EXTRACTION OF ACCENT INFORMATION FROM URDU SPEECH FOR FORENSIC SPEAKER RECOGNITION

by

Falak Tahir

Supervised By

Dr. Hanif Zauq

Co-Supervised By

Dr. Sajid Saleem

Submitted for partial fulfilment of the requirements of the degree of MSCS to the Faculty of Engineering and Computer Science

NATIONAL UNIVERSITY OF MODERN LANGUAGES,

ISLAMABAD

FEBRUARY 2019

EXTRACTION OF ACCENT INFORMATION FROM URDU SPEECH FOR FORENSIC SPEAKER RECOGNITION

by

Falak Tahir

Supervised By

Dr. Hanif Zauq

Co-Supervised By

Dr. Sajid Saleem

Submitted for partial fulfilment of the requirements of the degree of MSCS to the Faculty of Engineering and Computer Science

NATIONAL UNIVERSITY OF MODERN LANGUAGES,

ISLAMABAD

FEBRUARY 2019

THESIS AND DEFENSE APPROVAL FORM

The undersigned certify that they have read the following thesis, examined the defence, are satisfied with overall exam performance, and recommend the thesis to the Faculty of Engineering and Computer Sciences.

THESIS TITLE: Extraction of Accent Information from Urdu Speech for Forensic Speaker Recognition

Submitted By: Falak Tahir

Registration #: 7 MS/MCS-S16

Master of Science
MSCS

Computer Science
Name of Discipline

Dr.Hanif Zauq
Name of Research Supervisor

Signature: _____

Dr.Sajid Saleem
Name of Co-Supervisor

Signature: _____

Dr. Muhammad Akbar
Name of Dean (FE&CS)

Signature: _____

Brig. Muhammad Ibrahim
Name of Director General (NUML)

Signature: _____

14th February, 2019

CANDIDATE DECLARATION

I declare that this thesis entitled *"Extraction of Accent Information from Urdu Speech for Forensic Speaker Recognition"* is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : _____

Name : Falak Tahir

Date : February 14, 2019

ABSTRACT

This thesis presents a new method for extraction of accent information from Urdu speech signals. Accent is used in speaker recognition system especially in forensic cases and plays a vital role in identifying people of different groups, communities and origins due to their different speaking styles. Other applications of accent are telephone banking, voice dialing, e-health and biometric authentication. This thesis focuses on only the forensic applications of the accent. Forensic detection through accent helps in criminal investigation and provides additional information such as territorial origins of the suspects.

The proposed method is based on Gaussian Mixture Model-Universal Background Model (GMM-UBM) and a new Feature Mapping (FM) process. The proposed method is named as GMM-FM. The FM process maps Mel-Frequency Cepstral Coefficients (MFCC) features to higher dimensional space and improves the accent extraction and forensic speaker recognition performances of GMM-UBM.

In the proposed method, GMM-UBM is used to obtain accent independent model. For this purpose the training MFCC features of the training set are processed with the proposed FM method. The processed features of all the accent categories of the training set are combined and different GMM components are computed with GMM-UBM. Each GMM component is parameterized by a mean vector, mixture weight and covariance matrix.

In the second step, the GMM components estimated for accent independent model are used in a Bayesian process to adapt GMM components for each accent category of the training set. Such GMM components are referred to as accent dependent GMM. To classify accent in a speech sample the log-likelihood is computed using the GMMs of both accent dependent and independent models. Then accent is predicted for the test sample based on maximizing the log-likelihood values.

Experiments are performed on Urdu and Kaggle accent corpuses. The experimental results show that the proposed GMM-FM obtains on average 2.5% and 3.5% better equal error rate and accuracy than GMM-UBM, respectively.

Keywords: Accent, Urdu corpus, speech signals, Gaussian components, speech features, recognition and forensic.

DEDICATION

This thesis work is dedicated to my parents and my teachers throughout my education career who have not only loved me unconditionally but whose good examples have taught me to work hard for the things that I aspire to achieve.

ACKNOWLEGEMENT

First of all, I wish to express my gratitude and deep appreciation to Almighty Allah, who made this study possible and successful. This study would not be accomplished unless the honest espousal that was extended from several sources for which I would like to express my sincere thankfulness and gratitude. Yet, there were significant contributors for my attained success and I cannot forget their input, especially my research supervisors, Dr. Hanif Zauq and Dr. Sajid Saleem and my teacher Dr. Fazli Subhan who did not leave any stone unturned to guide me during my research journey.

I shall also acknowledge the extended assistance from the administrations of Department of Computer Sciences who supported me all through my research experience and simplified the challenges I faced. For all whom I did not mention but I shall not neglect their significant contribution, thanks for everything.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

LPC	Linear Predictive Coding
LPCC	Linear Predictive Cepstral Coefficient
MFCC	Mel-frequency Cepstral Coefficient
SDC	Shifted Delta Coefficient
KNN	K-Nearest Neighbor
SVM	Support Vector Machine
I-vector	Identity Vector
GMM	Gaussian Mixture Model
GMM-UBM	Gaussian Mixture Model-Universal Background Model
DCT	Discrete Cosine Transform
EER	Equal Error Rate
EM	Expectation Maximization
RBF	Radial Basis Function

CHAPTER 1

INTRODUCTION

1.1 Overview

Pronunciations varieties of a spoken language are known as accent. Accent refers to the sound and speaking style of a person [1]. Every person has a different accent. Accent generally refers to ways of pronouncing of a language words within a community. Remarkable attempts have been made to automatically identify the accent through speaker's utterances [2]. Recognition of accent prior to automatic speech recognition (ASR) helps in improving the performance of the system [3].

1.2 Accent Recognition

A pronouncing pattern of a speaker differentiates the speaker from other speakers [4]. Each speaker has different speaking style and resulting speech signals. The speech signals also carry other information about the speaker such as fitness, age, mental state, emotional state, gender, geographical and territorial origin [5, 6]. Other then linguistic variations like feelings, sensations, fitness, and generation the speaker uniqueness is the blend of his physiologic and lifestyle sense [6].

Accent recognition is used as a task to differentiate two or more persons on the basis of their accent [7]. Accent recognition has a wide range of applications such as telephone-based assistant systems [8], transcription [9], education [10], non-education [11], assistive living [12] and e-health [13]. Another application of accent is in forensic analysis such as speaker profiling, prison call monitoring and biometric authentication [14]. The forensic analysis is used by law enforcement agencies to identify the person through his accent [15, 16].

Audio forensics is a term used in literature to analyse audio recordings that may ultimately be presented as evidence in a court of law or some other official venue [17]. In other words, audio forensics refers to the use of scientific knowledge

and automation techniques for investigation and establishing of authentication or verification in the courtroom [18]. In forensic cases it is important to recognize the voice that listener overhear [19]. Speaker recognition in forensic cases often becomes difficult if distortions and noise occur. Distortions or manipulation of audio are due to transmission medium like telephone channel, recording material, or the speaker himself [20]. But by recognizing the accent before speaker recognition the performance of the system can be boosted [8].

Forensic speaker recognition is implemented by comparing the unknown recording with known recordings [20]. Accent in questioned recordings is recognized and provides additional information to recognize the suspect. The similarity of a questioned recording with a suspected one is known as evidence in forensic cases. Experts of forensic are anxious about the proof and evidence. They are not concerned with the liability or righteous of the accused person. The liability or righteous of the accused person is the decision of the court [20].

Figure 1.1 shows a general method which is used for accent recognition. The speech signals of different accents are recorded and the speech features are extracted. The training models are trained on speech features. In testing phase, a test signal is passed through same feature extraction step and provided to the training model to predicts the accent and then accuracy of the training model is computed.

Success of accent based forensic speaker recognition depends upon the quality of data and the training of the classifiers [17, 21, 22]. In the literature the accent recognition has been carried out on Chinese [2], English [19], French [23, 23] and other international languages [24]. Some researchers focus on English language only and use regional accents like American English and British language [25]. Some studies analyze regional accents within Britain [2, 7, 25, 26]. In all these studies, the speech corpus is first constructed. Corpuses are some time text-dependent where different speakers records the same text or sentences in their native accents. Some corpuses are text-independent [27], where different speakers record different text and sentences in their native accents. In text-dependent systems the utterances used in the training and testing phases are same [28]. Whereas in text-independent systems the utterances used in training and testing phases are different [29].

Some studies focuses on Asian languages like Hindi [30], Bahasa [17], Mandarin [31] and Tamil [31]. In contrast this thesis focuses on the progress and development of Urdu language, especially in the area of accent recognition. Urdu

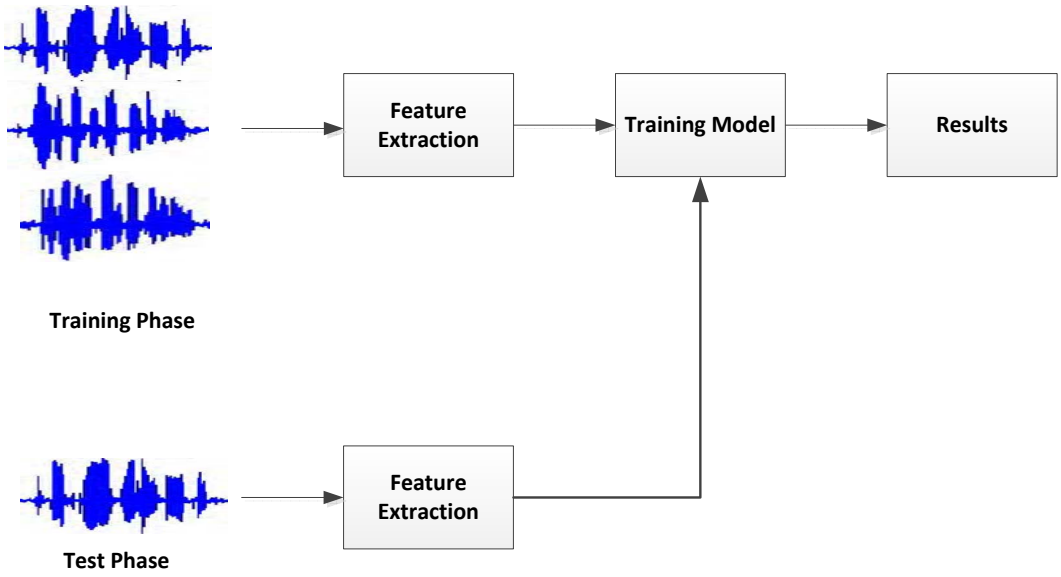


Figure 1.1: General block diagram used for accent recognition

is national language of Pakistan. Pakistan is a multilingual country. The purpose of this research is to analyze the influence of accents of regional language on Urdu. In Pakistan, people communicate and understand more than two languages including Urdu. Urdu is understandable by most of the Pakistanis. Although, Urdu is national language but people have different first language. Accents of native languages affect the accent of Urdu [32]. Majorly spoken regional/native languages are Punjabi, Sindhi, Balochi, and Pashto. Therefore, the Urdu accent varies geographically across Pakistan. These accent variations are due to pronunciation of Urdu words using the rule of regional languages.

In literature, the Urdu accent recognition is investigated on text dependent system in controlled environment [9, 32, 33]. Such systems employ a set of pre-defined sentences such as district names to generate a sequence of utterances to capture the accent. Then for recognition the acoustic similarity between the accents produced by the speakers is measured. The controlled environment is noise less environment with zero interruption and recordings are taken in a laboratory where there is no other voice than that of speaker [26].

In contrast this research focuses on text and speaker independent Urdu accent recognition. To implement such a system a corpus is constructed that contains Urdu utterances of native speakers of Punjabi, Sindhi, Balochi and Pashto languages. This research is beneficial for accent recognition especially in forensic analysis to identify the person's ethnicity and the territorial origin.

1.3 Motivation

Work available on Urdu accent recognition is either text or speaker dependent. There is a need to investigate text and speaker independent accent recognition and use the accent for forensic speaker recognition. Urdu is national language of Pakistan. Urdu accent is dominated by the accent of regional languages like Punjabi, Sindhi, Balochi and Pashto. Available work and datasets on Urdu accent are speaker and text dependent. These kinds of datasets do not provide variety of speech samples for analysis and accent recognition. For better results and to improve the performance of recognition system analysis of accent on speech and text independent dataset is required.

1.4 Research Questions

i. Given a speech utterance, is it possible to recognize the Urdu accent in text and speaker independent scenario?

ii. What is the best speech feature that can be used for Urdu accent recognition?

iii. What is the best machine learning technique for Urdu accent recognition?

1.5 Contributions

The main contributions of this thesis are as follows:

i. A new method for extraction of Urdu accent from speech signals

ii. A comparative study of state of the art features for Urdu accent recognition

iii. A performance evaluation of machine learning techniques for Urdu accent recognition

iv. Two new speech corpuses for Urdu accent recognition

1.6 Thesis Structure

Chapter 2 presents related work. It describes the baseline system and provides an overview of state of the art speech features and classifiers for accent recognition. Chapter 3 presents a detailed discussion on speech features, classifiers and components of accent recognition system. Chapter 4 presents the proposed method. Chapter 5 presents the experimental setup and results. Finally, the thesis is concluded in Chapter 6 with discussion, conclusion and future work.

CHAPTER 2

RELATED WORK

2.1 Overview

This chapter presents a comprehensive overview of existing accent recognition schemes. It briefly describes prominent features and classifiers that have been used for accent recognition. It also describes the accent based applications, especially using the accent in forensic cases.

2.2 Literature Review

The background of native speaker participates crucially in the speaker identification system [25]. The context of a language makes it more simple to distinguish a person's culture and territorial background [25]. The pattern in the speech signal carries information about speaker's geographical and social background. Information like trace of gender, age, region, background, or education of speaker are extracted from speech signal [5]. In forensic cases a person is recognized on the basis of his native language. Accent helps to determine the person background and ethnicity [34]. Every language has special pronunciation styles known as accents or dialects. Accents vary from language to language, person to person, and region to region [35].

2.2.1 Accent Based Speaker Recognition

Identifying the accent before the speaker identification improves the performance of the speaker recognition systems. Natural speech is a vocalized form of human language [30]. It is a primary means of communication between the people. Every language has unique sound structure, grammar syntax, and intonation pattern, which makes it distinct. Accent varies in the tone of voice, pronunciation of vowels, consonants, stress and prosody. That's why every accent differs from one another [36].

A deeper understanding of accent is done in [37]. Accent in the speech is

automatically identified and the speaker is recognized. Accent identification helps to understand the person background more clearly. A number of evaluation metrics are also used to evaluate the accent identification tasks. Similarly a frame work for forensic speaker is proposed in [12, 38]. These frameworks are used to identify the person for judiciary scenarios.

In [37] accent variations in native and foreign language are investigated. Author's focuses on the affects of listener's accent background. Speech comprehensibility is studied and its relation to accent perception is investigated. Listener's may sometime get the wrong idea about person accent due to his false perception [37].

In [30] accent recognition is implemented in a control environment. All interruptions or background noise are excluded from the recordings. Text-independent samples of one person are used. Forensic speech scientists are interested in fetching the information about an unknown speaker in a recording. With the help of accent recognition system, they are able to analyse such cases. The fetched information is used to identify the speaker, that either the speaker belongs to that recording or not [26].

2.2.2 Native and Non-native Accent Recognition

Variations in accuracy of a listener, identifying the accent is investigated for accent identification of native and non-native speakers [39]. Similarly accent identification in native and foreign listeners is explored in [22]. Identifying native accent over foreign one is comparatively simple. The social effect is carefully examined with additional data like voices samples. Listener social network provides additional information for accent identification.

It is found that the knowledge of the target language influence the accent recognition experiments. This is due to accent variations in the native and foreign listeners with knowledge of native language that identifies a native language better than foreign language. This study is conducted on different languages such as German, English, Chinese and Spanish [22].

A study on children speech is carried out to understand foreign accent and to identify the accent variations that is different from the native accent [12]. Children pronunciation is more complicated as compared to adults, but their accent is more accurate comparatively. As they only speak their mother tongue and use their native

accent. Two situations are observed, a new approach for automatically recognizing dialect and accent and a phonetic based kernel approach. The experimental results show that more consistent accent recognition accuracy is obtained with a perceptual learning.

The perceptual learning deals with learning better perception skills like differentiating two musical tones or categorizing patterns. Phonetic based kernel approaches use a phone recognizer for extracting GMM super-vectors for each phone kind, and then summarize the phonetic characteristics of speaker in one vector. By using the vectors, a kernel function is trained. The phonetic uniformity within pairs of speeches is evaluated with kernel function for training SVM classifiers. These similarities are used to identify accents obtaining phone hypotheses [40, 41].

