFCC) and vector quantization (VQ) techniques," in *22nd International Conference Electrical Communications and Computers (CONIELECOMP)*. IEEE, 2012. 4

[22] D. J. Mashao and M. Skosan, "Combining classifier decisions for robust speaker identification," *Pattern Recognition*, vol. 39, no. 1, pp. 147–155, 2006. 9

[23] J. J. Sroka and L. D. Braida, "Human and machine consonant recognition," *Speech Communication*, vol. 45, no. 4, pp. 401–423, 2005. 9

[24] S. Ben-Yacoub, Y. Abdeljaoued, and E. Mayoraz, "Fusion of face and speech data for person identity verification," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1065–1074, 1999. 9, 59, 105

[25] A. Higgins, L. Bahler, and J. Porter, "Speaker verification using randomized phrase prompting," *Digital Signal Processing*, vol. 1, no. 3, pp. 89–106, 1991. 12

[26] J. D. Markel and S. B. Davis, "Text-independent speaker recognition from a large linguistically unconstrained time-spaced database," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 27, no. 1, pp. 74–82, March 1979. 12

[27] R. Sarikaya, B. L. Pellom, and J. H. Hansen, "Wavelet packet transform features with application to speaker identification," in *IEEE Nordic Signal Processing Symposium*. Citeseer, 1998, pp. 81–84. 13, 68

[28] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 1448–1460, March 2007. 13

[29] A. G. Adami and D. A. C. Barone, "A speaker identification system using a model of artificial neural networks for an elevator application," *Information Sciences*, vol. 138, no. 1, pp. 1–5, 2001. 13

[30] A. Haydar, M. Demirekler, and M. K. Yurtseven, "Speaker identification through use of features selected using genetic algorithm," *Electronics Letters*, vol. 34, no. 1, pp. 39–40, 1998. 13

[31] C. Wutiwiwatchai, V. Achariyakulporn, and C. Tanprasert, "Text-dependent speaker identification using LPC and DTW for Thai language," in *IEEE Region 10 Conference (TENCON)*, vol. 1. IEEE, 1999, pp. 674–677. 13

[32] R. Vogt and S. Sridharan, "Explicit modeling of session variability for speaker verification," *Computer Speech Language*, vol. 22, no. 1, pp. 17–38, 2008. 13

[33] D. Hosseinzadeh and S. L. Krishnan, "Combining vocal source and MFCC features for enhanced speaker recognition performance using GMMs," in *Proc. Int. Conf. on Signal Processing*. IEEE, 2007, pp. 365–368. 13

[34] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," *International Speech Communication Association (ISCA)*, 2001. 15

[35] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Processing Magazine*, vol. 11, no. 4, pp. 18–32, 1994. 15

[36] M. Afify and O. Siohan, "Comments on vocal tract length normalization equals linear transformation in cepstral space," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1731–1732, 2007. 15

[37] J. S. Walker, Ed., *A Primer on Wavelets and Their Scientific Applications*, ser. CRC Press LLC. Kentucky 41042-2919, USA: Taylor and Francis, 2008, vol. 2. 15

[38] K. Daqrouq, W. Al-Sawalmeh, A. Al-Qawasmi, and I. N. Abu-Isbeih, "Speaker identification wavelet transform based method," in *5th International Multi-Conference on Systems, Signals Devices*. IEEE, 2008, pp. 20–23. 16

[39] S. Y. Lung, "Wavelet feature selection based neural networks with application to the text independent speaker identification," *Pattern Recognition*, vol. 39, no. 3, pp. 1518–1521, 2006. 16

[40] S. Nawab and T. Quatieri, "Short-time Fourier transform," *Advanced Topics in Signal Processing*, vol. 6, no. 2, pp. 289–337, 1988. 16

[41] C. Garcia, G. Zikos, and G. Tziritas, "Wavelet packet analysis for face recognition," *Image and Vision Computing*, vol. 18, no. 4, pp. 289–297, 2000. 22, 43

[42] W. Li and X. Zhu, "A new image fusion algorithm based on wavelet packet analysis and PCNN," in *International Conference on Machine Learning and Cybernetics*, vol. 9. IEEE, 2005, pp. 5297–5301. 22

[43] A. Croisier, D. Esteban, and C. Galand, "Perfect channel splitting by use of interpolation/decimation/tree decomposition techniques," in *International*

Conference on Information Sciences and Systems, vol. 2. Patras, Greece, 1976, pp. 443–446. 23

[44] R. Crochiere, S. Webber, and J. Flanagan, "Digital coding of speech in sub-bands," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1976, pp. 233–236. 23

