Acoustic Feature based Language Identification Using Single Word Utterances with Fixed Vocabulary

Session: 2012 – 2019

Submitted by:

Farah Adeeba 2012-PhD-CS-17

Supervised by:

Dr. Sarmad Hussain

Department of Computer Science and Engineering

University of Engineering and Technology

Lahore Pakistan

Acoustic Feature based Language Identification Using Single Word Utterances with Fixed Vocabulary

Submitted to the faculty of the Computer Science and Engineering Department of the University of Engineering and Technology Lahore in partial fulfillment of the requirements for the Degree of

Doctor of Philosophy

in

Computer Science.

Internal Examiner	**External Examiner**
Signature:	Signature:
Prof. Dr. Sarmad Hussain	Dr. Agha Ali Raza
Professor, KICS	Signature:
UET, Lahore	Dr. Muhammad Ali Tahir
Chairman	**Dean**
Signature:	Signature:
Prof. Dr. Shazia Arshad	Prof. Dr. Tahir Izhar

Department of Computer Science and Engineering

University of Engineering and Technology

Lahore Pakistan

Declaration

I declare that the work contained in this thesis is my own, except where explicitly stated otherwise. In addition this work has not been submitted to obtain another degree or professional qualification.

Signed: _____

Date: _____

Acknowledgments

Sometimes our light goes out but is blown into flame by another human being. Each of us owes deepest thanks to those who have rekindled this light. Among all those, first I would like to express my sincere gratitude and veneration to my supervisor, Prof. Sarmad Hussain, for giving me an opportunity to work in the exciting area of speech processing and providing his meticulous guidance throughout my research work. His guidance that maintained my focus and the freedom to develop and express my scientific ideas was of great importance to me to raise my individual abilities. His sage advices, insightful criticisms and always providing sincere help to the successful completion of this research project. I offer my humble thanks to him, whose guidance and advice is beyond compare, whose affection and encouraging attitude is matchless.

Special thanks to my contemporary and colleagues, Miss Sana Shams and Mr. Toqeer Ehsan for their generous help, moral support during our time in the same laboratory and constant readiness to help and share the lab resources. I would like to acknowledge Mr. Ashok Kumar Khatri, Mr. Asad Mustafa and Mr. Inaam-ullah Torwali of Center for Language Engineering, for their assistance in development of phonetic lexicon of Sindhi, Punjabi and Pashto languages. I am grateful to administrative staff of Department of Computer Science and Engineering(CS&E), especially to Dr. Muhammad Aslam for his all-time available help and concern and cooperative behavior during my stay in UET.

Words would never be able to fathom the depth of feeling for my reverend parents who raised me up to here, Mr. Mushtaq Ahmad and Mrs. Zarina Bibi, for the immeasurable love, encouragement and understanding. Many endless gratitude to my sisters, Attiya Muffarat, Nafeesa Mushtaq and Maria Mushtaq for their constant help, motivation, concern for my PhD encouragement and loving support in every thick and thin. You have a big contribution in making me who I am today.

I would like to especially thank my best friend Qurat-ul-Ain Akram who was always there to support and buck up me whenever I lost hope. Your friendship and encouragement means a lot to me.

Last but not least, my endless gratitude to my dear husband Mr. Shahid Saleem, for his constant help, motivation and concern for my PhD. His loving support in every thick and thin made this achievement possible for me. I would also like to express my deepest thanks and love to my kids (Noor ul Ain and Arham Shahid), mother in law, Zahid Saleem and Shahzad Saleem for being extremely understanding and constantly care towards my daily life.

Farah Adeeba

April, 2019

To my husband

Shahid Saleem

Contents

List of Figures

List of Tables

Abbreviations

ANN	**A**rtificial **N**eural **N**etwork
ASR	**A**utomatic **S**peech **R**ecognition
BLSTM	**B**idirectional **L**ong **S**hort **T**erm **M**emory
BW	**B**aum **W**elch
CE	**C**ross **E**ntropy
CNN	**C**onvolution **N**eural **N**etwork
DCT	**D**iscrete **C**osine **T**ransformation
DNN	**D**eep **N**eural **N**etwork
EER	**E**qual **E**rror **R**ate
EM	**E**xpectation-**M**aximization
ERB	**E**quivalent **R**ectangular **B**andwidth
FFT	**F**ast **F**ourier **T**ransformation
GFCC	**G**ammatone **F**requency **C**epstral **C**oefficients
GMM	**G**aussian **M**ixture **M**odel
HMM	**H**idden **M**arkov **M**odel
IPA	**I**nternational **P**honetic **A**lphabet
JFA	**J**oint **F**actor **A**nalysis
LID	**L**anguage **I**dentification
LCD	**L**inguistic **D**ata **C**onsortium
LRE	**L**anguage **R**ecognition **E**valuation
LSTM	**L**ong **S**hort **T**erm **M**emory
LFCC	**L**inear **F**requency **C**epstral **C**oefficients
LVCSR	**L**arge **V**ocabulary **C**ontinuous **S**peech **R**ecognition
MFCC	**M**el **F**requency **C**epstral **C**oefficients