2.2.3 Frameworks for Accent Recognition

Likelihood ratio framework is used for forensic analysis in automatic speaker recognition system [38]. This framework quantifies the strength of voice evidence. The system uses a database to learn speaker variability and speech information. The system's effectiveness is tested with two other databases. The experimental results indicate that automatic acoustic features produce acceptable results and facilitate in the evidence analysis [38].

In a study by Morrison et. al. [15] major differences in the perspective and structure is discussed. It is also noted that these perspectives and structures are used for speaker identification by law enforcement agencies around the world for forensic analysis. Differences between regions, countries and individual law enforcement agencies, and also within the reporting devices is investigated. Different approaches are used for speaker identification task. Lawyers, law enforcement agencies, and courtrooms use speech and other biometric features for recognizing suspects. Generally speaker recognition is helpful in identifying the person on the basis of their accents [15].

A method based on acoustic features such as pitch, density, and amplitude produces consistent results for forensic analysis and enable experts to make identification decision with high probability [42]. Moreover, formulation and generalization of actions undertaken by the experts are also considered. The experts aim to search for the assure value for their experimental outcomes. However, binary decisions are unavailable with their methods but this allows them to quantifying the confidence values for decision making.

Acoustic evidence need, is increasing day by day. At the same time the necessity for easy-to-apply and a reliable method is growing. A method that is consistent, straight and enables an expert to make decision with clearly determined probability is need of the time. An outline of the routine expert actions and methods that are successfully used by expert group is discussed in [42].

Similarly the Bayesian theory is used to interpret the results and to compute the confidence values [43]. Recorded samples are compared independently and evaluated for estimating the quality of accessible data. The continuity of the method depends on the results reliability. It is shown that the combination of methods into one expert report, making use of statistically independent, sets of parameters produces promising results. However some observations are noted where it is not possible to create a single quantitative measure for the final statement on speaker identity in the forensic cases. The study lie downs a frame work to make engineers and phoneticians to work together [20].

Bayesian framework is used for the evidence interpretation. Results obtained from non-experts perceptual tests are estimated with an automatic speaker recognition system. Noticeable higher accuracy rate is observed in ASR while conditions are matched of training and testing phase. In other hand performance of ASR systems degrade remarkably in mismatched conditions. Use of perceptual indications, that remains robust, for checking the possibility in accuracy increment of an automatic system in mismatched recording conditions is discussed [20].

Evidence of distinct forms are used and tested experimentally and forensically in solving the accent and forensic related problems [26]. This is an attempt to work in contrast to traditional approaches for extracting the information related to forensic. In traditional approach subject is examined by cross checking his recordings manually. While this approach helps to examine a person automatically through the stored recordings. A speech signal is entered and matched with the stored recordings in database. If features match with those of recordings then the person is considered guilty. Also there is an attempt to investigate automatic methods for filling the missing information for forensic analysis.

The missing information reported is forensic speech evidence strength is highlighted, with addition to multiple other issues involved in accurate estimation. By filling the missing information with higher level data it is possible to support either defence or prosecution [44]. Central idea of the study is evidence. Three

major discussions enrolled to verification are discussed: first one is numerous kinds of evidence that are being used in speaker recognition, second discusses accurate and logical structure for evolving the evidence, and third is up to what extent this evaluation can be tested to meet legal important standards [44].

In forensic analysis, the quality of audio recording plays an important role. The recorded speech samples are enhanced prior to experiments. The analysis of disputed utterances is carried out and the examination of the authenticity of audio recordings is done for forensic analysis [20]. The quality and quantity of the voice samples are crucial [30, 45]. An error free recording or recording that is taken in controlled environment gives a better result comparatively to a disputed recording. Voice filtering helps to improve the performance of a system.

2.2.4 Interruptions and Accent Recognition

Another common issue is the difference of transmission channels between samples of the same speaker [30]. The voice samples are mostly noisy for learning which usually mismatch with each other during testing [15]. Evidence is in the form of recorded speech interpretation in the forensic context and these evidences are presented for specific challenges [44]. In forensics, interruptions are examined separately to fetch information for evidence [16].

An approach is proposed to improve the accuracy and substantially lowers the time complexity for accent recognition. The approach uses a kernel function, computed faster than acoustics based methods [40]. In the case of matched recording it is noticed that the suspected and the questioned recordings gives better execution as compared to aural accent recognition systems. Thus automatic recognition presents more accurate results.

The accuracy of automatic systems is further enhanced with emotional clues, on which listeners depend like the utterance in a low or high frequency or slow or a fast utterance, that remain robust to mismatched conditions [37]. Similarly the forensic acoustics and their features are studied to fill the gap between the scientific and the legal world. Researchers are motivated to involve in improving the reliability and flexibility of acoustic forensic science [46]. This improvement helps to identify suspects more easily on the basis of better evidence fetched from recording.

In [47] feature selection methods are investigated for accent recognition.

The idea is to select more appropriate and robust features to improve the system performance and for identifying the most relevant speech signals for increasing the accent recognition accuracy for forensic application. Accent recognition helps to solve frameworks that involve classification tasks for accent attached with numerous other tasks.

The methods based on acoustic features exploit the variation within the delivery of sounds, while in phonotactic approach a sequence is attained everywhere these sounds occur and their role in the accent. These methods complement one another and confusion matrices are used for additional experiments. The work helps in constructing computationally efficient system for real-time applications [5]. Also accent recognition within a language helps forensic experts to execute results in profiling and comparison of speaker. Speaker recognition through accent supports personalising artificial utterance of text-to-speech systems.

2.2.5 Features and Classifiers for Accent Recognition

Recently the Mel-Frequency Cepstral Coefficients (MFCCs) [48] has attracted the attention of many researches for speaker and accent recognition tasks. Different classifiers like Bayesian, K-Nearest Neighbour (KNN), and Neural Network (NN) etc are trained on MFCC features. MFCC works efficiently as compared to traditional feature extraction methods that were used before. MFCC is considered most promising feature extraction method presently.

KNN is found as one of the simplest classifier. The mean vectors of MFCCs feature matrices are considered for simplicity. An alternative method such as Gaussian mixture models (GMM) to differentiate within microphones recording and enhances speaker recognition accuracy. It is taken into account to enhance MFCC prior to training. Method is tested on a small dataset [48, 49].

Using KNN algorithm with MFCC features is working because it is based on minimum Euclidean distance between the data to be tested and data present in the database. Within this work MFCC are also compared between the global group and the smaller sub groups. The classification of accents on the basis of MFCC is also investigated in [49]. It is necessary to differentiate human speech versus the spoofed speech [50]. Speech fusion and transformation of voice techniques are alarming situations for speaker identification systems [50].

Table 2.1: List of feature extraction methods and classifiers for recognition of accent

Method	Classifiers	Features	Applications
Lyn [34]	GMM	MFCC	Gender and Accent identification for Malaysian English
Huang et al [2]	GMM	MFCC	Accent adaptation and Speech recognition
Brown et. al [52]	i-vector	MFCC	Forensic accent recognition system
Abbas et. al [53]	Linear Discriminent Analysis (LDA)	MFCC	Pashto isolated digits database development
S. Afnan [27]	GMM-UBM	LPCC	Comparison of two different classifiers
S. Afnan [27]	SVM	MFCC	Comparison of two different classifiers
Huang et al [54]	GMM, Hidden Markov model (HMM)	MFCC	Decrease the computational cost for accent recognition
Sinha et al [4]	NN	MFCC	Accent identification with the contribution of different acoustic-phonetic features

Similarly MFCC and LPC techniques are compared for accent recognition task [51]. A database based on 10 sentences is used. No remarkable difference between the results obtained with both type of features are reported [51]. The objective of this investigation is to automatically recognize the speaker's accent among four regional accents for biometric identification. Focus is on features related to variations in pitch, intensity and rhythm. Gaussian Mixture Modelling (GMM) framework is used to achieve the goal. Table 2.1 shows a list of feature extraction methods and classifiers for accent recognition.

In [55] two different GMM-based algorithms are trained. Both perform well for accent recognition. Noise interruption is overcome with MFCC. This approach helps to test sample of a suspect to be recognized, within a noisy environment, using a few seconds of speech, and at different times of training and testing. Recording sessions mostly differs from one another.

Sessions do not contain enough data to recognise the accent, where as the voices are recorded with the help of mobile channel. Identifying a person on the basis of his speech accent for a forensic quality context is quite challenging. Recording is recorded in both noisy and controlled environment. This diverseness support recognition in forensic experimentation [15].

Speech style and gender is investigated in [17]. Speech style varies from person to person. And its variation in features like pitch and amplitude helps in finding speakers gender. The system performance is checked. Equal error rate (EER) performance is used as a metric to measure system performance. Furthermore, a robust

Table 2.2: Types of interruptions encountered in accent recognition

Interruption type	**Reason**
Environment [20]	Uncontrolled environment having interruptions like background noise or two people talking at same time
Medium [46]	Low quality microphone or telephone used for recording
Channel [58]	Band width, magnitude, misinterpretation, echo of recording
Speech styles [59]	Voice tone, rhythm, loudness, isolated words, speed (words per second)
Vocabulary [60]	Features available for training corpus

method for differentiating the gender and speech style is also presented.

In [56] it is determined whether a questioned voice belongs to the suspected person. Gaussian Mixture Models (GMMs) and the Bayesian Interpretation (BI) are combined for recognizing the speaker in forensic context. The challenges faced through Bayesian inference like interpretation of evidence recording in forensic scenarios, are investigated.

A new method is suggested called double-statistical approach (BI-GMM). That efficiently integrates the Gaussian mixture models (GMMs) with Bayesian interpretation (BI) framework. It provides a suitable result for analysing the recording as proof in the law enforcement agencies. The method produces promising results for the analysis of recording as scientific proof in the courtroom [16].

In [3] three basic speech modeling approaches are tried for accent recognition issue. For that three different classifiers are employed for each approach for finding perfect match within the speech modelling schemes and the classifiers. The experiments show promising results. Among all the classifiers GMM based i-vector and SVM gives the best result [3]. Table 2.2 discusses some interruptions that affect the recognition of accent and reasons of these interruptions. These interruptions have a clear impact on the results [57].

In [40] a kernel based SVM is trained for accent recognition. Series of experiments are conducted for different accents. Comparison of state-of-the-art methods are carried out from recording, first phone hypotheses is obtained using a audio identifier, then GMM-super-vectors are extracted for all the phone type and phonetic characteristics of speaker's are summarized effectively. Kernel function is modelled using a vector, which evaluates the phonetic similarities between the recordings pairs for training SVM classifiers for accent recognition.

A hybrid approach is discussed for identifying foreign accent in a spoken language [61]. The approach combines phonotactic and spectral features for handling the complication in accent recognitions. The main focus is to use accent recognition for foreign language as a language recognition system and use general speech features.

2.2.6 Accent Corpuses

In [62] a system is designed to overcome performance breakdown occurs due to speech variations. These variations are because of the different accents or dialects of speakers. Texas Instruments and Massachusetts Institute of Technology (TIMIT)[1] dataset [33] is used, which is designed corpus of phonetic American English speech by different genders and accents. This approach gives very appropriate result for accent recognition.

In [20] acoustic and automatic speaker recognition is compared for forensic analyses. This analysis is done with the help of a Bayesian structure for the evidence illustration. To estimate the strength of evidence a strategy correspondent to the Bayesian function is applied for forensic automatic speaker recognition. It uses emotional experiment that is done by non-professional and compares its working with an ASR network.

York ACComulation DISTribution (Y-ACCDIST) system is proposed in [26]. The system distinguishes between more similar accent varieties to one another than previous automatic accent recognition research. Y-ACCDIST inspects the interface of sociolinguistics and phonetics corpus as an introductory screening tool.

A subset of the Panjabi-English in Bradford and Leicester (PEBL) corpus is investigated that efficiently estimate accent similarities across different linguistics. This paper also applies automatic accent recognition technology to forensic casework [52]. A similar techniques is used on Chinese language for voice comparison in forensic cases [63]. The study demonstrates the comparison of audio in forensic, presented to courtroom in China where ratio of likelihood obtained by GMM is used. Data is tested under investigation. Hypothesized test is conducted on two sisters. A comparison of forensic audio review conducted for new method is discussed. Its result is granted to the courtroom for a civilian case [64].

Voice Comparison and Analysis of the Likelihood of Speech Evidence

[1] https://catalog.ldc.upenn.edu/LDC93S1

(VOCALISE) system has the capability of features comparison from a test acoustic recording of suspected speaker against features of an acoustic case of a targeted speaker [65]. VOCALISE makes a link between traditional phonetics-based and automatic speaker recognition systems for forensic cases. It enables the user to make objective estimates of the strength of the evidence in a speaker recognition case.

I-vector based recognition is used in dialect and accent classification [47]. It shows that recognizing of foreign accent over native accent is generally a difficult task. Foreign accent detection has many tasks to be done in future for new feature extraction and feature modeling schemes. In case of security applications these values of accent detection tasks are really high that are not acceptable [66]. Another modelling scheme Neural Networks (NN) gives very authentic results in case of recognizing pattern, by increasing its reliability. Further development is the sampling of native accents while uttering foreign languages, for creating equivalent accent recognition models [67].

Grapheme-based recognition voice models are trained by using hierarchical periodic neural network structure. Predictions of Grapheme model are obtained from a hierarchical model. The hierarchical model is practiced on language samples. When hierarchical model is compared with phoneme-based acoustic model trained on the same Grapheme prediction it gives an increment in result [68].

A new technique Null-Hypothesis is proposed for automatic forensic speaker recognition. Objective of this analysis is increment in the transparency level, consistency and connection with forensic while working on distinct automatic systems [21].

In some circumstances sounds around us become the subject of a law enforcement investigation, an accident review, or some other legal proceeding that ends up in a courtroom [46]. In many context acoustical scientists knowledge is helpful for legal and investigative proceedings. The sound at the recording background helps in different ways e.g. a clock voice or a voice of some machine, any emergency alarm interruption or plane landing sound help to detect the accuracy of recording.

Acoustical scientist's knowledge is helpful to legal and investigative proceedings in crime cases. Forensic acoustics bridge gap between scientific and the legal world [46]. Judiciary and other law agencies use language and biometric features

for identifying the culprits [15]. Accent recognition is also helpful in learning foreign languages and providing an answer to a student [67].

2.2.7 Urdu Accent recognition

A prediction framework is presented in [9] that provides weather information in Urdu language. It takes input verbally about the location of a place in Urdu. Corpus composed of 139 district names of Pakistan. Prime languages used in Pakistan are Punjabi, Pashto, Sindhi, Balochi, Seraiki and Urdu. Each language accent varies from one another. Therefore different techniques are implemented and evaluated to handle accent variations. Methodology discussed in this study is to construct an ASR system for weather forecast and obtains results from them. Results based testing and enhancement of field in system is compared.

In [9] demand of online verbal system is discussed in accent recognition scenario. As online information is need for development in this modern era and its significance increasing with the passage of time. Online systems are mostly beneficial to educate population. Un-educated or semi-educated people are not capable to completely avail this opportunity. Main barrier in availing online resources is low literacy rate and internet connectivity. Accent recognition and speech recognition is helping to overcome this barrier. Accent provides guidance in native languages. Online voice system is also helpful for visually handicapped in many terms. Accent recognition helps in speaking, communicating, dialogue, dictation, learning, understanding and translating a language.

In [33] a review of different voice corpora designed for different languages of the world is presented. Recognition system is designed to identify the voice of an individual talking through a microphone or with a telephone set and to transform the audio voice into text form. Dataset covers six major accents of Pakistan. Proposed methodology is a data pre-processing step for the development a successful integrated Urdu dialog system to provide weather information of Pakistan. Accent recognition is helpful in medical management [11, 33, 63], software development [69] and meteorology fields [9, 33].

2.3 Summary

This chapter summarizes the literature work on Accent recognition. Features and Classifiers used in literature for accent recognition are also briefly explained. This chapter else give an overview of accent recognition based application. Main focus is on forensic application with the help of accent recognition.

CHAPTER 3

FEATURES AND CLASSIFIERS

3.1 Overview

This chapter presents important features and classifier that have been used for accent recognition. Features are extracted from speech utterances. Features are extracted using different features extraction schemes (such as LPCC, MFCC). These futures are used in training of different classifiers like GMM, SVM and i-vector.

3.2 Linear Predictive Coding

Linear predictive coding (LPC) is an important speech feature [70]. It is considered a powerful and promising technique in speech analysis [71, 30]. Usually used for transferring spectral information and produce tolerant to transmission errors [72]. LPC is based on power spectrum of the signal [72].