[45] Y. Zhao, X. Lei, and F. Zhonghua, "Mask estimation and refinement for MFT-based robust speaker verification," in *13th Annual Conference of the International Speech Communication Association*. INTERSPEECH, 2012, pp. 2654–2657. 25

[46] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002.* c, 2002, pp. 4072–4075. 26, 30

[47] A. Srinivasan, "Speaker identification and verification using vector quantization and Mel frequency cepstral coefficients," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 4, no. 1, pp. 33–40, 2012. 27, 131

[48] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 3, pp. 19–41, 2000. 27, 28, 30, 57, 102, 131, 133

[49] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995. 28

[50] S. Memon, "Speaker verification based on different vector quantization techniques with Gaussian mixture models," in *Network and System Security*. IEEE, October 2009, pp. 403–408. 28

[51] R. V. Pawar, P. P. Kajave, and S. N. Mali, "Speaker identification using neural networks," in *International Enformatika Conference (IEC)*. World Academy of Science, Engineering and Technology, 2005, pp. 429–433. 28, 30, 81, 82

[52] A. Revathi, R. Ganapathy, and Y. Venkataramani, "Text independent speaker recognition and speaker independent speech recognition using iterative clustering approach," *International Journal of Computer Science & Information Technology*, vol. 1, pp. 30–40, 2009. 29, 30, 57, 131, 133

[53] K. Daqrouq, "Wavelet entropy and neural network for text-independent speaker identification," *Engineering Applications of Artificial Intelligence*, vol. 24, no. 5, pp. 796–802, 2011. 29

[54] K. Daqrouq, I. N. Abu-Isbeih, O. Daoud, and E. Khalaf, "An investigation of speech enhancement using wavelet filtering method," *International Journal of Speech Technology*, vol. 13, no. 2, pp. 101–115, 2010. 29

[55] J. N. Gowdy and Z. Tufekci, "Mel-scaled discrete wavelet coefficients for speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3. IEEE, 2000, pp. 1351–1354. 29

[56] X. Yuanyou, X. Yanming, and Z. Ruigeng, "An engineering geology evaluation method based on an artificial neural network and its application," *Engineering geology*, vol. 47, no. 1, pp. 149–156, 1997. 29

[57] D. F. Specht, "A general regression neural network," *IEEE Transactions on Neural Networks*, vol. 2, no. 6, pp. 568–576, 1991. 29, 69, 70, 72

[58] T. D. Ganchev, D. K. Tasoulis, M. N. Vrahatis, and N. D. Fakotakis, "Generalized locally recurrent probabilistic neural networks with application to text-independent speaker verification," *Neurocomputing*, vol. 70, no. 7, pp. 1424–1438, 2007. 29

[59] S. Chakroborty and G. Saha, "Improved text-independent speaker identification using fused MFCC and IMFCC feature sets based on Gaussian filter," *International Journal of Signal Processing*, vol. 5, pp. 11–19, 2009. 31, 57

[60] R. Saeidi, P. Mowlaee, T. Kinnunen, and Z. H. Tan, "Signal-to-signal ratio independent speaker identification for co-channel speech signals," in

Proceedings of International Conference on Pattern Recognition (ICPR). IEEE, 2010, pp. 4565–4568. 31, 57

[61] P. Gomez, "A text independent speaker recognition system using a novel parametric neural network," in *Proceedings of International Journal of Signal Processing, Image Processing and Pattern Recognition*, 2011, pp. 1–16. 31, 57, 131, 133

[62] S. F. Galton, "Personal identification and description-II," *Nature*, vol. 38, pp. 201–203, 1888. 32

[63] W. Zhao, "Robust image based 3D face recognition," Ph.D. dissertation, University of Maryland at College Park College Park, MD, USA, 1999. 32

[64] A. K. Jain, B. Klare, and U. Park, "Face recognition: Some challenges in forensics," in *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, 2011, pp. 726–733. 32

[65] O. Arandjelovic, "Unfolding a face: From singular to manifold," in *Proc. Asian Conference on Computer Vision*. Springer Lecture Notes in Computer Science, 2009, pp. 203–213. 32

[66] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, pp. 137–154, 2009. 32, 37, 111

[67] I. Paliy, "Face detection using Haar-like features cascade and convolutional neural network," in *Proceedings of International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science*. IEEE, 2008, pp. 375–377. 35

[68] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, pp. 589–591, 1991. 35

[69] R. Baron, "Mechanisms of human facial recognition," *International Journal of Man-Machine Studies*, vol. 15, no. 2, pp. 137–178, 1981. 35

[70] R. Brunelli and T. Poggio, "Face recognition: Features versus templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 10, pp. 1042–1052, 1993. 35