MAP	**M**aximum-**A**-**P**osteriori
NIST	**N**ational **I**nstitute of **S**tandard and **T**echnologies
PPRLM	**P**arallel **P**hone **R**ecognition and **L**anguage **M**odeling
PLP	**P**erceptual **L**inear **P**erdiction
PLDA	**P**robablistic **L**inear **D**iscriminant **A**nalysis
PPR	**P**arallel **P**hone **R**ecognizer
RNN	**R**ecurrent **N**eural **N**etwork
SAMPA	**S**peech **A**ssement **M**ethods **P**honetic **A**lphabet
SDC	**S**hifted **D**elta **C**epstral
SNR	**S**ignal to **N**oise **R**atio
SVM	**S**upport **V**ector **M**achine
UBM	**U**niversal **B**ackground **M**odel
VSM	**V**ector **S**pace **M**odel
WCCN	**W**ithin **C**lass **C**ovariance **N**ormalization
WER	**W**ord **E**rror **R**ate

Abstract

Human speech comprises of multiple aspects of information including what is being said i.e. message, who is speaking i.e. speaker (identity, gender, age etc.), language spoken, environment and emotions. The task of machine to infer language from a speech utterance independent of speaker and topic is called spoken language identification (LID). Generally, phonetic, phonological, morphological, syntactic and semantic information of a language is used to discriminate it from other languages. With the advancement in technology, communication among people around the world from different linguistic backgrounds is increasing gradually, resulting in the requirement of automatic speech recognition (ASR) service. To facilitate speech recognition it is helpful to identify which language is being spoken.

In a multilingual country like Pakistan, where around 69 languages are currently being spoken, the automatic language identification systems have special significance. According to the 1998 census of Pakistan, Balochi, Pashto, Punjabi, Saraiki, Sindhi and Urdu languages have large population as compared to the other languages. Different LID systems are available for different languages but no LID system is available for languages spoken in Pakistan. Hence, there is dire need to develop a LID system for these languages.

To cope with the issue of unavailability of publicly available speech corpora of aforementioned Pakistani languages for language identification research, a 10.43 hours speech corpus is designed. This speech corpus is recorded from 316 native speakers differing in gender, age, demographics and educational background. This speech data is recorded over telephonic channel with sampling frequency of 8 KHz. The transcription of speech corpus in X-SAMPA format and in orthographic form is also prepared. This corpus minimizes the barrier of data availability for the development of speech processing applications e.g. speaker recognition and speech recognition for these languages.

A variety of state-of-the-art language identification approaches are compared and effectiveness of these approaches for identification of Pakistani languages is analyzed. In addition, a set of different acoustic features are investigated and their

impact on system accuracy is observed. In order to increase the recognition accuracy, different configuration models are investigated. The performance of the systems is evaluated on the Dataset-1 (from 0.27 sec to 1.5 sec) and Dataset-2 (maximum duration of 3 sec).

The novel language identification approach based on bidirectional long short-term memory neural network is proposed. This approach is evaluated on two different datasets to examine the impact of utterance duration on accuracy. Effect of test data duration is also analyzed (from 0.27 sec to 1.5 sec) and it is observed that with very short duration as 0.4 sec an accuracy of over 50% can be achieved. Moreover, capsule network based language identification system is also proposed and Equal Error Rate of 14.42% and 10.27% is achieved on Dataset-1 and Dataset-2, respectively. Experiments demonstrated that proposed approaches perform better as compared to the existing approaches.

Chapter 1

Introduction

With the development in communication technology, communication among people around the all over the world having different linguistic backgrounds is gradually increasing, resulting into the need of services like automatic speech recognition (ASR) and speech to speech translation. For speech recognition in a multilingual context it is mandatory to identify which language is being spoken. Spoken language identification (LID) is the task to identify the language from a speech utterance [60]. In speaker identification and ASR tasks, only the speaker identity or information of the content of the utterance is unavailable. However, in language identification both the content of the utterance and the identification of speaker is not available, which is an added challenge [4]. The human listening depends on phonotactic and prosody cues for language identification. Similarly, automatic identification of spoken language is based on different cues, related to phonetic, phonological, morphological and syntactic information of a language. For automatic LID, these cues are extracted from the speech signal.

The number of known spoken languages in Pakistan is more than 69 [29]. Therefore, an excellent LID system should make use of different aspects of speech information that can discriminate languages from each other, very accurately and minutely. Moreover, LID system should be flexible enough to handle the diversity of different speakers. Although, amongst more than 69 spoken languages in Pakistan, the majority of the Pakistan's population speaks a set of six languages

including Balochi, Pashto, Punjabi, Saraiki, Sindhi and Urdu : among all speakers, about 95% of speakers use of only 8% of languages spoken in Pakistan [1]. This study focuses on language identification from very short utterances of closely related languages of Pakistan including Balochi, Pashto, Punjabi, Saraiki, Sindhi and Urdu.

1.1 Motivation

Spoken language identification plays a crucial role as a preliminary step of multilingual speech processing applications. Spoken language identification will enhance the multilingual speech recognition [68], the speech to speech translation performance [110], retrieval of spoken document [10], and user interaction with spoken dialog system [122]. Moreover, LID can also be utilized in international call centers as front-end application by routing the call to particular system or operator depending upon the caller language.