In LPC, speech signal is analyzed with the help of formant estimation. The formants effects are eliminated from the speech signal, after that potency (strength) and frequency is evaluated for the extra humming sound at the background [70]. Speech signal samples are fetched as linear fusion of the LPC's previous samples. The calculation achieved is known as a linear predictor that is why it is called Linear Predictive Coding (LPC). The format is defined by the difference equation's coefficient [70].

To obtain LPCC first LPC is computed. LPCC is a widely well-known algorithm for extracting audio features. Sound frames potency and frequency spectrum is derived with the help of LPC parameters. Audio signals spectrum, modeling and pattern recognition is set by the outcome of logarithm increment to stops the rapid alternation of frequency spectrum that is highly to the point and improved in case of short-time character. Frequency alternation is due to Cepstrum extracted from actual spectrum. Commonly used short-term spectral include cepstral coefficients derived

from LPC known as Linear Predictive Cepstral Coefficients (LPCC) and their reversion coefficients. Benefit of LPCC over LPC is that it removes the channel interruptions from Cepstral mean subtraction (CMS) on Cepstral coefficients.

For calculating LPC, Cepstrum is calculated first. Cepstrum is known as a numbers sequence of a speech frame. It is calculated by two means, one by periodogram estimate and other by AR power spectral estimate. Cepstrum computed with the help of power spectrum estimated periodogram is used for pitch tracking, where as the Cepstrum calculated with the help of power spectral estimate for audio identification. Somehow the Cepstrum for speech recognition are now replaced by MFCCs.

3.2.1 Cepstrum computation

For LPC Cepstrum is computed. The Cepstrum is imagined similar to the auto-correlation sequence. Auto-correlation sequence is computed by power spectrum using the Wiener-Khinchin theorem. It is defined as follows, $x(n)$ is represented as a time domain discrete signal where n is index, $X(k)$ is the complex spectrum where $k = 1, 2, 3....N$ and N is number of samples, $P(k)$ is power spectrum of $x(n)$ and $A(n)$ is the autocorrelation sequence of $x(n)$. For complex spectrum value Discrete Fourier Transform (DFT) of $x(n)$ is computed using Equation 3.1 as:

$$DFT(x(n)) \rightarrow X(k) \tag{3.1}$$

$$X(k) = \sum_{n=1}^{N} x(n) e^{-i2\pi kn/N} \tag{3.2}$$

where $k = 0 \ldots\ldots N$ and N is total number of samples. $e^{-i\theta}$ can be expanded using Equation 3.3

$$e^{-i\theta} = \cos\theta - i\sin\theta \tag{3.3}$$

where $\theta = 2\pi kn/N$. Equation 3.2 can be written as Equation 3.4 and Equation 3.5:

$$X(k) = \sum_{n=1}^{N} x(n)(\cos(2\pi kn/N) - i\sin(2\pi kn/N)) \tag{3.4}$$

$$X(k) = \sum_{n=1}^{N} x(n)\cos\left(\frac{2\pi kn}{N}\right) - i\sum_{n=1}^{N} x(n)\sin\left(\frac{2\pi kn}{N}\right) \tag{3.5}$$

Inverse of DFT gives the value of $x(n)$ using Equation 3.6 as:

$$IDFT(X(k)) \rightarrow x(n) \tag{3.6}$$

The $x(n)$ obtained form inverse of DFT can be represented as shown in Equation 3.7

$$x(n) = \frac{1}{N} \sum_{n=1}^{N} X(k) e^{i2\pi kn/N} \tag{3.7}$$

Power spectrum ($P(k)$) is obtained by taking the square root of the absolute of the signal domain's ($x(n)$) DFT as shown in Equation 3.8.

$$|DFT(x(n))|^2 \rightarrow P(k) \tag{3.8}$$

$P(k)$ is the power spectrum of frame. Periodogram spectral estimate contains information that is not needed for speech recognition. So this kind of information is removed. For this purpose, clump of periodogram bins help to find out the estimate of energy present in different frequency zone. Complex Fourier transform absolute value is calculated, and result is squared. Commonly 512 point Fast Fourier transform (FFT) are executed and just first 257 coefficients are stored. Periodogram-based speech frame power spectral estimate is calculated in Equation 3.9 as:

$$p(k) = \frac{1}{N} |S(k)|^2 \tag{3.9}$$

where N is sample number and $S(k)$ is DFT value calculated below. Complex Discrete Fourier Transform (DFT) gives $S_i(k)$. i represents the number of frame equivalent. DFT of frame is calculated by multiplying hamming window with the framed signal. It is obtained using Equation 3.10 as follows:

$$S_i(k) = \sum_{n=1}^{N} s_i(n) w(n) e^{-j2\pi kn/N} \qquad 1 \leq k \leq K \tag{3.10}$$

Hamming window $w(n)$ of an sample long analysis window, $s(n)$ is framed signal, K is the DFT length, and $n = 1 \ldots N$ Autocorrelation sequence is achieved by taking the IDFT of the power spectrum ($P(k)$) using Equation 3.11 as:

$$IDFT(P(k)) \rightarrow A(n) \tag{3.11}$$

Correlation is the uniformity within signals. When x and y are similar, $x(i)$ is positive and vice versa. Signals with similar attributes attain high connection; whereas different signals attain a low connection. Signals having both similarities and differences at

equal level attain a correlation score somewhere in between. Signals are similar when there is a large number of negative correlations are present. In this case one signal is inverted with respect to the other. It is represented in Equation 3.12 as:

$$A(n) = \sum_{i=0}^{N} x(i)y(i) \tag{3.12}$$

By taking the power spectrum logarithm before the IDFT, Cepstrum is obtained:

$$IDFT(\log(P(k))) \rightarrow C(n) \tag{3.13}$$

Cepstrum is defined as an autocorrelation sequence of compressed logarithm, it convey facts that resembles to the autocorrelation sequence. Cepstrum is derived from the power spectrum log alternate to the standard power spectrum. It is expressed in Equation 3.14 as:

$$c_x(n) = \left(\frac{1}{2\pi}\right) \int_{-\pi}^{\pi} ln(|X(k)|)e^{jwn}d\omega \tag{3.14}$$

where $X(k)$ is the represented as a Fourier transform of the sequence $x(n)$. This definition represents the inverse transform after applying a natural logarithm to the Fourier transform of $x(n)$.

3.3 Linear Prediction Cepstral Coefficients

Linear Prediction Cepstral Coefficients (LPCCs) are computed in two ways, first is the same way as Cepstral and second method is computation from LPC. The only difference is LPCC is computed using interruption free power spectrum while Cepstral is computed from the periodogram estimation of the power spectrum [28]. For calculating LPC a Autocorrelation coefficients are computed. Figure 3.1 shows a block diagram for LPCC feature extraction.

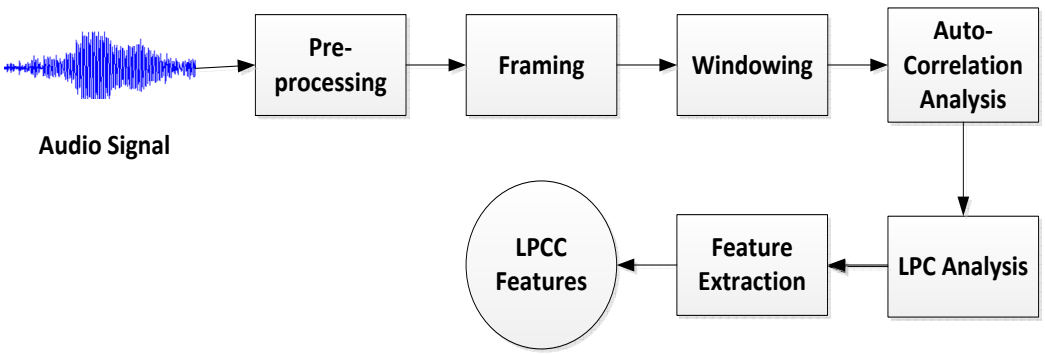


Figure 3.1: Block Diagram for extraction of LPCC features

3.3.1 Computing LPCCs from LPCs

LPCC is computed from LPCs with a simple repeated formula as shown in Equation 3.15 without doing any DFTs where a_n represents coefficients of linear prediction (p).

$$c(n) = \begin{cases} 0 & n < 0 \\ ln(G) & n = 0 \\ a_n + \sum_{k=1}^{n-1} \left(\frac{k}{n}\right) c(k) a_{n-k} & 0 < n \leq p \\ \sum_{k=n-p}^{n-1} \left(\frac{k}{n}\right) c(k) a_{n-k} & n > p \end{cases} \tag{3.15}$$

Cepstral coefficients infinite numbers are calculated with the help of LPC coefficients. Usually 12 to 20 Cepstral coefficients are used [73].

3.4 Discrete Fourier Transform

Discrete Fourier Transform (DFT) gives complete description of finite-duration signals in frequency domain. It determines the frequency content of the signal. From computation point of view, DFT uses lowest operation, which is based on involved multiplications numbers. It is also useful for two-dimensional signal or image processing. It can be efficiently applies on a portion of a long signal. Some remarkable properties of DFT are Linearity and Symmetry. DFT is computed with Equation 3.5.

3.5 Fast Fourier Transform

Fast Fourier Transform (FFT) is a fast computing method for DFT with reduced execution time. When N is large, then the number of computation savings is considerable. We can rewrite DFT (see Equation 3.2) as FFT (Equation 3.16 below):

$$X(k) = \sum_{n=1}^{N} x(n) W_N^{nk} \tag{3.16}$$

In computation same values of W_N^{nk} are calculated, for different fusion of k and n , nk is repeated as integer product whereas periodic function (W_N^{nk}) has N definite value.

3.6 Discrete Cosine Transform

A Discrete Cosine Transform (DCT) indicates data points of finite sequence as a sum of cosine functions vibrating at different frequencies. DCT is similar to DFT but use only the real numbers using Equation 3.17 as:

$$y(k) = \sqrt{\frac{2}{N}} \sum_{n=1}^{N} x(n) \frac{1}{\sqrt{1 + \partial_{K1}}} \cos\left(\frac{\pi}{2N}(2n-1)(k-1)\right) \tag{3.17}$$

where frame number is represented as N, DFT size is represented by K, While delta ∂_{K1} is given in Equation 3.18 as:

$$\partial_{ij} = \begin{cases} 0 & if \quad i \neq j, \\ 1 & if \quad i = j, \end{cases} \tag{3.18}$$

DCT is a short version for the FFT and considers the real part of FFT only.

3.7 Mel-frequency Cepstral Coefficients

A Mel-frequency Cepstral Coefficient (MFCC) is an algorithm for extracting distinct features from audio signal [19]. It is based on a short-term spectrum which is obtained through Fast Fourier Transform (FFT) [4].

MFCC extraction is carried out in few steps. In the first step signals are framed into short frames. Each frame is multiplied with a Hamming Window. Each frames power spectrum is obtained with the help of FFT. The power spectrum is then filtered with a filter bank. The filter bank operates in Mel frequencies. These frequencies are obtained by transforming the signal frequencies into Mel scale. The filter bank returns filtered power spectrum from which the filter bank energies are computed and DCT is applied. DCT gives multiple coefficients from which first 13 DCT coefficients are kept and the rest are discarded. Figure 3.2 shows a block diagram for extraction of MFCC features. It shows the steps that are used to extract the MFCC features form audio signal. All these steps are breifly explained below:

i. Let x be an audio signal sampled at 16 KHz. Sampled signal is framed in to frame of 25 milliseconds (ms). The frame length is 25ms x 16KHz = 400 samples. Each frame contains 400 samples.

ii. Let $s_i(n)$ be ith frame and n ranges from 1 to 400 and i is the number of frames. Hamming window and frames are multiplied. Formula for computing Hamming window is given in Equation 3.19 by:

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N}\right) \qquad 0 \leq n \leq N \tag{3.19}$$

where $N = 400$ is number of samples.

iii. Then FFT ($S_i(k)$) of every frame is computed with Equation 3.20 as follows:

$$S_i(k) = \sum_{n=1}^{N} s_i(n) w(n) e^{-j2\pi kn/N} \qquad 1 \leq k \leq K \tag{3.20}$$

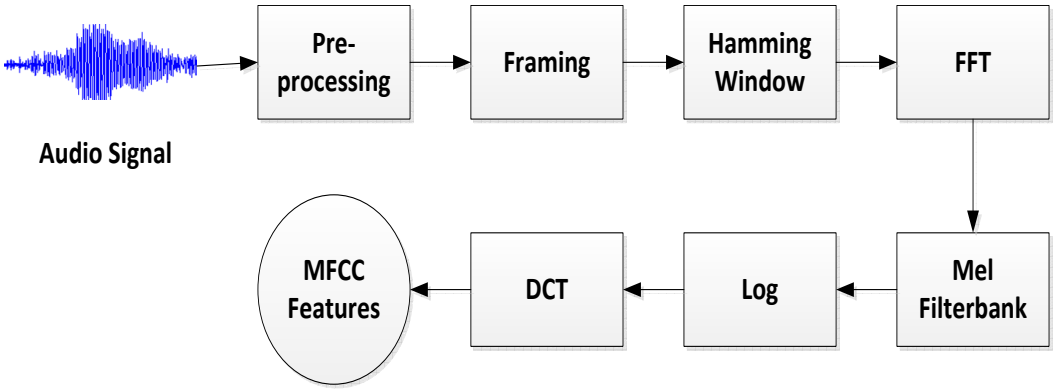


Figure 3.2: Block Diagram for extraction of MFCC

where $w(n)$ is hamming window, K is the FFT length, and N is sample number.

iv. Each frames power spectrum $P_i(k)$ is computed using Equation 3.21 as follows:

$$P_i(k) = \frac{1}{N}|S_i(k)|^2 \tag{3.21}$$

Generally, 512 point FFT is executed from which first 257 coefficient are kept.

v. Mel-frequency based filterbank is computed. For this purpose 0 Hz is selected as the lower frequency of signal and 8000 Hz as upper frequency. Twenty six equally space frequencies are obtained between $0 - 8000$ Hz as follows: F = 0, 296.3, 592.6, 888.9, 1185.2, 1481.5, 1777.8, 2074.1, 2370.4, 2666.7, 2963, 3259.3, 3555.6, 3851.9, 4148.1, 4444.4, 4740.7, 5037, 5333.3, 5629.6, 5925.9, 6222.2, 6518.5, 6814.8, 7111.1, 7407.4, 7703.7, 8000

vi. These frequencies are converted into Mel frequency (f) using Equation 3.22 as follows:

$$f = 1125ln(1 + F/700) \tag{3.22}$$

where ln is natural log. For instance, $F = 0$ Hz gives $f = 0$ Mels and $F = 8000$ Hz gives $f = 2834.99$ Mels. f = 0, 397.1, 690, 922.2, 1114.5, 1278.8, 1422.1, 1549.1, 1663.3, 1766.9, 1861.8, 1949.3, 2030.5, 2106.2, 2177.2, 2243.9, 2306.9, 2366.6, 2423.2, 2477.2, 2528.6, 2577.8, 2625, 2670.2, 2713.7, 2755.6, 2796, 2834.99 Equation 3.22 is invertable for which Equation 3.23 is used:

$$F = 700(exp(f/1125) - 1) \tag{3.23}$$

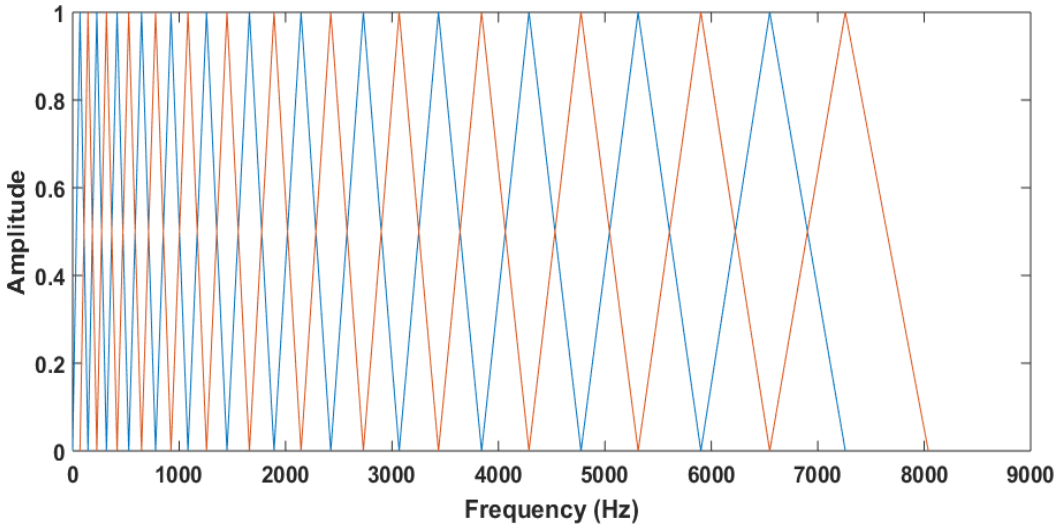


Figure 3.3: An illustration of Filterbanks used for MFCC features

vii. A set of 26 Mel-spaced filterbank ($H_m(k)$) is computed using Equation 3.24 where m represents frequency index:

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)} & f(m-1) \leq k \leq f(m) \\ \frac{f(m+1)-k}{f(m+1)-f(m)} & f(m) \leq k \leq f(m+1) \\ 0 & k > f(m+1) \end{cases} \tag{3.24}$$

where $m = 1, 2, 3, \ldots\ldots 26$ and $f(m-1) < k < f(m+1)$, In Figure 3.3, filter banks are graphically elaborated and illustrated with respected to frequency and amplitude on x-axis and y-axis, respectively.