[71] M. Bartlett, J. Movellan, and T. Sejnowski, "Face recognition by independent component analysis," *IEEE Transactions on Neural Networks*, vol. 13, no. 6, pp. 1450–1464, 2002. 35, 36, 43

[72] P. N. Belhumeur, J. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997. 35, 36

[73] Y. Barniv and D. Casasent, "Multisensor image registration: Experimental verification," in *Proceedings of the SPIE 292*, 1981, pp. 160–171. 35

[74] P. Comon, "Independent component analysis, a new concept?" *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994. 36

[75] C. Liu and H. Wechsler, "Comparative assessment of independent component analysis (ICA) for face recognition," in *International conference on audio and video based biometric person authentication*. Citeseer, 1999. 36, 43

[76] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," *JOSA A*, vol. 14, no. 8, pp. 1724–1733, 1997. 36, 43

[77] M. Do and M. Vetterli, "The contourlet transform: An efficient directional multiresolution image representation," *IEEE Transactions on Image Processing*, vol. 14, no. 12, pp. 2091–2106, 2005. 37, 43

[78] Y. Freund and R. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *Computational learning theory*. Springer, 1995, pp. 23–37. 37

[79] A. Demiriz, K. Bennett, and J. Shawe-Taylor, "Linear programming boosting via column generation," *Machine Learning*, vol. 46, no. 1-3, pp. 225–254, 2002. 37

[80] Y. Freund, "An adaptive version of the boost by majority algorithm," *Machine learning*, vol. 43, no. 3, pp. 293–318, 2001. 37

[81] G. Guo and H. Zhang, "Boosting for fast face recognition," in *IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*. IEEE, 2001, pp. 96–100. 37

[82] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995. 38

[83] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: An application to face detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 1997, pp. 130–136. 38

[84] B. Scholkopf, A. Smola, and K. Muller, "Kernel principal component analysis," *Artificial Neural Networks*, pp. 583–588, 1997. 38

[85] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Mullers, "Fisher discriminant analysis with kernels," in *IEEE Signal Processing Society Workshop: Neural Networks for Signal Processing*. IEEE, 1999, pp. 41–48. 38

[86] Q. Liu, R. Huang, H. Lu, and S. Ma, "Face recognition using kernel-based fisher discriminant analysis," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2002, pp. 197–201. 38

[87] F. Bach and M. Jordan, "Kernel independent component analysis," *The Journal of Machine Learning Research*, vol. 3, pp. 1–48, 2003. 38

[88] T. Martiriggiano, M. Leo, P. Spagnolo, and T. D'Orazio, "Facial feature extraction by kernel independent component analysis," in *IEEE Conference on Advanced Video and Signal Based Surveillance*. IEEE, 2005, pp. 270–275. 38

[89] Q. Liu, J. Cheng, H. Lu, and S. Ma, "Modeling face appearance with nonlinear independent component analysis," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2004, pp. 761–766. 38

[90] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, 1994. 38

[91] G. W. Cottrell and M. Fleming, "Face recognition using unsupervised feature extraction," in *Int. Neural Network Conf*, 1990, pp. 322–325. 39

[92] Z. Zhao, D. Huang, and B. Sun, "Human face recognition based on multi-features using neural networks committee," *Pattern Recognition Letters*, vol. 25, no. 12, pp. 1351–1358, 2004. 39

[93] S. Lawrence, C. L. Giles, A. Tsoi, and A. Back, "Face recognition: A convolutional neural-network approach," *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 98–113, 1997. 39

[94] M. J. Er, W. Chen, and S. Wu, "High-speed face recognition based on discrete cosine transform and RBF neural networks," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 679–691, 2005. 39, 43

[95] L. Wiskott, J. Fellous, N. Kuiger, and C. Von Der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775–779, 1997. 39

[96] G. Edwards, T. Cootes, and C. Taylor, "Face recognition using active appearance models," in *Computer Vision, ECCV*. Springer, 1998, pp. 581–595. 40

[97] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989. 40

[98] F. Samaria and S. Young, "HMM-based architecture for face identification," *Image and Vision Computing*, vol. 12, no. 8, pp. 537–543, 1994. 40

[99] A. Nefian and M. Hayes III, "Hidden Markov models for face recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5. IEEE, 1998, pp. 2721–2724. 41

[100] L. Bai and L. Shen, "Combining wavelets with HMM for face recognition," in *Applications and Innovations in Intelligent Systems XI*. Springer, 2004, pp. 227–233. 41

[101] I. Daubechies et. al., "Wavelet transforms and orthonormal wavelet bases," *Different Perspectives on Wavelets*, vol. 47, pp. 1–33, 1993. 41