In a multilingual country like Pakistan, where more than 69 languages are currently being spoken [29], the automatic language identification systems have special significance. Different LID systems are available in literature to fulfill this purpose for different languages but no LID system is available for Pakistani languages. This study will make available an LID system that can serve this purpose in context of Pakistani languages. This research will also be useful for retrieval and translation of multimedia content of Pakistani language as more and more local multimedia content is becoming available online.

1.2 Research Gaps

Mostly, research in speech processing specifically language identification is primarily addressed for the world's dominant languages, and for these languages abundant resources are available. However, the relative sparsity of resources of targeted languages (Pashto, Punjabi, Balochi, Saraiki, Urdu and Sindhi) is a major challenge to do research on LID system of these languages.

The LID system performance mainly depends on: (1) reliability of extracted

cues (2)linguistic differences among target languages and(3)amount of information available for feature extraction i.e. utterance duration. The current state-of-art LID system performance degrades as the utterance duration decrease. This becomes very challenging issue when LID systems are deployed in real-world applications such as call center when user response can be merely a word. From the literature, it is evident that, the languages which have different phoneme set can be easily identified as compared to the identification of similar languages. Targeted languages are acoustically related and share a common set of phonemes e.g. Sindhi and Saraiki languages have common set of phonemes.

The work presented in this thesis focuses on these shortcomings and results in availability of the linguistic resources such as text corpus, phonetic lexicon, and speech corpus to develop an LID system for targeted languages. Moreover, by using deep neural network based framework prominent language differences from very short speech signal apart are extracted for the development of a reliable LID system.

1.3 Major Contributions

This thesis focuses on acoustic features based automatic language identification of six major Pakistani languages, especially from short utterances. Major contributions of this thesis together with the scientific publications are listed as follows:

- Following linguistic resources are developed:
  - Developed phonetic lexicon of Sindhi, Pashto, Punjabi and Urdu with the size of 17239, 22305, 84422 and 91280 words, respectively. These lexicon provide pronunciation in IPA and X-SAMPA format.
  - Collected text corpora of Urdu, Punjabi, Pashto, Sindhi and Saraiki.
  - Phonetically rich text corpora are developed for Urdu, Punjabi, Pashto, Sindhi and Saraiki languages.
- Recorded and annotated multilingual speech corpus of Pakistani languages for LID task. This speech corpus comprised of 10 hours of speech recorded

from 316 native speakers. Beside language identification, the dataset can be utilized for other speech processing applications such as speech recognition and speaker recognition.

- Examined different acoustic features for the language identification task. Different combinations of the acoustic features are also investigated in this thesis.

- Investigated different language identification methods for LID of Pakistani languages. These methods are evaluated on the longer utterances and very short utterances i.e. single word.

- Proposed a novel appliance of the bidirectional long short term memory recurrent neural network for LID task. Network architecture search is also carried out to optimize system accuracy. BLSTM network trained on spectrogram and chochleagram features are merged together to take advantage of combined features.

- Proposed an end to end framework based on capsule network for LID. Proposed framework is evaluated on very short test utterance and promising results are achieved.

1.4 Thesis Organization

The organization of the thesis is as follows:

Chapter 2 provides the phonetic overview to understand the similarities and differences across six Pakistani languages namely Punjabi, Pashto, Saraiki, Urdu, Balochi and Sindhi. In phonetic overview of language, population size, major dialects and phonemic inventory (both consonantal and vocalic) is presented. Moreover, phonological differences among languages are also discussed.

Chapter 3 presents some background work relevant to the research of automatic spoken language identification (LID). A brief introduction of LID techniques is

provided, and various state-of-the-art techniques of LID are discussed. Finally, available linguistic resources are presented.

Chapter 4 describes speech corpus's design and collection for spoken language identification of languages spoken in Pakistan. Speech corpus for each language is basically gathered from native speakers. Moreover, transcription of each utterance orthographic form and X-SAMPA format is also prepared. The developed speech corpus (Dataset-2) is used for evaluation of the LID techniques. In addition, isolated word utterances speech corpus (Dataset-1) employed for the LID is also discussed.

Chapter 5 discusses different acoustic features that can be used for LID task. Gammatone frequency cepstral coefficients (GFCC), Mel-frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP) are introduced. Acoustic features based LID techniques i.e. Gaussian mixture model-universal background model (GMM-UBM) and i-vectors are adopted as baseline system. In order to increase the recognition accuracy, different configuration of models are investigated. The performance of the system is evaluated on the Dataset-1 and Dataset-2.

Chapter 6 proposes a novel application of bidirectional long short-term memory (BLSTM) neural network for spoken language identification. Two BLSTM models trained on spectrogram and cochleagram based features are merged together and forwarded to the fully connected network. The performance of the merged BLSTM system is evaluated on Dataset-1.

Chapter 7 proposes a capsule network (CapsNet) based approach for language identification. The proposed approach of capsule network use convolutional neural network as feature detector. Several capsule layers are designed to effectively select representative frequency bands for each individual language. Experiments showed that the proposed method outperformed the previous state-of-the-art i-vector, BLSTM and merged BLSTM methods.