The Filter bank consists of triangular filters. Filter bank is in the form of 26 vectors of length 257. The first filterbank starts at $f = 0$ Mel, reach its peak at $f = 397.1$ Mel, then return to zero at $f = 690$ Mel. The second filterbank start at $f = 397.1$ and so on.

viii. Filter bank and power spectrum is multiplied and added with coefficients for calculating the energies of filter bank, which results in 26 filter bank energies.

ix. Log of filterbank energies is computed and DCT is applied. This gives 26 cepstral coefficients. First 13 coefficients are kept and rests are discarded. These coefficients become MFCC coefficients.

3.8 Shifted Delta Coefficients

Shifted Delta Coefficient (SDC) feature are used for automatic language recognition. These are stacked version of delta coefficient over many frames [74].

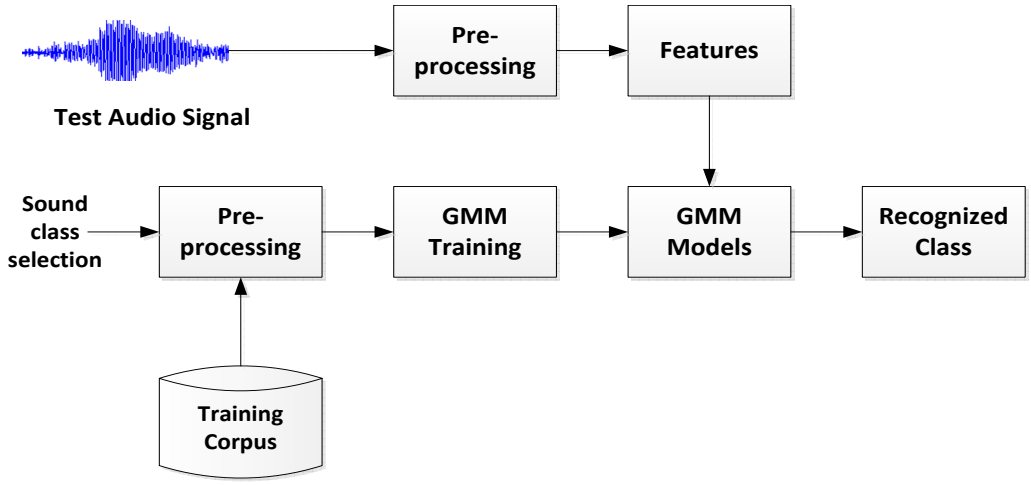


Figure 3.4: A block diagram for GMM classifier training

SDC is an improved feature set and work as an extension of delta-Cepstral coefficients. GMM-based language recognition was not up to mark then alternate approaches used before than SDC [75]. SDC coefficients capture difference over several frames of data. SDC coefficients are based upon four parameters, typically written as N-d-P-k [74]. MFCCs are calculated for each data frame based on N; i.e., $c_0, c_1, \ldots\ldots, c_{N-1}$. The parameter d determines the spread over which deltas are calculated, and the parameter P determines the gaps between successive delta computations. For a given time, t, we obtain through Equation 3.25 as:

$$\Delta c(t,i) = c(t+iP+d) - c(t+iP-d) \tag{3.25}$$

as an intermediate calculation. The SDC coefficients are then k stacked versions of Equation 3.25

$$SDC(t) = [\Delta c(t,0)^t, \quad \Delta c(t,1)^t \ldots\ldots \Delta c(t,k-1)^t]^t \tag{3.26}$$

3.9 Gaussian Mixture Models

Gaussian Mixture Models (GMM) are probabilistic model for representing normally distributed sub-populations within an overall population. GMMs are trained on features extracted from speech data and used in wide variety of speech related applications like accent recognition [70, 76], language identification [31] and speaker recognition [77]. A block diagram for GMM classifier training is presented in Figure 3.4

GMMs are parameterized by weights, means and co-variances of the GMM components. These parameters are used to compute probability (p) for a feature x, using Equation 3.27, Equation 3.28 and Equation 3.29

$$p(x) = \sum_{i=1}^{K} \omega_i N(x|\mu_i, \sigma_i) \tag{3.27}$$

$$N(x|\mu_i, \sigma_i) = \frac{1}{\sigma_i\sqrt{2\pi}} exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right) \tag{3.28}$$

$$\sum_{i=1}^{K} \omega_i = 1 \tag{3.29}$$

where N represents normal distribution. μ_i, $\sum_i$ and ω_i are mean, standard deviation and weight of ith component of the mixture. In case of multidimensional feature x, then $\sum$ is used as co-variance matrix and the probability is computed with Equation 3.30, Equation 3.31 and Equation 3.32 as:

$$p(x) = \sum_{i=1}^{K} \omega_i N(x|\mu_i, \sum_i) \tag{3.30}$$

$$N(x|\mu_i, \sum_i) = \frac{1}{\sqrt{(2\pi)^K |\sum_i|}} \quad exp\left(-\frac{1}{2}(x-\mu_i)^T \sum_i^{-1} (x-\mu_i)\right) \tag{3.31}$$

$$\sum_{i=1}^{K} \omega_i = 1 \tag{3.32}$$

3.9.1 Expectation Maximization Algorithm

Expectation Maximization (EM) algorithm is used for finding components for GMM. It is an iterative algorithm. This iteration consists of an E-step (expectation step) that come after an M-step (maximization step), that why the algorithm is called Estimation Maximization. EM starts from some initial estimate of mean co-variance and weights, and then proceeds iteratively to update until convergence is detected. In E-step the component expectation (C_j) assignments for each feature $x_i \in X$ given the model parameters, ω_j , μ_j and $\sum_j$, are computed. In M-step the expectations are maximized and the values of ω_j , μ_j and $\sum_j$, are updated. This iterative process is repeated until the algorithm merges and gives a maximum likelihood estimate. Following steps are used in EM algorithm:

i. Initialize: ω_j , μ_j and $\sum_j$ are initial estimates and $j = 1, 2, \ldots\ldots k$. Initial log-likelihood is computed with the help of Equation 3.33 as follows:

$$l = \sum_{i=1}^{n} \log(\sum_{j=1}^{k} \omega_j P_i(x)) \tag{3.33}$$

ii. E-step: compute with Equation 3.34 as:

$$\gamma_{ij} = \frac{\omega_j P_i(x)}{\sum_{j=1}^{k} \omega_j P_i(x)} \tag{3.34}$$

where $i = 1, \ldots n$, $j = 1, \ldots k$, and $\sum_{j=1}^{k} \gamma_{ij} = n_j$

iii. M-step: new estimates computation with Equation 3.35, Equation 3.36 and Equation 3.37 as:

$$\omega_j = \frac{n_j}{n} \tag{3.35}$$

$$\mu_j = \frac{1}{n_j} \sum_{i=1}^{n} \gamma_{ij} x_i \tag{3.36}$$

$$\sum_j = \frac{1}{n_j} \sum_{i=1}^{n} \gamma_{ij} (x_i - \mu_j)(x_i - \mu_j)^T \tag{3.37}$$

iv. Checking convergence: New log likelihood computation from Equation 3.38 as:

$$l^{new} = \sum_{i=1}^{n} \log(\sum_{j=1}^{k} \omega_j P_i(x)) \tag{3.38}$$

Step ii is repeated if $|l^{new} - l| > \delta$ for a threshold δ; otherwise algorithm ends.

3.10 GMM-Universal Background Model

GMM-Universal Background Model (GMM-UBM) method is widely used in speech related [28]. It is a predominant approach i.e, a basic hypothesis test within two hypotheses. A likelihood ratio is estimated and compared between the two hypotheses for a threshold decision [78]. Gaussian distribution is used for GMM due to better performances [79, 58]. GMM-UBM computation structure is shown in Figure 3.5.

Two hypothesis are shown in Figure 3.5. Let a test signal S that encounter with front end processing. Front end processing deals with feature extraction e.g., MFCC [29]. After feature extraction, the features are tested against (i) hypothesis speaker model (λ_{hyp}) and (ii) background model ($\lambda_{\overline{hyp}}$). The two results are combined

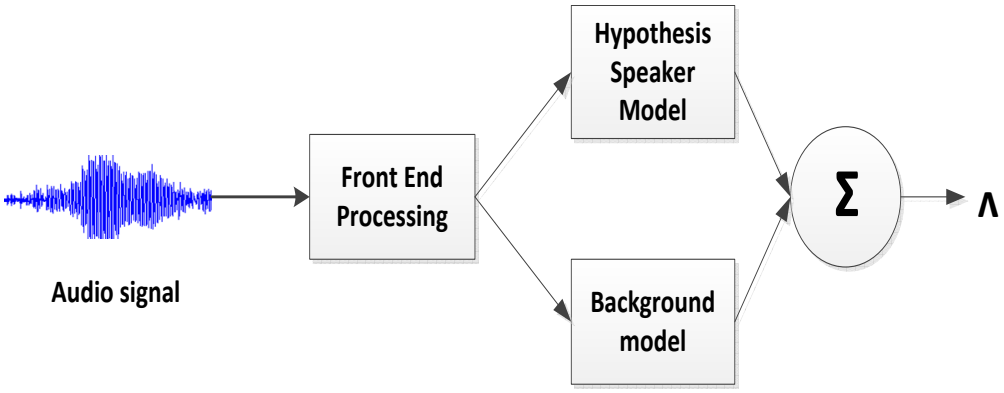


Figure 3.5: Illustration of GMM-UBM computation process

to produce an answer (Λ). Ratio of likelihood is computed with Equation 3.39 for producing the answer:

$$\frac{P(X|\lambda_{hyp})}{P(X|\lambda_{\overline{hyp}})} \begin{cases} \geq 0 & accept \quad \lambda_{hyp} \\ < 0 & reject \quad \lambda_{hyp} \end{cases} \tag{3.39}$$

After taking log Equation 3.39 becomes Equation 3.40:

$$\Lambda(X) = \log p(X|\lambda_{hyp}) - \log p(X|\lambda_{\overline{hyp}}) \tag{3.40}$$

Model is used for estimating only one model so it is well defined, whereas model λ_{hyp} is less defined as it is estimated for all the models and represents possible alternatives to $\lambda_{\overline{hyp}}$. Two common approaches are used for the hypothesis modelling. In the first approach a speaker model set is used for covering alternative hypothesis space. This other speakers set are known as sets of likelihood ratio and background speakers. Alternative hypothesis model for background speaker model set $(\lambda_1, \ldots, \lambda_N)$ with N numbers is given in Equation 3.41:

$$p(X|\lambda_{\overline{hyp}}) = F(p(X|\lambda_1), \ldots\ldots, p(X|\lambda_N)) \tag{3.41}$$

Background set likelihood values for average or maximum function is given by $F()$. For large number of hypothesized speakers, background speaker set is required for each set. The second approach is to pool a speaker's speech for training a single model. This pooling is also known as world, general or universal background (UBM) model. Its main advantage is that a single trained model is used for all hypothesis speakers. Mixture density for D-dimensional feature vector X is obtained with Equation 3.42:

$$p(X|\lambda) = \sum_{i=1}^{M} \omega_i p_i(x) \tag{3.42}$$

A fusion of linear M Gaussian mixtures is defined as density. $p_i(x)$ is parameterized by a mean $D \times 1$ vector (μ_i) and a $D \times D$ co-variance ($\sum_i$);

$$p_i(x) = \frac{1}{(2\pi)^{D/2} |\sum_i|^{1/2}} \quad exp\left(-\frac{1}{2}(x-\mu_i)^{'}(\sum_i)^{-1}(x-\mu_i)\right) \tag{3.43}$$

Mixture weight (ω_i) satisfy the constraints $\sum_{i=1}^{M} \omega_i = 1$. Density models parameters are $\lambda = \{\omega_i, \mu_i, \sum_i\}$, where $i = \{1, \ldots, M\}$. Feature vector of X are presumed independent, where $X = \{x_1, \ldots, x_i\}$ Likelihood $\log$ for model λ derived from Equation 3.42 is defined in Equation 3.44:

$$\log p(X|\lambda) = \sum_{i=1}^{M} \log p(x_i|\lambda) \tag{3.44}$$

GMM-UBM used single background model, given as $P(X|\lambda_{\overline{hyp}})$. Simplest approach is to simply combine (pool) gathered data for training the Universal background model with the help of EM algorithm as shown in Figure 3.6(a). Additionally, an approach for training distinctive UBMs known as dependent model with the sub-populations in the data is shown in Figure 3.6(b).

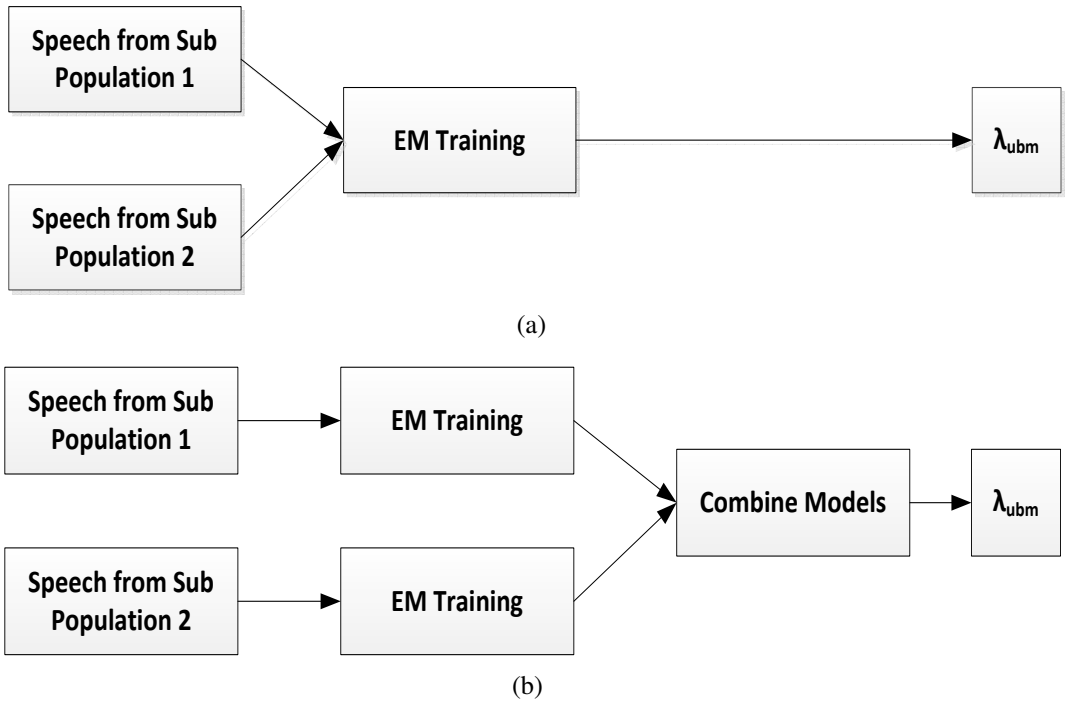


Figure 3.6: Illustration of GMM-UBM single model (a) background model (b) dependent model

After estimating UBM, the probability (Pr) for each speaker model is estimated in Equation 3.45 as:

$$Pr(t|x_i) = \frac{\omega_t p_t(x_i)}{\sum_{j=1}^{M} \omega_j p_j(x_i)} \tag{3.45}$$

where t represents mixture in the UBM. Use $Pr(t|x_i)$ and x_i to compute weight, mean and variance using Equation 3.46 and Equation 3.47 as:

$$n_t = \sum_{i=1}^{M} Pr(t|x_i) \tag{3.46}$$

$$E_t(x) = \frac{1}{n_t} \sum_{i=1}^{M} Pr(t|x_i)(x_i) \tag{3.47}$$

It is defined as Bayesian learning or Max a posterior (MAP) estimations. These are the counts for GMM mixture for computing the weight, mean, and variance as shown in Equation 3.48.