[102] M. Bicego, U. Castellani, and V. Murino, "Using hidden Markov models and wavelets for face recognition," in *12th International Conference on Image Analysis and Processing*. IEEE, 2003, pp. 52–56. 41

[103] H.-S. Le and H. Li, "Recognizing frontal face images using hidden Markov models with one training image per person," in *17th International Conference on Pattern Recognition*, vol. 1. IEEE, 2004, pp. 318–321. 41

[104] F. S. Samaria, "Face recognition using hidden Markov models," Ph.D. dissertation, University of Cambridge, 1994. 41

[105] H. Othman and T. Aboulnasr, "A separable low complexity 2D HMM with application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1229–1238, 2003. 41

[106] S. Fine, Y. Singer, and N. Tishby, "The hierarchical hidden Markov model: Analysis and applications," *Machine learning*, vol. 32, no. 1, pp. 41–62, 1998. 41

[107] J. Kittler et. al., "3D assisted face recognition: A survey of 3D imaging, modelling and recognition approachest," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2005, pp. 114–114. 41

[108] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1063–1074, 2003. 41

[109] A. Bronstein, M. Bronstein, and R. Kimmel, "Expression-invariant 3D face recognition," in *Audio-and Video-Based Biometric Person Authentication*. Springer, 2003, pp. 62–70. 41

[110] S. Helgason, *The Radon Transform*. Springer, 1999, vol. 5. 42

[111] A. Webb and G. Kagadis, "Introduction to biomedical imaging," *Medical Physics*, vol. 30, p. 2267, 2003. 42

[112] A. Kadyrov and M. Petrou, "The trace transform and its applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 8, pp. 811–828, 2001. 42

[113] S. Srisuk, M. Petrou, W. Kurutach, and A. Kadyrov, "Face authentication using the trace transform," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2003, pp. I–305. 42

[114] D. P. Huttenlocher, G. A. Klanderman, and W. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993. 42

[115] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002. 42

[116] J. Matas et. al., "Comparison of face verification results on the XM2VTFS database," in *15th International Conference on Pattern Recognition*, vol. 4. IEEE, 2000, pp. 858–863. 42

[117] C. Park and H. Park, "Fingerprint classification using fast Fourier transform and nonlinear discriminant analysis," *Pattern Recognition*, vol. 38, no. 4, pp. 495–503, 2005. 42

[118] J. Lai, P. Yuen, and G. Feng, "Face recognition using holistic Fourier invariant features," *Pattern Recognition*, vol. 34, no. 1, pp. 95–109, 2001. 42

[119] G. Wallace, "The JPEG still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992. 42

[120] S. Eickeler, S. Muller, and G. Rigoll, "Recognition of JPEG compressed face images based on statistical methods," *Image and Vision Computing*, vol. 18, no. 4, pp. 279–287, 2000. 42

[121] Z. Hafed and M. Levine, "Face recognition using the discrete cosine transform," *International Journal of Computer Vision*, vol. 43, no. 3, pp. 167–188, 2001. 42

[122] B. Achermann, "The face database of University of Bern," *Institute of Computer Science and Applied Mathematics, University of Bern*, 1995. 42

[123] L. Zhao, Y. Cai, J. Li, and X. Xu, "Face recognition based on discrete cosine transform and support vector machine," in *International Conference on Neural Networks and Brain*, vol. 2. IEEE, 2005, pp. 1248–1252. 43

[124] A. Amira and P. Farrell, "An automatic face recognition system based on wavelet transforms," in *IEEE International Symposium on Circuits and Systems*. IEEE, 2005, pp. 6252–6255. 43

[125] J. Chien and C. Wu, "Discriminant waveletfaces and nearest feature classifiers for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1644–1649, 2002. 43

[126] M. Lades et. al., "Distortion invariant object recognition in the dynamic link architecture," *IEEE Transactions on Computers*, vol. 42, no. 3, pp. 300–311, 1993. 43

[127] C. Liu, "Gabor-based kernel PCA with fractional power polynomial models for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 5, pp. 572–581, 2004. 43

[128] P. Yang, S. Shan, W. Gao, S. Li, and D. Zhang, "Face recognition using ADA-boosted Gabor features," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2004, pp. 356–361. 43

[129] E. J. Candes, "Ridgelets: Theory and applications," Ph.D. dissertation, Stanford University, 1998. 43

[130] E. Candes and D. Donoho, "Curvelets: A surprisingly effective nonadaptive representation for objects with edges," DTIC Document, Tech. Rep., 2000. 43