Chapter 8 summarizes the language identification systems discussed in this thesis. Several problems that have not been addressed, as well as possible solutions that could direct future work, are also discussed.

Chapter 2

Phonetic Overview of Targeted Languages

In Pakistan more than 60 languages are spoken [13], among them Urdu is declared as the national language of Pakistan. According to the 1998 census of Pakistan [1], Punjabi, Balochi, Saraiki, Urdu, Pashto and Sindhi languages have more population as compared to the other languages [1] (see Figure 2.2). The classification of these languages is shown in the Figure 2.1. As Urdu is the lingua franca and national language of Pakistan, it is widely utilized in education, media, government institutions; resulting in Urdu as second language of the community.

There is a paucity of linguistic work on these languages in order to determine how much these languages are acoustically similar or dissimilar to each other. In order to understand the similarities and differences across these targeted languages, comprehensive details of these languages, phonetic inventory and phonological variations among them are explained.

[1] http://www.pbs.gov.pk/sites/default/files//tables/POPULATION%20BY%20MOTHER%20TONGUE.pdf

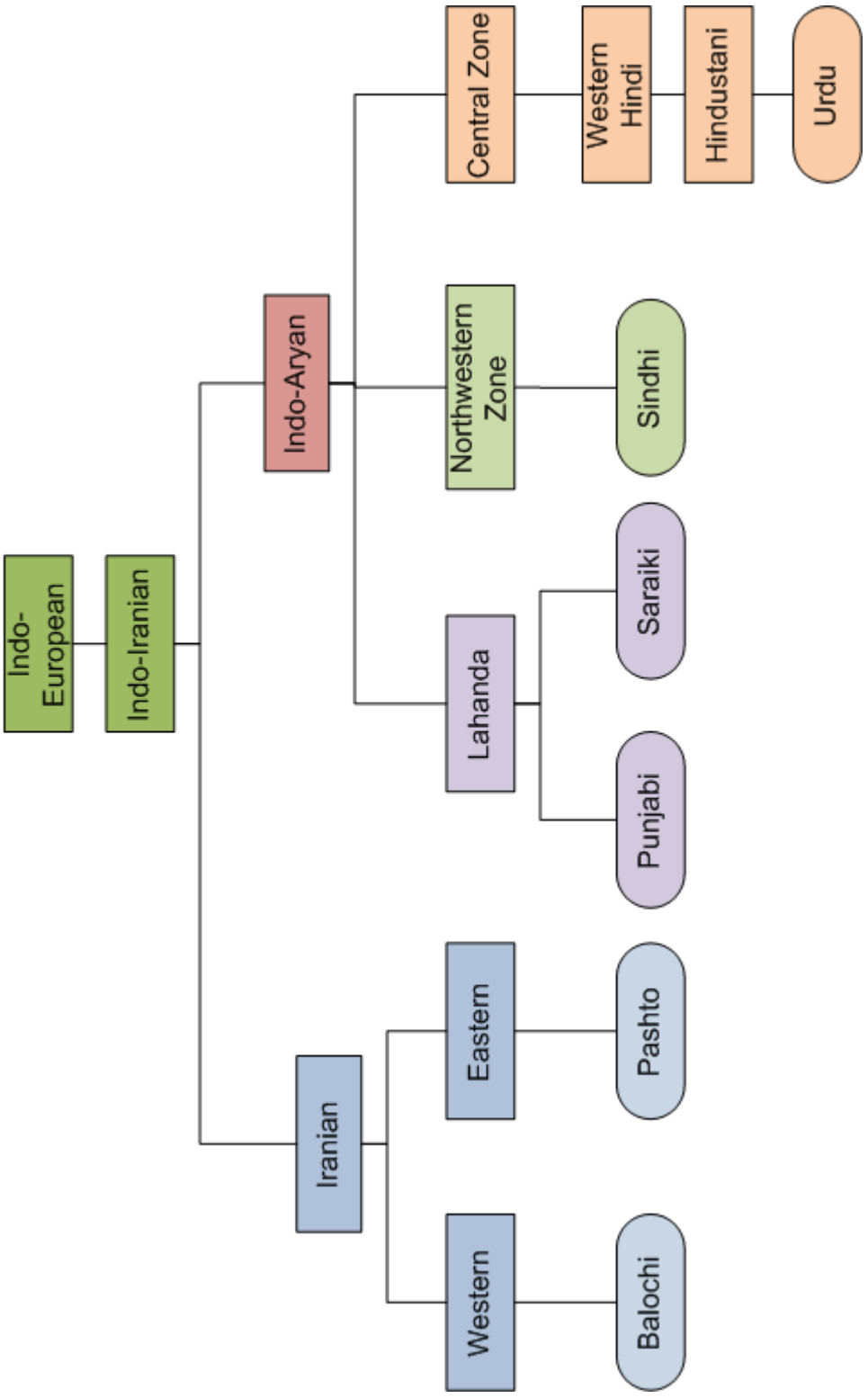


FIGURE 2.1: Pakistan languages and their classification [113]