$$E_t(x^2) = \frac{1}{n_t} \sum_{i=1}^{M} Pr(t|x_i)(x_i^2) \tag{3.48}$$

New sufficient statistics are computed from speaker specific training data as shown in Equation 3.49, Equation 3.50 and Equation 3.51 respectively.

$$\hat{\omega}_t = [\alpha_t^{\omega} n_t / M + (1 - \alpha_t^{\omega})\omega_t]\gamma \tag{3.49}$$

$$\hat{\mu}_t = \alpha_t^m E_t(x) + (1 - \alpha_t^m)\mu_t \tag{3.50}$$

$$\hat{\sigma}_t = \alpha_t^v E_t(x^2) + (1 - \alpha_t^v)(\sigma_t^2 + \mu_t^2) - \hat{\mu}_t^2 \tag{3.51}$$

where $\{\alpha_t^{\omega}, \alpha_t^m, \alpha_t^v\}$ are adaptation coefficients, γ is scale vector, relevance vector is 16 by default and $\alpha_t^p, p \in \{\omega, m, v\}$ defined in Equation 3.52 as:

$$\alpha_t^p = \frac{n_t}{n_t + r^p} \tag{3.52}$$

3.11 I-vector (Identity Vector)

I-vector is a sequence of vectors (generally Cepstral coefficients) derived from a speech recording. In I-vector three basic steps for speaker recognition are used. First step is extraction of I-vector, second is modelling of extracted vectors and last one is computing likelihood ratio. Once an I-vector is derived, the mechanism for extraction is ignored [80]. Figure 3.7 shows an illustration for computation of I-vector. The steps

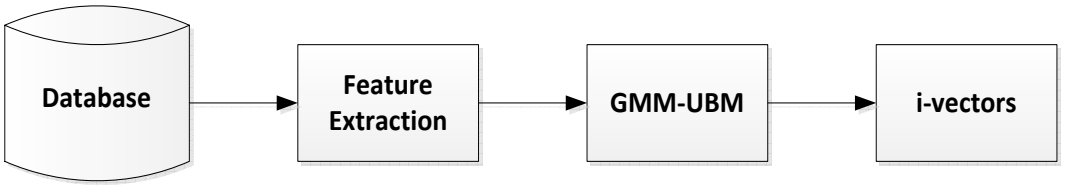


Figure 3.7: Illustration of i-vector method

involved in i-vector based recognition systems are:

i. Classifier training (with the help of UBM using EM)

ii. Computing zero and first order statistics

iii. Training Total Variability Space

iv. Utilizing UBM and TVS for i-vector extraction

v. Speaker (i-vectors) are separated for training and remaining for testing

vi. Use of highest mean, and sum to integrate i-vector

vii. True three tests are calculated from four tests for identifying speaker separately

3.12 Support Vector Machine

Support Vector Machine(SVM) is known as an algorithm for supervised machine learning. It is used for divergent challenges including regression and classification. Mostly it is used for classification problem. An SVM tries to find out a margin between data classes to generalizes the test data points [3]. In accent recognition the SVM with the discriminate classifiers are of great use [19].

In SVM the data items are projected onto n-dimensional space. A hyper-plane is picked that discriminate two classes positive and negative for classification step. SVM uses different kernels, such as kernels that use polynomial and radial basis to improve the classification accuracy. SVM includes some important features:

i. In start kernel uses clear evolution for feature space extracted with the help of SVM for low complexity calculation.

ii. Then a simple mean-squared error classifier is build with the help of SVM for generating more precise system.

iii. Lastly, a compatible and approved system as compared to other approaches (like Gaussian) is proposed.

An SVM classifier creates a partition within speaker and pretender [27]. Two group classification problems are solved through a Support Vector Machine (SVM) in machine learning. Input vectors are non-linearly mapped with a high dimension feature space. Feature space then makes a decision surface that is linear. Learning machine potential is measured by the decision surface properties [81].

Main idea is to implement the training data that are separated without errors. SVM model is presented with the help of a clear gap that separate categories. This gap is as wide as possible. Machine learning basic task is data classification. Given a training dataset of n points of the form:

$$(\overrightarrow{x}_1, y_1), \ldots\ldots\ldots, (\overrightarrow{x}_n, y_n) \tag{3.53}$$

where y_i is given as 1 or -1, it indicates the class of $\overrightarrow{x}_i$. Each $\overrightarrow{x}_i$ is a p-dimensional real vector. Hyper-plane is found with a maximum margin. Margin is the division of two groups of points $\overrightarrow{x}_i$ for which $y_i = 1$ or $y_i = -1$. Any hyper-plane is written as the set of points $\overrightarrow{x}$ satisfying as in Equation 3.54:

$$\overrightarrow{w}.\overrightarrow{x} - b = 0, \tag{3.54}$$

where $\overrightarrow{w}$ is the normal vector of hyper-plane. $\overrightarrow{w}$ is not necessarily a unit vector. The offset of hyper-plane parameter $\frac{b}{(||\overrightarrow{w}||)}$ from the origin along the normal vector $\overrightarrow{w}$. Two parallel hyper-planes are selected that clearly separates the two data classes, so that the gap between them is as large as possible.

The bounded region within two hyper-planes is ”margin”, and the hyper-plane with maximum margin is the one that is between them. With an ordered dataset, hyper-planes is explained with the help of Equation 3.55 and Equation 3.56 as follows:

$$\overrightarrow{w}.\overrightarrow{x} - b = 1 \tag{3.55}$$

$$\overrightarrow{w}.\overrightarrow{x} - b = -1 \tag{3.56}$$

where anything on or above the boundary is of class 1 else class -1.

Kernel methods are considered as an algorithms class for pattern recognition in machine learning, and SVM is considered as its best member. Geometrically the distance between these two hyper-planes is measured by $\frac{2}{(||\overrightarrow{w}||)}$, therefore for maximizing the planes distance minimize$||\overrightarrow{w}||$. It is also prevented that data points fall upon the margin. This can be rewritten in Equation 3.57 as:

$$y_i(\overrightarrow{w}.\overrightarrow{x} - b) \geq 1, \quad \forall \quad 1 \leq i \leq n \tag{3.57}$$

3.12.1 Radial Basis Kernel function

Radial Basis Kernel function (RBF) are mostly used as a kernel in SVM. The RBF kernel for two samples x and $x^{'}$, represented as feature vectors in some input space, is defined in Equation 3.58 as:

$$K(x, x^{'}) = exp\left(-\frac{||x - x^{'}||^2}{2\sigma^2}\right) \tag{3.58}$$

$||x - x^{'}||^2$ is the Euclidean distance within the feature vectors and (σ) is a free parameter. An equivalent definition involves a $\gamma = \frac{1}{2\sigma^2}$. γ value is inserted in Equation 3.59:

$$K(x, x^{'}) = exp\left(-\gamma||x - x^{'}||^2\right) \tag{3.59}$$

3.12.2 Polynomial Function

The polynomial kernel is a kernel that uses a polynomial function. The polynomial kernel is defined in Equation 3.60 as:

$$K(x, y) = (x^T y + c)^d \tag{3.60}$$

where x and y are the input feature, $c \geq 0$ is a free parameter and d is a polynomial degree. Kernel is homogeneous when $c = 0$.

3.13 Summary

This chapter concludes the features and classifiers used in this thesis. Features including MFCC and LPCC are widely used. The classifiers like GMM-UBM, SVM and i-vector are discussed and shown their mathematical method for calculating the m. And feature extraction and training of classifier are also explained.

CHAPTER 4

METHODOLOGY

4.1 Overview

This chapter presents the proposed method for Urdu accent recognition. The proposed method is based on MFCC and GMM-UBM and a Feature Mapping (FM) process. The proposed method is named as GMM-FM

4.2 Proposed method

The block diagram for proposed method is shown in Figure 4.1. First of all an Urdu speech corpus is constructed. The corpus is randomly divided into training and test sets. Speech features are computed on training and test set samples. A feature map is applied, which maps features to higher dimension space. Then classifier is trained on mapped features. The GMM-UBM is trained on mapped features and the test samples are classified. Finally performance is measured using the accuracy and equal error rate metrics.

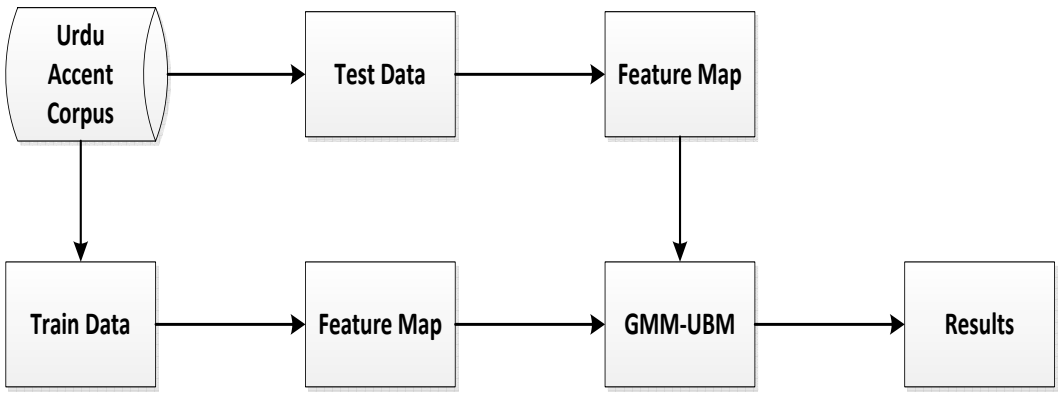


Figure 4.1: Block Diagram for proposed GMM-FM method

The Urdu speech corpus contains the samples of both male and female speakers of four different regional languages of Pakistan. These languages are Punjabi, Sindhi, Balochi and Pashto. The speakers are selected randomly. Accent helps identification of a person. For this purpose another dataset is constructed where speaker identification is carried out on basis of accent. Accent recognition on Kaggle dataset [1] is also impalement. The Kaggle dataset is a dataset of English language.

4.3 Feature map

Feature map is a process through which the speech features are projected onto higher dimensional space to increase the accent extraction and speaker recognition performance of GMM-UBM. Similar sort of mapping is also used in Support Vector Machines [81]. In contrast, this thesis uses feature map to increase the accuracy of GMM-UBM classifier.

To understand the process of feature mapping, let $x = [x_1, x_2, \ldots, x_n]$ be an n-dimensional feature vector and $x \in R$. A transformation φ is applied on x to obtain a feature vector $x^{'}$ in Equation 4.1 as:

$$x^{'} = \varphi(x) = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & \ldots & x_n \\ x_1^2 & x_2^2 & x_3^2 & x_4^2 & \ldots & x_n^2 \end{bmatrix} \tag{4.1}$$

The mapping process transforms each $1 \times n$ size feature vector x into $x^{'}$ i.e., a new feature vector is obtained of size $2 \times n$, where x_i^2 is obtained by squaring the i^{th} element of x. The experimental results given in the next chapter show that such a mapping improves the accuracy of GMM-UBM classifier and also outperforms linear SVM and SVM based on polynomial and RBF kernels.

Let X be a sequence of training feature vectors i.e, $X = [x_1, x_2, x_3, \ldots, x_m]$. Each feature vector x_m is of size $1 \times n$. The size of X is $m \times n$; where m is number of training feature vectors. Through proposed mapping process $X^{'}$ is obtained i.e, $X^{'} = \varphi(X)$, where the size of $X^{'}$ is $2m \times n$ as each $1 \times n$ feature vector is mapped to $2 \times n$ feature vectors.

4.4 Urdu Speech Corpus

Urdu speech corpus is consist of recording are collected from different internet sources in four major Urdu accents; Punjabi, Pashto, Sindhi and Balochi. Each accent category of Urdu corpus has 50 speakers. These speakers are selected randomly. The

[1] https//www.kaggle.com/datasets

reason behind construction of this type of corpus is to implement text and speaker independent accent recognition. Total number of recording are 70 per category, which are randomly divided in to two disjoint sets, the training set, which consists of 50 recordings and the test set consists of remaining 20 recordings.

The corpus consist of total 280 recordings (70 speakers $\times 4$ accents = 280). Each recording is approximately 15 seconds long in .wav format with 16 KHz as sampling rate. Recordings of each accent category is randomly divided into training and test set. Total number of recording in the training set are $50 \times 4 = 200$. Test data consist of remaining 20 Recordings. Total number of recordings in the test set is $20 \times 4 = 80$. Table 4.1 summarizes the Urdu speech corpus.

Table 4.1: Urdu speech corpus

Accents	Number of Speakers per Category	Number of Samples per Speaker	Training Samples per Speaker	Test Samples per Speaker	Nature of Speech Samples
Balochi	70	70	50	20	Speaker and text independent
Pashto	70	70	50	20	
Punjabi	70	70	50	20	
Sindhi	70	70	50	20	

4.5 Forensic Urdu speech Corpus

The Forensic Urdu speech Corpus is used for accent based forensic speaker recognition. The recordings for this corpus are gathered from different internet sources. The corpus consists of 4 speakers per accent category. There are 60 recordings per speaker and which are randomly divided into two sets training and test. The training set consists of 40 recording and the test set consists of remaining 20 recordings. Table 4.2 summarized the Urdu forensic speech recognition datasets.

Table 4.2: Forensic speaker recognition dataset

Accents	Number of Speakers per Category	Number of Samples per Speaker	Training Samples per Speaker	Test Samples per Speaker	Nature of Speech Samples
Balochi	4	60	40	20	Speaker and text independent
Pashto	4	60	40	20	
Punjabi	4	60	40	20	
Sindhi	4	60	40	20	

4.6 Kaggle Accent Corpus

This corpus is for English language and consists of five different accent categories which are Arabic, English, French, Spanish and Mandarin. For this corpus

a specific English paragraph is recorded by different accent category speaker. It is text-dependent and speaker-independent corpus. The paragraph is given below:

> *"Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station."*

This corpus is used for accent recognition and to evaluate the performance of the proposed method. Table 4.3 summarize the Kaggle corpus.

Table 4.3: Kaggle Speech Corpus

Accent	**Number of samples per speaker**	**Training samples per speaker**	**Test samples per speaker**	**Nature of speech samples**
Arabic	63	43	20	Speaker and text independent
English	63	43	20	
French	63	43	20	
Spanish	63	43	20	
Mandarin	63	43	20	

4.7 Classifier

GMM-UBM is used as a classifier in the proposed method. GMM-UBM compute two models from the MFCC features of the training samples. These models are accent independent and accent dependent models.

4.7.1 Accent Independent Model

The training data after passing through the feature mapping process is provided to GMM-UBM classifier for training purpose. GMM-UBM is trained with different mixture components starting from $M = 2$ up to $M = 256$ components. GMM-UBM provides an accent independent model known as a background model (λ_{ubm}). To obtain this model the features of training sample of all the accent categories are combined and M-different Gaussian mixture components are computed on combined samples using the GMM algorithm [79] as illustrated in Figure 4.2.

λ_{ubm} is parameterized by M-mixture components having mixture weight ω_i , mean vector μ_i of size $n \times 1$, and $n \times n$ sized covariance matrices $\sum_i$ and $\sum_{i=1}^{M} \omega_i = 1$. The mixture density in case of a feature vector $x \in R^n$ is computed from i^{th} mixture

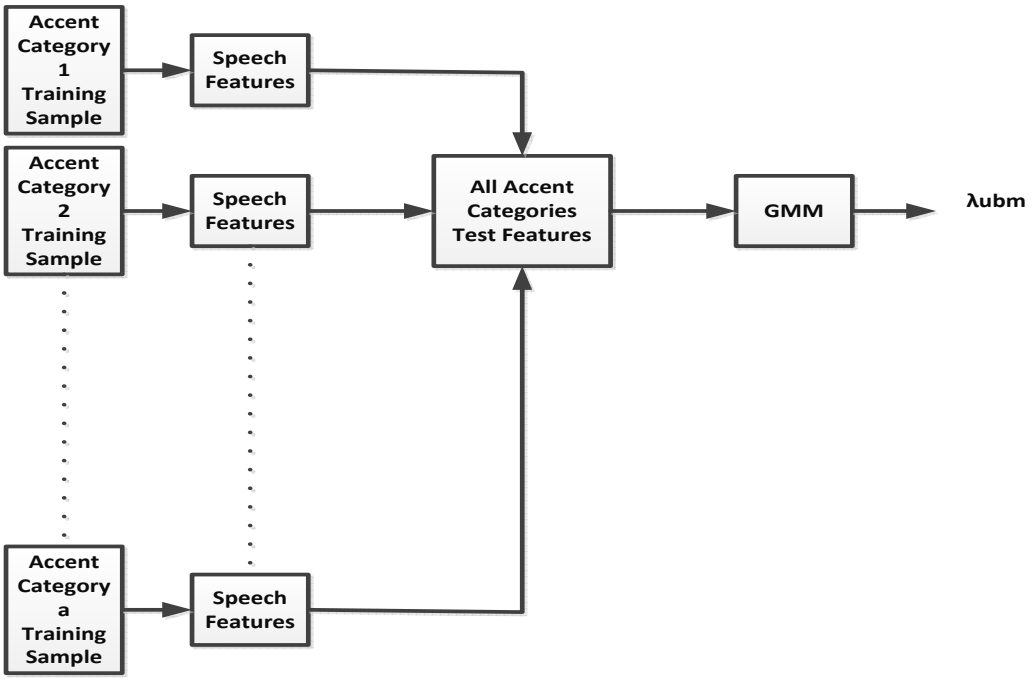


Figure 4.2: Illustration of GMM-UBM method for calculation of accent independent model

of λ_{ubm} is as shown in Equation 4.2:

$$p(x|\lambda_{ubm}) = \sum_{i=1}^{M} \omega_i p_i(x) \tag{4.2}$$

The density is a linear combination of M Gaussian densities $p_i(x)$. In case of a sequence of feature vectors $X = [x_1, x_2, x_3, ...x_T]$, the log likelihood is computed as in Equation 4.3:

$$\log p(X|\lambda ubm) = \sum_{t=1}^{T} \log p(x_t|\lambda_{ubm}) \tag{4.3}$$

GMM-UBM independent model is illustrated in Figure 4.2.