[131] Y. Ariki and W. Ishikawa, "Integration of face and speaker recognition by subspace method," in *Pattern Recognition, Proceedings of the 13th International Conference. Dept. of Electron. & Inf.*, 1996, pp. 456–460. 44, 58

[132] W. Adams and G. Iyengar et. al., "Semantic indexing of multimedia content using visual, audio, and text cues," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 2, pp. 170–185, 2003. 45

[133] D. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, 1997. 45

[134] C. Snoek, M. Worring, and A. Smeulders, "Early versus late fusion in semantic video analysis," in *ACM international conference on Multimedia*. ACM, 2005, pp. 399–402. 45, 48, 107

[135] Y. Wang, Z. Liu, and J. Huang, "Multimedia content analysis-using both audio and visual clues," *IEEE Signal Processing Magazine*, vol. 17, no. 6, pp. 12–36, 2000. 46

[136] R. Yan, "Probabilistic models for combining diverse knowledge sources in multimedia retrieval," Ph.D. dissertation, IBM, 2006. 46

[137] A. Nefian and L. Liang et. al., "Dynamic bayesian networks for audio-visual speech recognition," *EURASIP Journal on Advances in Signal Processing*, vol. 2002, no. 11, pp. 1274–1288, 2002. 48

[138] G. Chetty and M. Wagner, "Audio visual speaker verification based on hybrid fusion of cross modal features," in *Pattern Recognition and Machine Intelligence*. Springer, 2007, pp. 469–478. 48

[139] ——, "Investigating feature-level fusion for checking liveness in face-voice authentication," in *Eighth International Symposium on Signal Processing and Its Applications*, vol. 1. IEEE, 2005, pp. 66–69. 48

[140] P. Atrey, M. Kankanhalli, and J. Oommen, "Goal-oriented optimal subset selection of correlated multimedia streams," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 3, no. 1, p. 2, 2007. 49

[141] G. Iyengar, H. Nock, and C. Neti, "Audio-visual synchrony for detection of monologues in video archives," in *International Conference on Multimedia and Expo*, vol. 1. IEEE, 2003, pp. I–329. 49, 106

[142] A. Bendjebbour and Y. Delignon et. al., "Multisensor image segmentation using Dempster-Shafer fusion in markov fields context," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 8, pp. 1789–1798, 2001. 50

[143] J. Ni, X. Ma, L. Xu, and J. Wang, "An image recognition method based on multiple BP neural networks fusion," in *International Conference on Information Acquisition*. IEEE, 2004, pp. 323–326. 50

[144] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998. 50

[145] H. Nock, G. Iyengar, and C. Neti, "Assessing face and speech consistency for monologue detection in video," in *ACM International conference on Multimedia*. ACM, 2002, pp. 303–306. 50

[146] M. Kankanhalli, J. Wang, and R. Jain, "Experiential sampling in multimedia systems," *IEEE Trans. Multimedia*, vol. 8, no. 5, pp. 937–946, 2006. 50, 51, 106

[147] G. L. Foresti and L. Snidaro, "A distributed sensor network for video surveillance of outdoor environments," in *International Conference on Image Processing*, vol. 1. IEEE, 2002, pp. I–525. 51, 105, 106

[148] M. Yang, S. Wang, and Y. Lin, "A multimodal fusion system for people detection and tracking," *International Journal of Imaging Systems and Technology*, vol. 15, no. 2, pp. 131–142, 2005. 51, 106

[149] C. Neti and B. Maison et. al., "Joint processing of audio and visual information for multimedia indexing and human-computer interaction." in *RIAO*, 2000, pp. 294–301. 51, 106

[150] V. Radova and J. Psutka, "An approach to speaker identification using multiple classifiers," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2. IEEE, 1997, pp. 1135–1138. 51

[151] N. Pfleger, "Context based multimodal fusion," in *International conference on Multimodal interfaces*. ACM, 2004, pp. 265–272. 51

[152] S. Ayache, G. Quenot, and J. Gensel, "Classifier fusion for SVM-based multimedia semantic indexing," in *Advances in Information Retrieval*. Springer, 2007, pp. 494–504. 51

[153] V. Pitsikalis, A. Katsamanis, G. Papandreou, and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation." in *INTERSPEECH*, 2006. 52

[154] H. Xu and T. Chua, "Fusion of AV features and external information sources for event detection in team sports video," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, vol. 2, no. 1, pp. 44–67, 2006. 52, 107

[155] P. Atrey, M. Kankanhalli, and R. Jain, "Information assimilation framework for event detection in multimedia surveillance systems," *Multimedia systems*, vol. 12, no. 3, pp. 239–253, 2006. 52