2.1 Balochi

Balochi belongs to Northwestern Iranian language family with three to five million speakers [103], mostly residing in Pakistan, Iran and Afghanistan. Balochi is provincial language of Balochistan, Pakistan; written in Arabic script using Nastaliq style. Balochi dialects are classified into three categories [114]: (1) Eastern Balochi dialects, (2) Western Balochi dialects and (3) Southern Balochi dialects. Eastern Balochi dialects are influenced by Pashto and Sindhi and primarily spoken in India and Pakistan. Western Balochi dialects are influenced by Persian and primarily spoken in Afghanistan, Iran and Pakistan. Southern Balochi dialects are influenced by Arabic language and primarily spoken in Iran, Pakistan and United Arab Emirates. According to Elfenbein [24], there are six major dialects of Balochi (1) Rakhshani, (2) Kechi, (3) Coastal/ Mekrani, (4) Sarawani, (5) Lashari and (6) Eastern Hill. In Pakistan Kechi, Rakhshani, Coastal dialects and the eastern hill dialects are spoken, whereas, Lashari and Sarawani accents are prominent in Iran. Details of where Balochi dialects are primarily spoken are provided in Table 2.1.

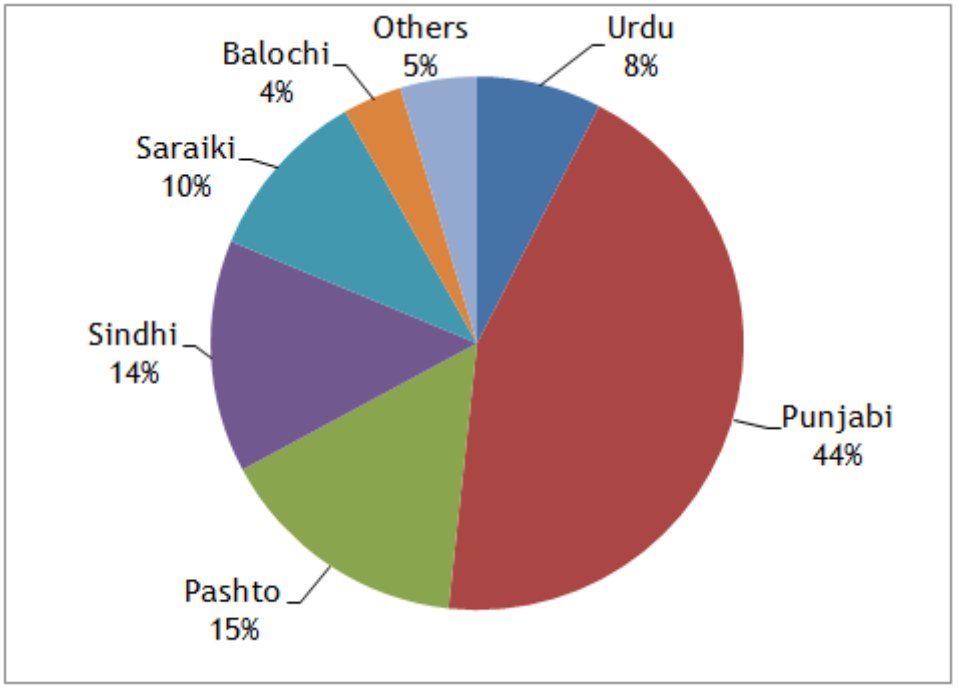


FIGURE 2.2: Pakistan population by mother tongue [1]

TABLE 2.1: Balochi dialects

Dialect	**Where primarily spoken**
Raksani	Makran, Lasbela, Quetta, Chagai and Nushki
Kechi	Kech
Coastal dialects	Gwadar and Karachi
Eastern Hill Balochi	East Quetta, Bugti tribes, North Jacobabad, Upper Sindh frontier, areas in between Dera Ghazi khan, Dera Ismail Khan and Sibi

Balochi consonantal inventory common in all balochi dialects [23] is shown in the Figure 2.3. Lenition of consonants feature of Eastern Balochi dialects distinguish them from the Southern and Western Balochi dialects. Consonants shift at word initial, postconsonantal position and postvocalic position is mentioned in Table 2.2.

	Bilabial	**Dental**	**Alveolar**	**Retroflex**	**Palato-Alveolar**	**Palatal**	**Velar**	**Glottal**
Stops	p b	t̪ d̪		ʈ ɖ			k g	ʔ
Affricates					t͡ʃ d͡ʒ			
Fricatives	f	s z			ʃ ʒ			h
Nasals	m		n					
Lateral	w					j		
Flaps		r	l	ɽ				

FIGURE 2.3: Consonant set common in all Balochi dialects

TABLE 2.2: Eastern Balochi Consonant Shift

Southern and Western Balochi	**Eastern Balochi**	
	word initial and post-consonantal position	**postvocalic position**
p, t̪, k	aspiration: p^h, $\underset{\frown}{t}^h$, k^h	fricatives: f, x
b, d, g	no change	β, ɣ
t͡ʃ , d͡ʒ	aspiration: $t\widehat{ʃ}^h$, no change in d͡ʒ	ʃ, ʒ
w	aspiration: w^h	no change

Phonologically Balochi has five long and three short vowels [114], high and low vowels have short and long contrast, while the middle vowels don't have short

version. Balochi vowel's inventory [114] is shown in Figure 2.4. The vowels /i/, /a/, /u/ can be short or long and vowels /e/ and /o/ are always long.