4.7.2 Accent Dependent Model

The accent dependent models (λ_a) are computed from λ_{ubm} with Bayesian adaptation [82]. The adaption process adapts the parameters of λ_{ubm} i.e., mean, covariance and mixture weights, for each accent category of the Urdu speech corpus one by one as shown in Figure 4.2.

There are four accent categories, so $a = \{1, 2, 3, 4\}$. Let X_a be a set of training feature vectors that belong to accent category a, where $X_a = [x_{a_1}, x_{a_2}, x_{a_3}, \ldots, x_{a_T}]$ and let i be i^{th} Gaussian mixture of λ_{ubm} the probability is computed with Equation 4.4:

$$Pr(i|x_{a_t}) = \frac{\omega_i p_i(x_{a_t})}{\sum_{j=1}^{M} \omega_j p_j(x_{a_t})} \tag{4.4}$$

Then sufficient statistics for parameter adaptation is obtained with Equation 4.5:

$$s_i = \sum_{t=1}^{T} Pr(i|x_{a_t}) \tag{4.5}$$

$$E_i(x) = \frac{1}{s_i} \sum_{t=1}^{T} Pr(i|x_{a_t}) x_{a_t} \tag{4.6}$$

$$E_i(x^2) = \frac{1}{s_i} \sum_{t=1}^{T} Pr(i|x_{a_t}) x_{a_t}^2 \tag{4.7}$$

These sufficient statistics create the adapted parameters for i^{th} mixture of accent model λ_a from the i^{th} mixture of λ_{ubm} as Equation 4.8, Equation 4.9 and Equation 4.10 below:

$$\hat{\omega}_i = [\alpha_i^{\omega} n_i / T + (1 - \alpha_i^{\omega}) \omega_i] \gamma \tag{4.8}$$

$$\hat{\mu}_i = \alpha_i^{m} E_i(x) + (1 - \alpha_i^{m}) \mu_i \tag{4.9}$$

$$\hat{\sigma}_i^2 = \alpha_i^{v} E_i(x^2) + (1 - \alpha_i^{v})(\sigma_i^2 + \mu_i^2) - \hat{\mu}_i^2 \tag{4.10}$$

where $\{\alpha_i^{\omega}, \alpha_i^{m}, \alpha_i^{v}\}$ are adaptation coefficients, γ is a scale vector computed over all adapted mixture weights to ensure they sum to unity and α_i^{p}, $p \in \{\omega, m, v\}$ is defined as in Equation 4.11:

$$\alpha_i^{p} = \frac{s_i}{s_i + r^p} \tag{4.11}$$

where r^p is a fixed relevance factor for parameter p and r=16 is used as default. Illustration of GMM-UBM dependent model is shown in Figure 4.3.

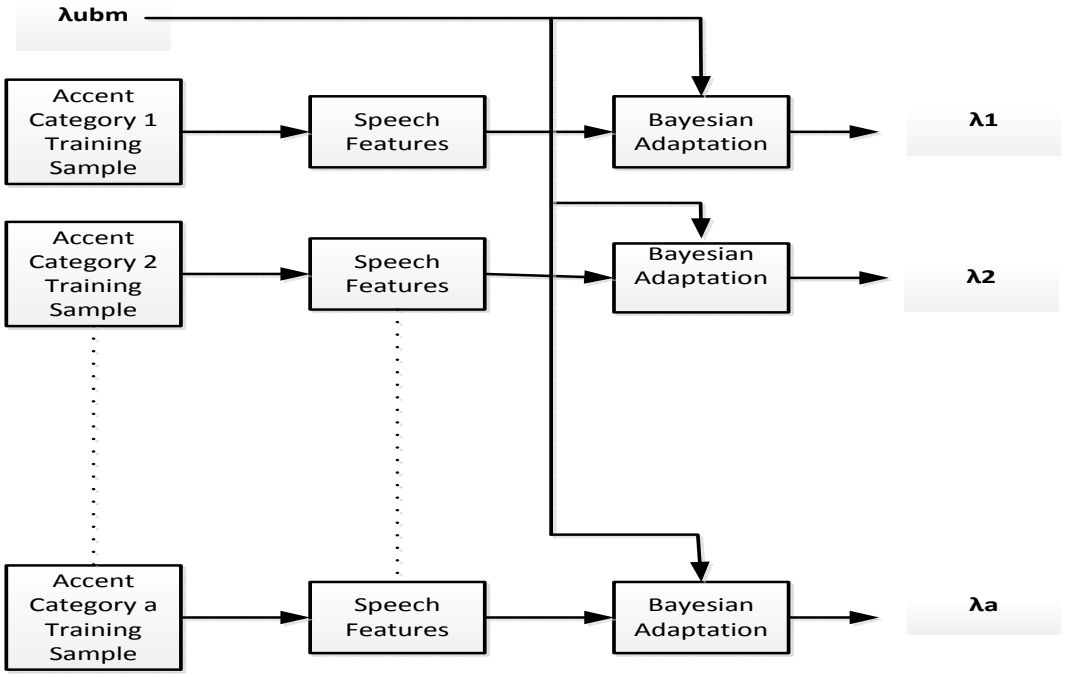


Figure 4.3: Illustration of GMM-UBM method for calculation of accent dependent model

4.8 Classification

Each accent dependent model λ_a is parameterized by $\hat{\omega}_i, \hat{\mu}_i, \hat{\sigma}_i^2$. Now test samples are applied for accent recognition. Let a test sample comprises of feature vectors $Y = [y_1, y_2, y_3, ..., y_T]$, the log-likelihood is computed in Equation 4.12 as:

$$\Lambda_a(Y) = \log p(Y|\lambda_a) - \log p(Y|\lambda_{ubm}) \tag{4.12}$$

The accent is predicted for the test sample and it is from the accent category that maximizes $\Lambda_a(Y)$. After that accuracy and EER are computed.

4.9 Confusion Matrix

It is a matrix to evaluate the performance description of classifier on a test dataset. It is also known as a summary of prediction. Numbers of correct and incorrect predictions are counted class wise. Each row of the confusion matrix corresponds to a predicted class. And each column of the matrix corresponds to an actual class.

The counts of correct and incorrect classification are then filled into the table. The total number of correct prediction for a class goes into the expected row for that class value and the predicted column for that class value. In the same way, the total number of incorrect predictions for a class goes into the expected row for that class

value and the predicted column for that class value.

Basic terms used in the confusion matrix are as follows:

i. True positives (TP): the predicted accent is same as actual accent.

ii. True negatives (TN): The predicted accent is false and the actual accent is also false.

iii. False positives (FP): The predicted accent is true, but actual accent is false.

iv. False negatives (FN): The predicted accent is false, but actual accent is true.

There are two possible predicted classes: ”true” and ”false”. While predicting the Punjabi accent, if the accent is ”true” it means accent is Punjabi, and ”false” means accent is not Punjabi. The classifier makes a total of 80 predictions. Table 4.4 illustrates a confusion matrix, where column represents number of predictions made and row shows ground truth. Values in diagonal shows true positive (TP), therefore, $TP = 15 + 16 + 19 + 17 = 67$. Whereas other values are either TN, FP or FN. Accuracy is equal to $(TP)/total = (67)/80 = 83\%$. Accuracy is expressed in percentage.

Table 4.4: A confusion matrix for four provisional languages

	Balochi	**Pashto**	**Punjabi**	**Sindhi**
Balochi	15	4	1	0
Pashto	3	16	1	0
Punjabi	0	0	19	1
Sindhi	0	1	2	17

4.10 Equal Error Rate (EER)

Equal error rate (EER) is a value where false acceptance rate becomes equal to false rejection rate. When the rates are equal, the common value is referred to as the equal error rate.

4.11 Forensic Speaker Recognition

This section presents experimental results for forensic speaker recognition. Accent classification (AC) prior to speaker recognition is investigated and the experimental results for forensic speaker recognition with and without AC are presented. The experiments are performed on an Urdu Forensic Speech Corpus. Figure 4.4 shows the block diagram of forensic speaker recognition without

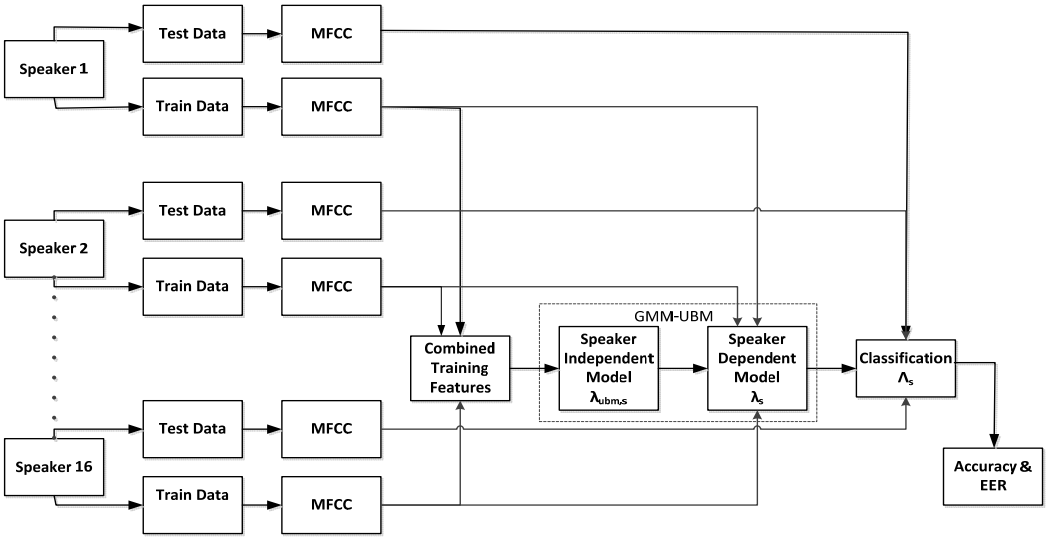


Figure 4.4: Block diagram for forensic speaker recognition without accent classification

Accent Classification (AC) using the GMM-UBM. The same block diagram is also implemented for the proposed method only difference is the proposed feature mapping.

First the background model (i.e. speaker independent model) is computed ($\lambda_{ubm,s}$) for both GMM-UBM and the proposed GMM-FM method by combining the MFCC features of training samples of the speaker categories. Then speaker dependent model (λ_s), for each speaker category is adapted from $\lambda_{ubm,s}$ using the Bayesian adaptation [79]. Since there are four speakers per accent, so total number of λ_s adapted are 4×4 (accents) = 16. Having computed $\lambda_{ubm,s}$ and λ_s , the next step is to recognize the speaker in a test sample. Let Y be a set of MFCC feature vectors obtained from a test speech sample. The log-likelihood for Y is computed in Equation 4.13 as:

$$\Lambda_s(Y) = \log p(Y|\lambda_s) - \log p(Y|\lambda_{ubm,s}) \tag{4.13}$$

The predicted speaker for the test sample belongs to s^{th} speaker category of the corpus if it maximizes $\Lambda_s(Y)$. After that accuracy and EER are computed. The same process is used for the proposed method. But the only difference is the feature mapping step which is used only in the proposed method to map the training and test samples.

For GMM-UBM based speaker recognition with AC the background model

(λ_{ubm}) and the accent independent models λ_a used for accent recognition in Section 3.6 are used. For a test sample, first accent is recognized from the MFCC feature vectors Y of the test sample in Equation 4.14 as:

$$\Lambda_a(Y) = \log p(Y|\lambda_a) - \log p(Y|\lambda_{ubm}) \tag{4.14}$$

The accent category that maximizes $\Lambda_a(Y)$ is used as a predicted accent for the test samples. Once the accent is identified then Y is processed for speaker recognition in Equation 4.15 as follows:

$$\Lambda_{s,a}(Y) = \log p(Y|\lambda_{s,a}) - \log p(Y|\lambda_{ubm,a}) \tag{4.15}$$

where $\lambda_{s,a}$ and $\lambda_{ubm,a}$ are the speaker dependent and independent models of the predicted accent category, respectively. The speaker is identified as the speaker of the predicted a^{th} accent category that maximizes $\Lambda_{s,a}(Y)$.

To compute $\lambda_{ubm,a}$, the MFCC features of training samples of all the speakers of the a^{th} accent category are combined. The combined features belong to different speakers but all of them have accent. For instance, the Balochi accent category (see Table 4.4) contains four speakers. All the speakers have the same Balochi accent. So $\lambda_{ubm,a}$ is speaker independent model with in accent category a.

The speaker dependent models $\lambda_{s,a}$ are then adopted from $\lambda_{ubm,a}$ using the Bayesian adaptation process. To recognize a speaker from a test sample first accent is identified using Equation 4.14 and then speaker are recognized using Equation 4.15.

4.12 Summary

This chapter summarizes the methodology used for Urdu accent recognition. It concludes the data gathering and briefly explains the corpora used in this thesis. It shows how the data is collected and how different techniques are applied. Test and training data of corpora are also defined. It also describes the performance measures.

CHAPTER 5

EXPERIMENTAL SETUP AND RESULTS

5.1 Overview

This chapter presents accent recognition results on three corpuses with different feature extraction schemes and classifiers. It also presents performance evaluation of the proposed GMM-FM method with respect to state of the art. The result consist of two parts in the first part Urdu accent recognition is presented and in the second part forensic speaker recognition result are presented with and without using the accent classification as a pre-processing set to speaker recognition.

5.2 Accent Recognition

Accent recognition is method to identify the specific features of speaker. These features are used in different ways to gather different information. In this thesis three corpuses are tested with different feature extraction schemes and classifiers. Accent recognition helps to evaluate these features and classifiers performances on the basis of their results accuracy and error rate. Accent recognition presents the comparison between different features and classifiers to judge the accuracy and equal error rate of proposed system and post schemes.

5.2.1 Comparison between speech features

In this section, a comparison of different speech features and classifiers is presented for accent recognition Two different corpuses are used (i) Urdu speech corpus (as explained in Section 4.6) (ii) Kaggle accent corpus. Table 5.1 shows a comparison between MFCC, LPC, LPCC and SDC features for accent recognition. EER(%) is used as a metric for performance comparison. GMM-UBM is used as a classifier. The objective of this comparison is to identify the best features for accent recognition. The GMM-UBM is trained using different mixture components i.e., $2, 4, 8, 16, 32, 64, 128$ and 256. The comparison shows that MFCC compared to others, demonstrates better EER on both Urdu and Kaggle corpuses. MFCC achieves

Table 5.1: EER(%) based comparison between MFCC, LPCC, SDC, LPC using GMM-UBM with different components for accent recognition

GMM Components		**2**	**4**	**8**	**16**	**32**	**64**	**128**	**256**
Urdu	**MFCC**	31.8	30.5	29.3	26.8	21.8	18.0	13.0	9.7
	SDC	33.6	32.8	31.2	29.6	25.7	21.4	17.9	12.3
	LPCC	38.2	36.7	34.5	31.6	29.8	26.9	22.1	19.4
	LPC	45.4	45.4	46.3	44.6	40.0	35.0	32.1	26.3
Kaggle	**MFCC**	34.0	34.0	34.0	31.0	30.0	29.0	29.3	29.0
	SDC	38.2	37.5	36.1	35.3	32.9	31.5	31.2	31.2
	LPCC	43.2	42.2	41.9	39.1	38.6	36.2	34.9	33.9
	LPC	49.0	49.5	48.3	47.0	47.3	45.0	44.0	42.2

Table 5.2: Accuracy (%) based comparison between MFCC, LPC, SDC and LPCC using GMM-UBM with different components for accent recognition

GMM Components		**2**	**4**	**8**	**16**	**32**	**64**	**128**	**256**
Urdu	**MFCC**	54.5	58.3	63.3	64.5	69.5	74.5	85.8	90.8
	SDC	47.1	49.2	53.4	60.2	63.8	68.2	73.4	79.9
	LPCC	40.1	43.2	45.8	49.9	53.7	59.1	63.2	65.9
	LPC	36.3	31.3	31.3	35.0	40.0	47.5	57.5	71.3
Kaggle	**MFCC**	41.0	46.0	43.0	46.0	50.0	50.0	50.0	52.0
	SDC	37.2	38.4	40.9	41.2	43.7	44.1	45.5	47.8
	LPCC	30.5	32.7	34.1	36.7	38.3	39.6	41.1	42.2
	LPC	24.0	31.0	32.0	33.0	33.0	18.0	24.0	24.0

minimum equal rate of 9.7% with 256 components whereas on Kaggle corpus it achieves EER of 29% with 256 mixture components.