[156] R. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960. 52

[157] A. Loh, F. Guan, and S. Ge, "Motion estimation using audio and video fusion," in *Control, Automation, Robotics and Vision Conference*, vol. 3. IEEE, 2004, pp. 1569–1574. 52

[158] J. Luettin, "Visual speech and speaker recognition," Ph.D. dissertation, University of Sheffield, 1997. 52, 53

[159] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner, "Acoustic-labial speaker verification," *Pattern Recognition Letters*, vol. 18, no. 9, pp. 853–858, 1997. 53

[160] R. Kuhn, P. Nguyen, J. C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, and K. Field, "Eigenfaces and Eigenvoices: Dimensionality reduction for specialized pattern recognition," in *IEEE Second Workshop on Multimedia Signal Processing*, 1998, pp. 71–76. 54, 58

[161] F. X. Tieyan, L. L. Xing, L. Hong, P. Xiaobo, and A. V. Nefian, "Audio-visual speaker identification using coupled hidden Markov models," in *International Conference on Image Processing, Vol.2*. IEEE, 2003, pp. 29–32. 54, 59

[162] H. Cheng, Y. Chao, S. L. Yeh, C. S. Chen, H. M. Wang, and Y. P. Hung, "An efficient approach to multimodal person identity verification by fusing face and voice information," in *IEEE International Conference on Multimedia and Expo, ICME*, 2005, pp. 542–545. 54, 60

[163] P. H. Lee, L. J. Chu, Y. P. Hung, S. W. Shih, C. S. Chen, and H. S. Wang, "Cascading multimodal verification using face, voice and iris information," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2007, pp. 847–850. 55, 61

[164] K. A. Lee, C. You, H. Li, and T. Kinnunen, "A GMM-based probabilistic sequence kernel for speaker verification," in *Proc. Interspeech 2007 (ICSLP)*. IEEE, 2007. 55

[165] K. Ban, K. C. Kwak, H. S. Yoon, and Y. K. Chung, "Fusion technique for user identification using camera and microphone in the intelligent service robots," in *IEEE International Symposium on Consumer Electronics*. IEEE, 2007, pp. 1–6. 55, 61, 104

[166] P. Kartik, R. Vara Prasad, and P. S. R. Mahadeva, "Noise robust multimodal biometric person authentication system using face, speech and signature features," in *INDICON 2008*. IEEE, 2008, pp. 23–27. 55, 62, 104

[167] A. Shukla, R. Tiwari, H. K. Meena, and R. Kala, "Speaker identification using wavelet analysis and modular neural networks," *Journal of Acoustic Society of India (JASI)*, vol. 36, no. 1, pp. 14–19, 2009. 55

[168] J. Kittler and K. Messer, "Fusion of multiple experts in multimodal biometric personal identity verification systems," in *12th IEEE Workshop on Neural Networks for Signal Processing*. IEEE, 2002, pp. 3–12. 60

[169] S. Mallat, *A Wavelet Tour of Signal Processing*. Access Online via Elsevier, 1999. 67

[170] S. Y. Lung and C. Chen, "Further reduced form of Karhunen-Loeve transform for text independent speaker recognition," *Electronics Letters*, vol. 34, no. 14, pp. 1380–1382, 1998. 67

[171] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding*. Prentice Hall PTR Englewood Cliffs, New Jersey, 1995, vol. 87. 67

[172] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Transactions on Image Processing*, vol. 1, no. 2, pp. 205–220, 1992. 67

[173] S. Mallat and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 7, pp. 710–732, 1992. 67

[174] S. Mallat, "Zero-crossings of a wavelet transform," *IEEE Transactions on Information Theory*, vol. 37, no. 4, pp. 1019–1033, 1991. 67

[175] Z. Tufekci and J. N. Gowdy, "Feature extraction using discrete wavelet transform for speech recognition," in *Proceedings of the IEEE Southeastcon*. IEEE, 2000, pp. 116–123. 67

[176] C. J. Long and S. Datta, "Wavelet based feature extraction for phoneme recognition," in *Proceedings of Fourth International Conference on Spoken Language (ICSLP)*, vol. 1. IEEE, 1996, pp. 264–267. 67

[177] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980. 68

[178] S. Y. Lung, "Feature extracted from wavelet Eigenfunction estimation for text-independent speaker recognition," *Pattern recognition*, vol. 37, no. 7, pp. 1543–1544, 2004. 68

[179] ——, "Improved wavelet feature extraction using kernel analysis for text independent speaker recognition," *Digital Signal Processing*, vol. 20, no. 5, pp. 1400–1407, 2010. 68