	Front	Central	Back
High	ɨ , ɨ:		u , u:
	e:		o:
Low		a, a:	

FIGURE 2.4: Balochi vowels

2.2 Pashto

Pashto is member of the East Indo-Iranian languages mainly spoken in Afghanistan, Pakistan and Iran. It is written in Perso-Arabic script. Pashto dialects are mainly divided into three categories: (1) Northern Pashto (spoken in Pakistan), (2) Southern Pashto (mainly spoken in Afghanistan) and (3) Central Pashto (spoken in Pakistan). Pashto is the provincial language of Khyber Pakhtunkhwa, Pakistan. Major dialects of Pashto in Pakistan are Yusufzai, Kandahari and Northern Pashto. Details of where Pashto dialects are primarily spoken are given in following Table 2.3.

TABLE 2.3: Pashto dialects

Dialect	**Where primarily spoken**
North-Eastern (Yousafzai) Pashto	Mardan, Peshawar, Sawat, Charsada, Swabi and Dir
South-Eastern Pashto	Quetta
Waziri	North Waziristan
South-Western (Kandahari) Pashto	Some areas in Balochistan, Khattak and Bannu
Central Pashto	Kabul Afghanistan and South Waziristan

Pashto consonantal phonetic inventory [33] [42] is shown in the Figure 2.5

In literature various vowel inventories are proposed for Pashto language, related to the particular Pashto dialect. According to the acoustic analysis of Yusuafzai

	Bilabial	Dental	Alveolar	Retroflex	Palatal	Velar	Glottal
Stop	p b	t̪ d̪		ʈ ɖ		k g	
Affricate				ʧ ʤ			
Fricatives	f	s z			ʃ ʒ	x ɣ	h
Nasals	m	n		ɳ			
Trill			r				
Flaps				ɽ			
Approximate	w	l			j		

FIGURE 2.5: Pashto consonants

accent, Ijaz [42] reported a total of ten vowels i.e. seven short and three long vowels in Pashto language. Whereas, Tegey and Robson [33] reported nine vowels (three long and six short) in southern dialect of Pashto (spoken in Afghanistan). Pashto vowel system based on [42] is shown in the Figure 2.6, given below.

	Front	Central	Back
High	i: , i		u: , u
Mid	e	ə	o
Low		a , a:	

FIGURE 2.6: Pashto vowels

2.3 Punjabi

Punjabi is the most spoken language of Pakistan covering 44.15% of whole population [1]. In all over the world, Punjabi stood at 10th position on the basis of largely spoken language with 92 million speakers[2]. Maximum number of Punjabi speakers are residing in India and Pakistan. In India, Punjabi is written in Gurmukhi whereas, in Pakistan, Arabic script is being used for Punjabi writing and termed as Shahmukhi. In Pakistan, several dialects of Punjabi exists with

[2]https://www.ethnologue.com/statistics/size

phonetic similarity within the dialects. Main dialects of Punjabi are (1) Doabi, (2) Lahndi, (3) Majhi, (4) Malwai, (5) Multani and (6) Pothohari [73]. Detail of some of Punjabi dialects with reference to location is provided in Table 2.4.

TABLE 2.4: Punjabi dialects

Dialect	**Where primarily spoken**
Majhi	Lahore, Shekhupura, Gujranwala, Gujrat and Kasur
Malwai	Bahawalnagar and Vehari
Doabi	Toba Tek Singh and Faisalabad
Shahpuri	Sargodga, Khushab, Mianwali and Bhakkar
Changvi (Jhangochi)	Khanewal and Jhang
Jandali (Rohi)	Jand Tehsil and Mianwali
Pothohari	North Pakistani Punjab, Azad Kashmir,Rawalpindi, Muree Hills, Jhelum, Muzfarabad
Dhani	Chakwal
Jafri/Khetrani	Musakhel, Barkhan

Punjabi phonetic inventory consists of 32 consonant and 10 vowels. Punjabi consonants [47] depending on place and the manner of articulation are shown in Figure 2.8. Punjabi vowels based on [47] is shown in Figure 2.7.

	Front	**Central**	**Back**
High	i , ɪ		ʊ , u:
	e:	ə	o
	ɛ		ɔ
Low		a	

FIGURE 2.7: Pujabi vowels

	Bilabial	Labiodental	Dental	Alveolar	Retroflex	Palatal	Velar	Glottal
Stops	p b pʰ		t̪ d̪ t̪ʰ		ʈ ɖ ʈʰ	ʧ ʤ ʧʰ	k g kʰ	
Fricatives		f v		s z		ʃ	x ɣ	h
Nasals	m			n	ɳ		ŋ	
Approximates				l	ɭ			
Flaps				r	ɽ	ɹ		

FIGURE 2.8: Punjabi consonants

2.4 Saraiki

Saraiki language belongs to the Indo-Aryan family and is written in Perso-Arabic script. Being North-Western Zone language (see Figure 2.1), Saraiki dialects have similarity with Punjabi and Sindhi [6]. Saraiki is mainly spoken in India, Pakistan and United Kingdom [6]. According to 1998 census, there are 21 million Saraiki speakers residing in Pakistan. Shackle [98] classifies Saraiki into the six dialects: (1) Multani, (2) Riasti (Bahawalpuri), (3) Derawali, (4) Jhangi, (5) Shahpuri and (6) Thali. Regions where Siraiki dialects are primarily spoken are listed in Table 2.5 given below.