Table 5.2 shows a comparison between MFCC, LPC, LPCC and SDC features for accent recognition where Accuracy(%) is used as a metric for performance comparison between them. As a classifier GMM-UBM is used. The objective of this comparison is to identify the best features for accent recognition. The comparison shows that MFCC compared to others demonstrates better Accuracy on both Urdu and Kaggle corpuses.

The experimental results show that MFCC outperforms all other specific features on both Kaggle and Urdu speech corpuses.

5.3 Comparison between GMM-UBM and I-vector using MFCC features

Table 5.3 shows a comparison between GMM-UBM and I-vector methods for accent recognition. Both are trained using MFCC features and different mixture components. It can be seen that GMM-UBM achieves minimum EER of 9.7% on Urdu corpus with 256 components whereas I-vector method demonstrate min EER of 42.5%. Similarly GMM-UBM outperforms I-vector on Kaggle corpus.

Table 5.3: EER(%) based comparison between GMM-BM and I-vector methods using MFCC features

GMM Components		2	4	8	16	32	64	128	256
Urdu	**GMM**	31.8	30.5	29.3	26.8	21.8	18.0	13.0	9.7
	I-vector	52.5	37.9	40.0	41.3	42.1	43.8	40.0	42.5
Kaggle	**GMM**	34.0	34.0	34.0	31.0	30.0	29.0	29.3	31.0
	I-vector	38.0	32.5	36.0	36.0	32.3	32.0	34.0	33.5

Table 5.4: EER(%) obtained with MFCC using GMM-UBM and the proposed GMM-FM method

GMM Components		2	4	8	16	32	64	128	256
Urdu	**GMM-UBM**	31.8	30.5	29.3	26.8	21.8	18.0	13.0	9.7
	GMM-FM	38.8	31.3	25.1	22.6	19.7	15.1	11.3	8.4
Kaggle	**GMM-UBM**	34.0	34.0	34.0	31.0	30.0	29.0	29.3	31.0
	GMM-FM	37.8	31.5	30.0	29.0	27.5	26.0	27.0	26.0

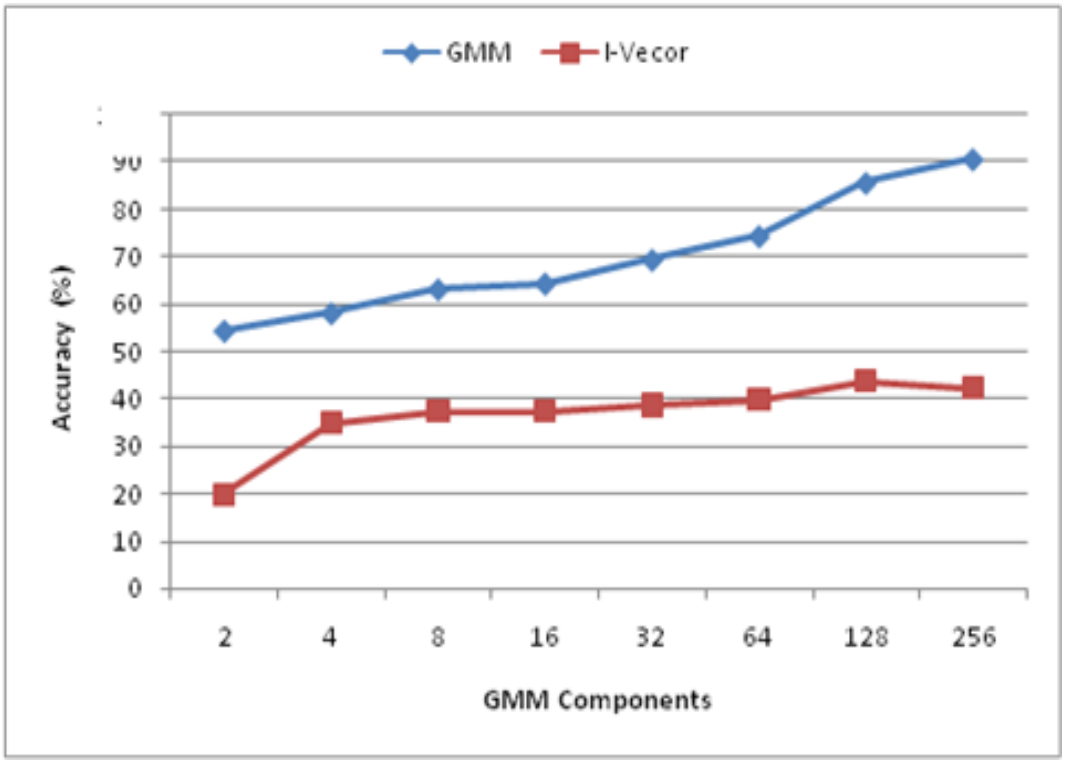
Figure 5.1 shows accuracy based comparison between GMM-UBM and I-vector methods using the MFCC features. The experimental results show that accuracy varies with respect to different mixture components and GMM-UBM outperforms I-vector method.

5.4 Comparison between GMM-UBM and the proposed GMM-FM method

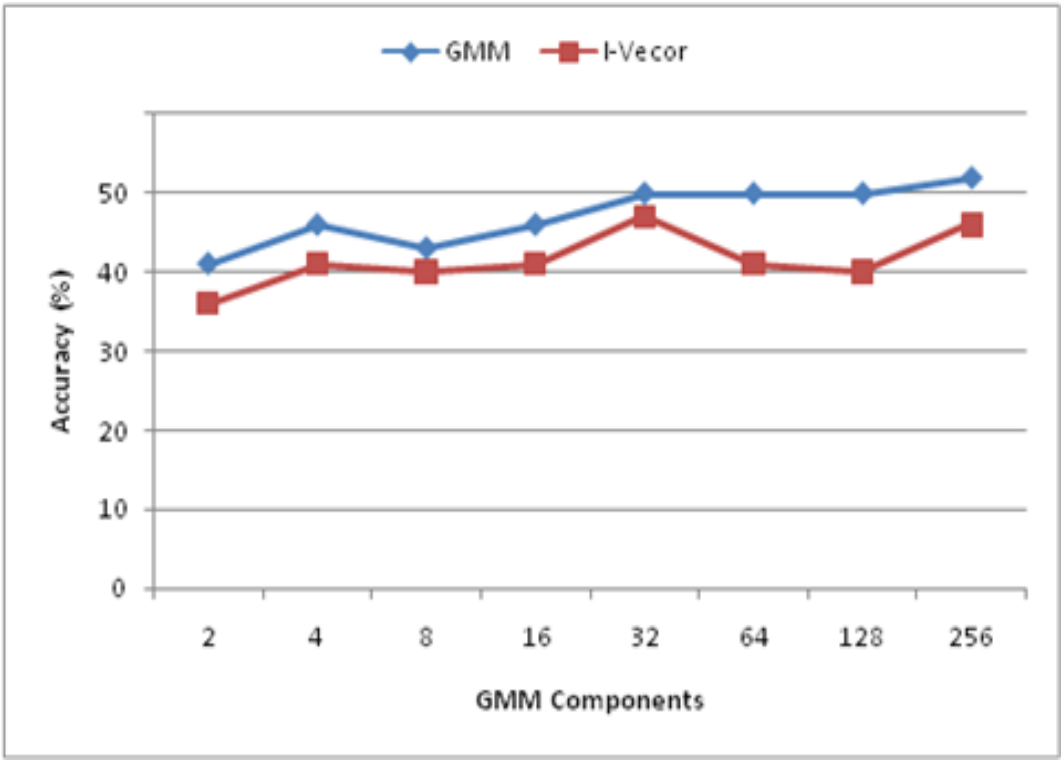
Table 5.4 shows the EER(%) based results obtained with the GMM-UBM method and the proposed methods. Difference between them is the proposed feature mapping which is only used in the GMM-FM method compared to GMM-UBM. It can be seen that GMM-UBM on Urdu corpus with 256 components gives EER of 9.7% whereas the proposed method gives min EER of 8.4%. So, the proposed method provides an improvement of almost 1.3% EER.

On Kaggle corpus, GMM-UBM and the GMM-FM method achieve 26% and 31% EER, respectively. Improvement is almost 5%. This shows that the proposed feature mapping efficiently improves the performance of GMM-UBM for accent recognition. Based on above experimental results it is observed that GMM-UBM and MFCC are best classifier and feature pairs for accent recognition, respectively. In the proposed GMM-FM method we use MFCC features the GMM-UBM classifier.

Table 5.5 shows the accuracy based comparison between GMM-UBM and the GMM-FM method using the MFCC features. It can be seen that the GMM-FM method demonstrate 1.2% and 3% better accuracy rates compared to GMM-UBM on Urdu and Kaggle corpuses, respectively.



(a)



(b)

Figure 5.1: Accent recognition accuracy achieved on (a) Urdu (b) Kaggle corpus with GMM-UBM and I-vector methods using MFCC features

Table 5.5: Accuracy(%) obtained with proposed GMM-FM method and GMM-UBM for accent recognition

GMM Components		2	4	8	16	32	64	128	256
Urdu	**GMM-UBM**	54.5	58.3	63.3	64.5	69.5	74.5	85.8	90.8
	GMM-FM	57.0	58.3	64.5	70.8	74.5	83.3	89.5	92.0
Kaggle	**GMM-UBM**	41.0	46.0	43.0	46.0	50.0	50.0	50.0	50.0
	GMM-FM	32.0	42.0	42.0	45.0	46.0	51.0	53.0	53.0

Table 5.6: Accent recognition accuracy achieved with GMM-UBM, I-vector, SVM and the proposed GMM-FM method

Corpus	GMM-UBM	I-vector	Linear SVM	SVM-Poly-Degree-2	SVM-Poly-Degree-3	SVM-RBF	GMM-FM
Urdu	90.8	43	55	31.25	30.2	61.25	92
Kaggle	50	47	42	27	20	44.3	53

5.5 Comparison between proposed GMM-FM method and SVM

Table 5.6 summarizes the accuracy rates achieved on both corpuses using GMM-UBM, I-vector, Linear SVM, SVM-RBF, SVM-Polynomial and the proposed method. MFCC features are used. On Urdu corpus the proposed method demonstrates the best accuracy rate of 92% followed by GMM-UBM (90.8%), SVM-RBF (61.25%) and Linear SVM (55%). The SVM with polynomial kernels with degrees 2 and 3 do not perform well. It can be seen that the SVM accuracy decreases with increase in the polynomial degree.

Similarly on Kaggle corpus the proposed method achieves accuracy of 53% and outperforms all other classifiers. The accuracy achieved on Kaggle is low compared to Urdu because the Kaggle corpus is text dependent where different speakers record the same English paragraph in their native accents. A comparison between GMM-FM method and SVM is shown in Table 5.6. GMM-FM method gives more accurate result on both Urdu and Kaggle corpus. It gives 92% accuracy on Urdu accent corpus and 53% on Kaggle corpus.

5.6 Forensic Speaker Recognition with and without Accent classification

This section presents experimental results for forensic speaker recognition. Accent classification (AC) prior to speaker recognition is investigated and the experimental results for forensic speaker recognition with and without AC are presented. The experiments are performed on an Urdu Forensic Speech Corpus. Comparison between GMM-FM GMM-FM method and GMM-UBM is shown with and without Accent Classification.

Table 5.7: EER(%) based ASR results obtained with GMM-UBM and the proposed GMM-FM method with and without Accent Recognition

GMM Components		2	4	8	16	32	64	128	256
Without AC	**GMM-UBM**	28.8	21.3	18.1	16.8	15.0	15.2	12.5	11.4
	GMM-FM	20.3	18.8	16.3	15.0	13.8	12.5	11.8	10.4
With AC	**GMM-UBM**	25.0	20.0	17.1	14.6	12.9	10.4	10.0	9.6
	GMM-FM	20.0	12.9	11.3	9.2	8.8	7.1	6.7	7.1

Table 5.8: Accuracy(%) based ASR results obtained with GMM-UBM and the proposed GMM-FM method with and without Accent Recognition

GMM Components		2	4	8	16	32	64	128	256
Without AC	**GMM-UBM**	42.5	50.0	56.3	68.8	68.8	71.3	71.3	68.8
	GMM-FM	53.8	62.5	66.3	66.3	70.0	68.8	72.5	70.0
With AC	**GMM-UBM**	61.3	71.3	73.8	82.5	83.8	84.0	82.5	83.8
	GMM-FM	77.5	78.8	86.3	85.0	85.0	86.3	87.5	87.5

Table 5.7 shows EER rate achieved for speaker recognition with and without AC using different mixture components. GMM-UBM and the proposed GMM-FM with 256 mixtures components achieve 11.4% and 10.4% EER without AC (see Table 5.7). Where as they achieve 9.6% and 7.1% EER with AC, respectively. So using AC as a pre-processing step, the improvement in speaker recognition EER is 1.8% and 3.3% for GMM-UBM and GMM-FM respectively.

Similarly the accuracy rates shown in Table 5.8 shows that with AC better speaker recognition rates are obtained compared to without AC based speaker recognition. Table 5.8 shows accuracy achieved for speaker recognition with and without AC. GMM-UBM and the proposed GMM-FM method with 256 mixtures components achieve 68.8% and 70% Accuracy without AC. Where as they achieve 83.8% and 87.5% accuracy with AC, respectively. The GMM-FM method in both cases with and without AC outperforms GMM-UBM by achieving better accuracy rates.

5.7 Summary

This chapter concludes all the results and findings of this thesis. Results obtained with MFCC, SDC, LPC and LPCC are compared on Urdu corpus. These results are obtained with different classifiers and evaluated with accuracy and equal error rate. It explains the experimental setup that all the collected recordings are divided into test and training parts and then passed though MFCC for feature extraction. These features are trained on classifiers and then tested recordings are tested on these trained features for results. Results then shows which feature gives better results with which classifier is more preferable for accent recognition system.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Overview

This chapter summarizes the accent recognition results. It presents conclusion and future work.

6.2 Conclusion

In this thesis, we investigated the accent recognition for Urdu language. We have also discussed some prominent features and classifiers and highlighted their accent recognition performances. A new method for extraction of accent information from Urdu speech signals is also presented. Four different Urdu accent are recognized which are Balochi, Pashto, Punjabi and Sindhi and then use the accent for forensic speaker recognition.

A person accent shows the person's background which helps to identify the person cultural and the territorial background. Experiments are conducted using different features and classifiers on three corpuses whereas classifier are trained for accent recognition are GMM-UBM, I-vector and SVM classifiers are trained. The experimental results show that MFCC features compared to LPCC, LPC and SDC features demonstrates better Urdu accent recognition performances.

The GMM-UBM classifier compared to I-vector and SVM methods achieves better Urdu accent recognition results. The proposed method which is based on GMM-UBM and a feature mapping process outperform the GMM-UBM classifier by 1.3% (EER) and 1.2% (Accuracy). Compared to RBF-SVM, Linear-SVM and Polynomial-SVM it achieves 30.7%, 37%, and 60% better accuracy rates, respectively.

The forensic speaker recognition results show that GMM-UBM and the proposed method with accent classification as a pre-processing step improve the

speaker recognition rates. However, the proposed method demonstrates 2.5% and 3.7% better EER and accuracy rates compared to GMM-UBM.