[180] C. T. Chen, S. Y. Lung, C. F. Yang, and M. C. Lee, "Speaker recognition based on 80/20 genetic algorithm," in *IASTED International Conference on Signal Processing, Pattern Recognition and Application, Greece*, 2002, pp. 547–549. 68

[181] K. S. Nathan and H. F. Silverman, "Time-varying feature selection and classification of unvoiced stop consonants," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3, pp. 395–405, 1994. 68

[182] D. Avci, "An expert system for speaker identification using adaptive wavelet sure entropy," *Expert Systems with Applications*, vol. 36, no. 3, pp. 6295–6300, 2009. 68

[183] W. Lu, W. Sun, and H. Lu, "Robust watermarking based on DWT and non-negative matrix factorization," *Computers & Electrical Engineering*, vol. 35, no. 1, pp. 183–188, 2009. 68

[184] B. Xiang and T. Berger, "Efficient text-independent speaker verification with structural Gaussian mixture models and neural network," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 447–456, 2003. 69

[185] J. C. Wells, "A study of the formants of the pure vowels of British English," Master's thesis, University of London, UK, 1962. 80, 92, 99, 110, 112, 128

[186] M. N. Stuttle and M. J. F. Gales, "Combining a Gaussian mixture model front end with MFCC parameters." in *INTERSPEECH*, 2002. 81

[187] B. S. Atal, "The history of linear prediction," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 154–161, 2006. 86

[188] G. S. University. (2013, *http://hyperphysics.phy-astr.gsu.edu/hbase/music/vowel2.html*, September) Hyperphysics education website. 87, 90

[189] M. J. Miles, "Speaker identification based upon an analysis of vowel sounds and its applications to forensic work." Master's thesis, University of Auckland, Australia, 1989. 87

[190] J. P. Campbell, T. E. Tremain, and V. C. Welch, "The DoD 4.8 Kbps standard (proposed federal standard 1016)," *Advances in Speech Coding - The Springer International Series in Engineering and Computer Science*, vol. 114, no. 1, pp. 121–133, 1991. 89

[191] L. Rabiner and B. H. Juang, Eds., *Fundamentals of Speech Recognition*. New Jersey, USA: Pearson Education, 2008. 89

[192] R. Mammone, X. Zhang, and R. Ramachandran, "Robust speaker recognition: A feature based approach," *Signal Process Magazine*, vol. 13, no. 1, pp. 58–71, 1996. 90

[193] G. Fant, *Acoustic theory of speech production*. Walter de Gruyter, 1970, no. 2. 90

[194] P. Ladefoged and S. F. Disner, *Vowels and consonants*. Wiley. com, 2012. 90

[195] M. Ter Keurs, J. M. Festen, and R. Plomp, "Effect of spectral envelope smearing on speech reception II," *The Journal of the Acoustical Society of America*, vol. 93, p. 1547, 1993. 90

[196] T. Baer and B. C. J. Moore, "Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech," *The Journal of the Acoustical Society of America*, vol. 95, p. 2277, 1994. 90

[197] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, p. 2670, 1994. 90

[198] R. Drullman, J. M. Festen, and T. Houtgast, "Effect of temporal modulation reduction on spectral contrasts in speech," *The Journal of the Acoustical Society of America*, vol. 99, p. 2358, 1996. 90, 91

[199] Q. J. Fu and G. Nogaki, "Noise susceptibility of cochlear implant users: The role of spectral resolution and smearing," *Journal of the Association for Research in Otolaryngology*, vol. 6, no. 1, pp. 19–27, 2005. 91

[200] C. Liu and Q. J. Fu, "Estimation of vowel recognition with cochlear implant simulations," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 1, pp. 74–81, 2007. 91

[201] H. Kuwabara, K. Takeda, Y. Sagisaka, S. Katagiri, S. Morikawa, and T. Watanabe, "Construction of a large-scale Japanese speech database and its management system," in *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1989, pp. 560–563. 92

[202] P. D. Templeton and B. J. Guillemin, "Speaker identification based on vowel sounds using neural networks," in *Proceedings of the 3rd International Conference on Speech Science and Technology*. Australian Association, 1990, pp. 280–285. 96

[203] M. Wagner, "Speaker identification using glottal-source waveforms and support-vector-machine modelling," in *SST 2012*. Macquarie University, Sydney, Australia, 2012, pp. 49–52. 97

[204] M. J. Han, J. H. Hsu, K. T. Song, and F. Y. Chang, "A new information fusion method for SVM-based robotic audio-visual emotion recognition," in

IEEE International Conference on Systems, Man and Cybernetics. IEEE, 2007, pp. 2656–2661. 104