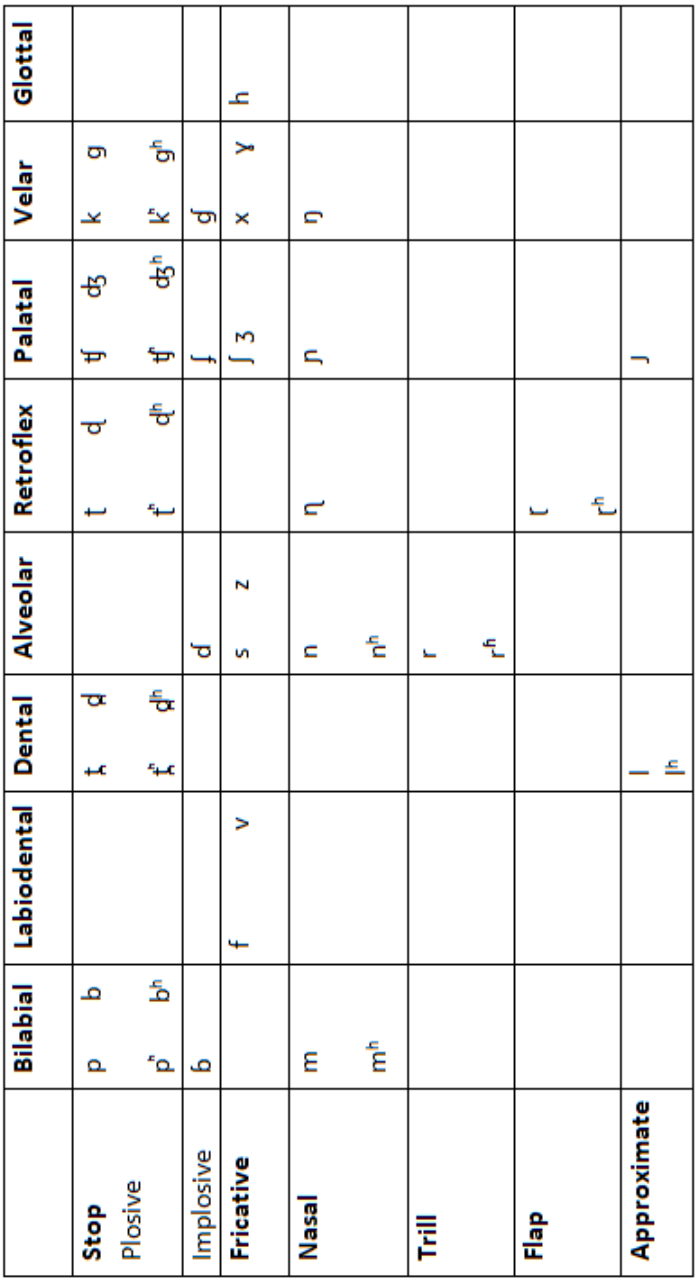
TABLE 2.5: Saraiki dialects

Dialect	**Where primarily spoken**
Derawali	Dera Ghazi Khan, Dera Ismael Khan
Multani	Multan, Muzaffargah, Rahim Yar Khan
Thali	Mianwali, Bannu
Riasti	Bahawalpur
Jhangi	Jhang
Shahpuri	Sargodha

Saraiki phonetic inventory [57] has 42 consonants as shown in the Figure 2.10. Just like Sindhi, Saraiki phonetic inventory contains implosive stops i.e. /ɓ/(voiced bilabial), /ɗ/(voiced alveolar), /ʄ /(voice palatal), /ɠ/(voiced velar). According to Latif [57], Saraiki vowel system comprises six nasal vowels, three short vowels and seven long vowels, whereas, Shackle [98] lists three short and six long vowels in Saraiki vowel system. Saraiki vowels system based on [98] is shown in Figure 2.9, given below.