6.3 Future work

This thesis has some limitations. Features examined in this thesis are most commonly used for accent recognition and speaker recognition. Other features can also be examined for training of classifiers. Classifiers used in this thesis can be extended to deep learning algorithms. Deep learning is not used because of the short corpus.

Corpuses constructed have short databases due to which classifiers with deep learning are not corporative. In future a large database can be constructed so that more tests can be applied and better results can be gained using proposed methodology.

It covers only four different Urdu accents. Other accents can be incorporated and the accent recognition can be evaluated. For example, Punjabi have many accents like Potohari, Saraiki, Hindko and more. In the same way other languages like Sindhi and Balochi have their own accents and dialects. These accents can also be examined for regional language. In future Punjabi and Pashto regional accents can also be tested for Urdu accent recognition. Moreover in future we can incorporate more accents e.g Pashto language can also be recognized using proposed method.

REFERENCES

[1] Bartkova, K. and Jouvet, D. On using units trained on foreign data for improved multiple accent speech recognition. *Speech Communication*, 2007. 49(10-11): 836–846.

[2] Huang, C., Chen, T. and Chang, E. Accent Issues in Large Vocabulary Continuous Speech Recognition. *International Journal of Speech Technology*, 2004. 7(2): 141–153.

[3] Bahari, M. H., Saeidi, R., Van hamme, H. and Van Leeuwen, D. Accent recognition using i-vector, Gaussian Mean Supervector and Gaussian posterior probability supervector for spontaneous telephone speech. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013. 2: 7344–7348.

[4] Sinha, S., Jain, A. and Agrawal, S. S. Acoustic-Phonetic Feature Based Dialect in Hindi Speech. *International Journal on Smart Sensing and Intelligent Systems*, 2015. 8(1): 235–254.

[5] Najafian, M., Safavi, S., Weber, P. and Russell, M. Identification of British English regional accents using fusion of i-vector and multi-accent phonotactic systems. *Odyssey*. 2016. 132–139.

[6] Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V. and Wellekens, C. Automatic speech recognition and speech variability: A review. *Speech Communication*, 2007. 49(10-11): 763–786.

[7] Mannepalli, K., Narahari Sastry, P. and Rajesh, V. Accent detection of Telugu speech using supra-segmental features. *International Journal of Soft Computing*. 2015, vol. 10. 287–292.

[8] Behravan, H. *Dialect and accent recognition*. Ph.D. Thesis. 2012.

[9] Qasim, M., Nawaz, S., Hussain, S. and Habib, T. Urdu speech recognition system for district names of Pakistan: Development, challenges and solutions. *Conference on Coordination and Standardization of Speech Databases and Assessment Techniques*. 2016. 28–32.

[10] Amino, K., Osanai, T., Kamada, T., Makinae, H. and Arai, T. Historical and procedural overview of forensic speaker recognition as a science. In: *Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism*. 3–20. 2012.

[11] T. Rahman. Language Policy and Localization in Pakistan: Proposal for a Paradigmatic Shift. *SCALLA Conference on Computational Linguistics*, 2004: 100.

[12] Gordon-Salant, S., Yeni-Komshian, G. H. and Fitzgibbons, P. J. Recognition of accented English in quiet by younger normal-hearing listeners and older listeners with normal-hearing and hearing loss. *The Journal of the Acoustical Society of America*, 2010. 128(1): 444–455.

[13] Vacher, M., Portet, F., Fleury, A. and Noury, N. Development of audio sensing technology for ambient assisted living: Applications and challenges. *International Journal of E-Health and Medical Communications*, 2011. 2(1): 35–54.

[14] Poorjam, A. H. Speaker Profiling for Forensic Applications, 2014.

[15] Algabri, M., Mathkour, H., Bencherif, M. A., Alsulaiman, M. and Mekhtiche, M. A. Automatic Speaker Recognition for Mobile Forensic Applications. *Mobile Information Systems*, 2017: 1–6.

[16] Morrison, G. S. and Enzinger, E. Score based procedures for the calculation of forensic likelihood ratios – Scores should take account of both similarity and typicality. *Science and Justice*, 2018. 58(1).

[17] Stefanus, I., Sarwono, R. J. and Mandasari, M. I. GMM based automatic speaker verification system development for forensics in Bahasa Indonesia. *5th IEEE International Conference on Instrumentation, Control, and Automation*. 2017. 56–61.

[18] Petkar, H. J. and Kakkad, N. P. ASR In Pursuit Of Forensics Investigation. *International Journal of Electronics, Communication and Soft Computing Science & Engineering*, 2015. 1: 159–163.

[19] Eriksson, A. Aural/acoustic vs. Automatic methods in forensic phonetic case work. In: *Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism*. 41–69. 2012.

[20] Alexander, A., Dessimoz, D., Botti, F. and Drygajlo, A. Aural and automatic forensic speaker recognition in mismatched conditions. *International Journal of Speech, Language and the Law*, 2005. 12(2): 214–234.

[21] Solewicz, Y. A., Jessen, M. and Van Der Vloed, D. Null-Hypothesis LLR: A

proposal for forensic automatic speaker recognition. *Annual Conference of the International Speech Communication Association*. 2017. 2849–2853.

[22] Eriksson, A. Tutorial on forensic speech science. Part I: Forensic phonetics. *International Speech Communication Association*, 2005. (2002): 83–96.

[23] Hansen, J. H., Gray, S. S. and Kim, W. Automatic voice onset time detection for unvoiced stops (/p/,/t/,/k/) with application to accent classification. *Speech Communication*, 2010. 52(10): 777–789.

[24] Mannepalli, K., Sastry, P. N. and Suman, M. MFCC-GMM based accent recognition system for Telugu speech signals. *International Journal of Speech Technology*, 2016. 19(1): 87–93.

[25] Lazaridis, A. and Khoury, E. Swiss French Regional Accent Identification, 2014.

[26] Brown, G. Exploring forensic accent recognition using the Y-ACCDIST system. *Sixteenth Annual Conference of the International Speech Communication Association*, 2016: 305–308.

[27] Afnan, S. *Comparison GMM and SVM Classifier for Automatic Speaker Verification*. Ph.D. Thesis. 2015.

[28] Bimbot, F. E. E., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska-Delac, D. and Reynolds, D. A. A Tutorial on Text-Independent Speaker Verification. *EURASIP Journal on Applied Signal Processing*, 2004. 4: 430–451.

[29] Bhattacharjee, U. and Sarmah, K. GMM-UBM Based Speaker Verification in Multilingual Environments. *International Journal of Computer Science Issues*, 2012. 9(6): 373–380.

[30] Chaudhari, R. H., Waghmare, K. and Gawali, B. W. Accent Recognition using MFCC and LPC with Acoustic Features. *International Journal of Innovative Research in Computer and Communication Engineering*, 2015. 3(3): 2128–2134.

[31] Torres-Carrasquillo, P. A., Singer, E., Kohler, M. A., Greene, R. J., Reynolds, D. A. and Deller, J. R. Approaches to language identification using Gaussian mixture models and shifted delta cepstral features. *Int. Conf. Spoken Language Processing*, 2002. 2: 89–92.

[32] Rafaqat Ali, A., Irtza, S., Farooq, M. and Hussain, S. Accent Classification among Punjabi, Urdu, Pashto, Saraiki and Sindhi accents of Urdu Language. *The Conference on Language and Technology*. 2014. 1–7.

[33] Rauf, S., Hameed, A., Habib, T. and Hussain, S. District names speech corpus for Pakistani Languages. *IEEE International Conference on Asian Spoken Language Research and Evaluation*. 2015. 207–211.

[34] LYN, G. *Gender and accent identification for Malaysian English using MFCC and Gaussian mixture model*. Ph.D. Thesis. 2013.

[35] Gfrörer, S., Gfroerer, S. and Wiesbaden, D. Auditory-instrumental forensic speaker recognition. *Eighth European Conference on Speech Communication and Technology*. 2003. 705–708.

[36] Dileep, A. D. and Chandra Sekhar, C. Hmm based intermediate matching kernel for classification of sequential patterns of speech using support vector machines. *IEEE Transactions on Audio, Speech and Language Processing*, 2013. 21(12): 2570–2582.

[37] Maher, R. Audio forensic examination. *IEEE Signal Processing Magazine*, 2009. 26(2): 84–94.

[38] Wang, H. and Zhang, C. Forensic Automatic Speaker Recognition Based on Likelihood Ratio Using Acoustic-phonetic Features Measured Automatically. *Journal of Forensic Science and Medicine*, 2015. 1(2): 119.

[39] Bent, T. and Frush Holt, R. The influence of talker and foreign-accent variability on spoken word identification. *The Journal of the Acoustical Society of America*, 2013. 133(3): 1677–1686.

[40] Biadsy, F., Hirschberg, J. and Ellis, D. P. Dialect and accent recognition using phonetic-segmentation supervectors. *Annual Conference of the International Speech Communication Association*. 2011, vol. 2. 745–748.

[41] Kinnunen, T., Saeidi, R., Sedlak, F., Lee, K. A., Sandberg, J., Hansson-Sandsten, M., Li, H., Sedlák, F., Lee, K. A., Sandberg, J., Hansson-Sandsten, M. and Li, H. Low-variance multitaper MFCC features: A case study in robust speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 2012. 20(7): 1990–2001.

[42] Ilina, O., Koval, S. and Khitrov, M. Phonetic Analysis in Forensic Speaker Identification . an Example of Routine Expert Actions . *International Conference of Phonetic Sciences*. 1999. 157–160.

[43] Marescal, F. Adding a Parametric Approach To Forensic. *Problems of Forensic Sciences*, 2001: 254–267.

[44] Rose, P. Technical forensic speaker recognition: Evaluation, types and testing of evidence. *Computer Speech & Language*, 2006. 20(2-3): 159–191.

[45] Barry, W. J., Hoequist, C. E. and Nolan, F. J. An approach to the problem of regional accent in automatic speech recognition. *Computer Speech and Language*, 1989. 3(4): 355–366.

[46] Rober C. Maher and Maher, R. C. Lending an Ear in the Courtroom: Forensic Acoustics. *Acoustics Today*, 2015. 11(3): 1–9.

[47] Brown, G. Automatic Accent Recognition Systems and the Effects of Data on Performance. *Odyssey Speech*, 2016: 94–100.

[48] Ma, Z. and Fokoué, E. A Comparison of Classifiers in Performing Speaker Accent Recognition Using MFCCs. *Open Journal of Statistics*, 2014. 04(04): 258–266.

[49] Bhatia, M., Singh, N. and Singh, A. Speaker Accent Recognition by MFCC Using K- Nearest Neighbour Algorithm: A Different Approach. *International Journal of Advanced Research in Computer and Communication Engineering*, 2015. 4(1): 153–155.

[50] Patel, T. B. and Patil, H. A. Combining Evidences from Mel Cepstral , Cochlear Filter Cepstral and Instantaneous Frequency Features for Detection of Natural vs . Spoofed Speech. *Annual Conference of the International Speech Communication Association*, 2015: 1–5.

[51] Hanani, A., Russell, M. and Carey, M. J. Speech-based identification of social groups in a single accent of British English by humans and computers. *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2011. 4876–4879.

[52] Brown, G. and Wormald, J. Automatic sociophonetics: Exploring corpora with a forensic accent recognition system. *The Journal of the Acoustical Society of America*, 2017. 142(1): 422–433.

[53] Abbas, A. W., Ahmad, N. and Ali, H. Pashto Spoken Digits Database for the Automatic Speech Recognition Research. *IEEE International Conference on Automation and Computing*. 2012, 2013.

[54] Huang, R., Hansen, J. H. L. and Angkititrakul, P. Dialect/Accent Classification Using Unrestricted Audio. *IEEE Transactions on Audio, Speech and Language Processing*, 2007. 15(2): 453–464.

[55] Lazaridis, A., Goldman, J.-p. and Avanzi, M. Syllable-based Regional Swiss French Accent Identification using Prosodic Features. *In Cahiers of French Linguistics*. 2014, vol. 31. 297–307.

[56] Drygajlo, A. Automatic speaker recognition for forensic case assessment

and interpretation. In: *Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism*. 21–39. 2012.

[57] Reddy, D. Speech recognition by machine: A review. *Proceedings of the IEEE*, 1976. 64(4): 501–531.

[58] Kumar, G. S., Raju, K. A. P., Rao, M. and Satheesh, C. P. Speaker Recognition Using Gmm. *International Journal of Engineering Science and Technology*, 2010. 2(6): 2428–2436.

[59] Stevenage, S. V., Clarke, G. and Mcneill, A. The accent effect in voice recognition. *Journal of Cognitive Psychology*, 2012: 37–41.

[60] Huang, C., Chang, E., Zhou, J. and Lee, K.-f. Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition. *International Conference on Spoken Language Processing*. 2000. 818–821.

[61] Behravan, H., Hautamäuki, V., Siniscalchi, S. M., Kinnunen, T., Lee, C.-h. H., Hautamauki, V., Siniscalchi, S. M., Kinnunen, T. and Lee, C.-h. H. Introducing attribute features to foreign accent recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2014. 5332–5336.

[62] Rizwan, M. and Anderson, D. V. A weighted accent classification using multiple words. *Neurocomputing*, 2017. 277: 1–9.

[63] Morrison, G. S. Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison. *Science & Justice*, 2014. 54(3): 245–256.

[64] Vloed, D. V. D., Bouten, J. and van Leeuwen, D. A forensic speaker recognition database and some first experiments. *NFI-FRITS*, 2014: 6–13.

[65] Alexander, A., Forth, O., Atreya, A. A. and Kelly, F. VOCALISE : A forensic automatic speaker recognition system supporting spectral , phonetic , and user-provided features. *Odyssey*, 2016.

[66] Behravan, H., Hautamäki, V. and Kinnunen, T. Factors affecting i-vector based foreign accent recognition: A case study in spoken Finnish. *Speech Communication*, 2015. 66: 118–129.

[67] Tverdokhleb, E., Dobrovolskyi, H., Keberle, N. and Myronova, N. Implementation of accent recognition methods subsystem for eLearning systems. *IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*. 2017, vol. 2.

1037–1041.

[68] Rao, K. and Sak, H. Multi-accent speech recognition with hierarchical grapheme based models. *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2017. 4815–4819.

[69] Alexander, A., Forth, O., Jessen, M. and Jessen, M. Speaker recognition with Phonetic and Automatic Features using VOCALISE software. *International Conference of Association for Forensic Phonetics and Acoustics*. 2005. 3–4.

[70] CR, R. Review of Algorithms and Applications in Speech Recognition System. *International Journal of Computer Science and Information Technologies*, 2014. 5(4): 5258–5262.

[71] Shrawankar, U. and Thakare, V. M. Techniques for Feature Extraction In Speech Recognition System : A Comparative Study. *arXiv preprint arXiv*, 2013.

[72] Paliwal, K. K., Paliwal, K. K., Kleijn, W. B. and Kleijn, W. B. Quantization of LPC Parameters. *Speech Coding and Synthesis*, 1995: 433–466.

[73] Maesa, A., Garzia, F., Scarpiniti, M. and Cusani, R. Text Independent Automatic Speaker Recognition System Using Mel-Frequency Cepstrum Coefficient and Gaussian Mixture Models. *Journal of Information Security*, 2012. 03(04): 335–340.

[74] Hanani, A., Russell, M. J. and Carey, M. J. Human and computer recognition of regional accents and ethnic groups from British English speech. *Computer Speech and Language*, 2013. 27(1): 59–74.

[75] Campbell, W., Campbell, J., Reynolds, D., Singer, E. and Torres-Carrasquillo, P. Support vector machines for speaker and language recognition. *Computer Speech & Language*, 2006. 20(2-3): 210–229.

[76] Ken Chen, Hasegawa-Johnson, M. and Cohen, A. An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model. *IEEE International Conference on Acoustics, Speech, and Signal Processing*. 2004, vol. 1. I–509–12.

[77] Campbell, W., Sturim, D. and Reynolds, D. Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, 2006. 13(5): 308–311.

[78] Scheffer, N. and Bonastre, J.-f. UBM-GMM Driven Discriminative Approach for Speaker Verification. *The Speaker and Language Recognition Workshop*. 2006, vol. 00. 1–7.

[79] Reynolds, D. A., Quatieri, T. F. and Dunn, R. B. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 2000. 10(1-3): 19–41.

[80] Garcia-romero, D. and Espy-wilson, C. Y. Analysis of I-vector Length Normalization in Speaker Recognition Systems. *International Conference of Speech Communication Association.* 2011. 249–252.

[81] Cortes, C. and Vapnik, V. Support-Vector Networks. *Machine Learning*, 1995. 20(3): 273–297.

[82] Gauvain, J.-L. and Chin-Hui Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 1994. 2(2): 291–298.