[205] Y. M. Zhang, L. Ma, and B. Li, "Face and ear fusion recognition based on multi-agent," in *International Conference on Machine Learning and Cybernetics*, vol. 1. IEEE, 2008, pp. 46–51. 104

[206] T. S. Ibiyemi, J. Ogunsakin, and S. A. Daramola, "Bi-modal biometric authentication by face recognition and signature verification," *International Journal of Computer Applications*, vol. 42, no. 20, pp. 17–21, 2012. 104

[207] T. S. Ibiyemi, "Automatic face recognition by computer," *Abacus: Mathematics Series*, vol. 30, no. 2B, 2003. 104

[208] ——, "On computation of optimum basis vector for face detection and recognition," *Abacus: Mathematics Series*, vol. 29, no. 2, 2002. 104

[209] H. Soliman, A. S. Mohamed, and A. Atwan, "Feature level fusion of palm veins and signature biometrics." *International Journal of Video, Image Processing and Network Security*, vol. 12, no. 1, 2012. 104

[210] H. AlMahafzah, M. Imran, and H. Sheshadri, "Multibiometric: Feature level fusion using FKP multi-instance biometric," *International Journal of Computer Science Issues*, vol. 9, no. 3, 2012. 105

[211] K. Bowyer, K. Chang, and P. Yan et. al., "Multi-modal biometrics: an overview," in *Second Workshop on Multi-Modal User Authentication*, 2006. 105

[212] R. Brunelli and D. Falavigna, "Person identification using multiple cues," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 17, no. 10, pp. 955–966, 1995. 105

[213] R. Duin, "The combining classifier: to train or not to train?" in *Int. Conf. Pattern Recognition*, vol. 2. IEEE, 2002, pp. 765–770. 105

[214] L. Hong and A. Jain, "Integrating faces and fingerprints for personal identification," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1295–1307, 1998. 105

[215] A. Ross and A. Jain, "Information fusion in biometrics," *Pattern Recognition Letters*, vol. 24, no. 13, pp. 2115–2125, 2003. 105

[216] X. Hua and H. Zhang, "An attention-based decision fusion scheme for multimedia information retrieval," in *Advances in Multimedia Information Processing*. Springer, 2005, pp. 1001–1010. 106

[217] K. McDonald and A. Smeaton, "A comparison of score, rank and probability-based fusion methods for video shot retrieval," in *Image and video retrieval*. Springer, 2005, pp. 61–70. 107

[218] R. Yan, J. Yang, and A. Hauptmann, "Learning query-class dependent weights in automatic video retrieval," in *International conference on Multimedia*. ACM, 2004, pp. 548–555. 107

[219] A. Corradini, M. Mehta, N. Bernsen, J. Martin, and S. Abrilian, "Multimodal input fusion in human-computer interaction," *NATO Science Series Sub Series III Computer and Systems Sciences*, vol. 198, p. 223, 2005. 107

[220] J. R. Jang, "Audio signal processing and recognition," 2011. 109

[221] N. A. Meseguer, "Speech analysis for automatic speech recognition," *Norwegian University of Science and Technology, Master's Thesis*, 2009. 109

[222] W. Holmes, *Speech Synthesis and Recognition*, 2001. 109

[223] A. Girgensohn, J. Boreczky, and L. Wilcox, "Keyframe-based user interfaces for digital video," *Computer*, vol. 34, no. 9, pp. 61–67, 2001. 111

[224] T. Moeslund, A. Hilton, and V. Kruger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 90–126, 2006. 111

[225] M. Cooper and J. Foote, "Discriminative techniques for keyframe selection," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2005, pp. 4–pp. 111

[226] C. Sanderson and K. Paliwal, "Polynomial features for robust face authentication," in *Int. Conf. Image Processing*, vol. 3. IEEE, 2002, pp. 997–1000. 116, 119

[227] P. Aleksic and A. Katsaggelos, "Audio-visual biometrics," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 2025–2044, 2006. 118, 119

[228] C. Sanderson, "Automatic person verification using speech and face information," Ph.D. dissertation, Griffith University, 2007. 118, 119

[229] L. Bui, D. Tran, X. Huang, and G. Chetty, "Face gender recognition based on 2D principal component analysis and support vector machine," in *Fourth International Conference on Network and System Security*, 2010, pp. 579–582. 119

[230] N. Almaadeed, A. Aggoun, and A. Amira, "Audio-visual feature fusion for speaker identification," in *19th International Conference on Neural Information Processing (ICONIP), Part I, Doha, Qatar*, 2012, pp. 56–67. 124

[231] A. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001. 141

[232] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. IEEE, 2005, pp. 65–72. 141