	Front	**Central**	**Back**
High	i:	ɪ ʊ	u:
	e:	ʌ	o:
Low	æ:	a	a:

FIGURE 2.9: Saraiki vowels



	Bilabial	Labiodental	Dental	Alveolar	Retroflex	Palatal	Velar	Glottal
Stop Plosive	p b pʰ bʰ		ʈ ɖ ʈʰ ɖʰ		ʈ ɖ ʈʰ ɖʰ	ʧ ʤ ʧʰ ʤʰ	k g kʰ gʰ	
Implosive	ɓ			ɗ		ʃ	ɠ	
Fricative		f v		s z		ʃ ʒ	x ɣ	h
Nasal	m mʰ			n nʰ	ɳ	ɲ	ŋ	
Trill				r rʰ				
Flap					ɽ ɽʰ			
Approximate			ɭ ɭʰ			ɻ		

FIGURE 2.10: Saraiki consonants

2.5 Sindhi

According to 1998 census of Pakistan, 14.5% of Pakistani population use Sindhi as first language, with maximum number of speakers residing in Sindh province. Sindhi is written in Arabic script. Major dialects of Sindhi in Pakistan are (1) Vicholi (spoken in Hyderabad), (2) Thareli (spoken in the Thar Desert region), (3) Lasi (in Kohistan and Las Bela), (4) Lari (in the lower Sindh delta and coastal

areas) and (5) Kachchi (in the Rann of Kutch)[48]. Details of where Sindhi dialects are primarily spoken are given in the following Table 2.6.

The phonetic inventory of Sindhi consists of 47 consonants [84] whereas according to Keerio [48] it is composed of 46 consonants. Sindhi consonants are shown in the Figure 2.12, given below. One consonant sound missing in [48] is /v^h/ sound marked with * in Figure 2.12.

The Unique feature of the Sindhi phonetic inventory is the occurrence of the implosive stops i.e. /ɓ/ (voiced bilabial), /ɗ/ (voiced alveolar), /ʄ / (voice palatal), /ɠ/(voiced velar). Sindhi vowel system has ten vowels [84] [48] and their counterparts with nasalization. Sindhi vowel system can be constructed in pairs, in terms of basic length contrast. Oral vowels based on [84] are shown below in Figure 2.11 .

TABLE 2.6: Sindhi dialects

Dialect	**Where primarily spoken**
Vicholi	Central areas of Sindh
Siroli	Upper parts of Sindh, Jacobabad
Lari	Lower Sindh, areas of Hyderabad, Thatta, Badin, Indus Delta
Thareli	Tharparkar
Kutchi	Some areas of lower Sindh and Kuch
Lasi	Lasbela (areas of Balochistan)

	Front	**Central**	**Back**
High	i , ɪ		ʊ , u
	e	ə	o
	ɛ		ɔ
Low			ɑ

FIGURE 2.11: Sindhi vowels

	Bilabial	Labiodental	Dental	Alveolar	Retroflex	Palato-Alveolar	Palatal	Velar	Glottal
Stop Plosive	p b p^h b^h		t̪ d̪ $t̪^h$ $d̪^h$		ʈ ɖ $ʈ^h$ $ɖ^h$			k g k^h g^h	
Implosive	ɓ				ɗ	ʄ		ɠ	
Affricate						ʧ ʤ $ʧ^h$ $ʤ^h$			
Fricatives		f		s z		ʃ		x ɣ	h
Nasals	m m^h			n n^h	ɳ $ɳ^h$		ɲ	ŋ	
Flaps				r	ɽ $ɽ^h$				
Laterals			l l^h						
Glides		ʋ $ʋ^{h*}$					j		

FIGURE 2.12: Sindhi consonants

2.6 Urdu

Urdu langauge is an Indo-Aryan language which is written in Perso-Arabic script using Nastalique writting style. Urdu is also the national language of Pakistan, spoken by more than 163 million people globally[3]. Urdu is one of the top six spoken languages of Pakistan, declared as national language of Pakistan [1]. According to 1998 census, 7.57% of Pakistani population use Urdu as first language [1]. Urdu has four core dialects[4], i.e. Pakistani, Dakhani, Modern Vernacular Urdu (based on the Khariboli dialect of the Delhi region) and Rekhta dialect. Dakhani dialect is being used in Deccan in India and Rekhta is used in Urdu poetry.

Phonetic inventory of Urdu consist of 44 consonants sounds [41], seven long nasal vowels, three short vowels and eight long oral vowels. Urdu consonant chart is shown in the Figure 2.14. List of oral (long and short) and nasal vowels is provided in Figure 2.13

	Front	Central	Back
High	i , ɪ		ʊ , u:
	e	ə	O
			ɔ
Low	æ		ɑ

FIGURE 2.13: Urdu vowels

[3]https://www.ethnologue.com/language/urd
[4]https://en.wikipedia.org/wiki/Urdu#Dialects

	Bilabial	Labio-dental	Dental	Alveolar	Retroflex	Palatal	Velar	Uvular	Glottal
Stop	p b pʱ bʱ	ʈ ɖ ʈʱ ɖʱ			ʈ ɖ ʈʱ ɖʱ	ʧ ʤ ʧʱ ʤʱ	k g kʱ gʱ	q	ʔ
Fricative		f v		s z		ʃ ʒ		x ɣ	h
Nasal	m mʰ		n nʰ						
Approximate				l lʰ		j			
Flap				r rʰ	ɽ ɽʱ				

FIGURE 2.14: Urdu consonants

2.7 Phonological Differences among Targeted Languages

Although these languages have similarities and also differ with each other on many levels such as morphology, orthography, syntax and phonology etc. This section will focus on the phonological differences of the languages. Some of the phonological differences are shown in the Figure 2.15. It summarizes the consonants present in a language (shown in rows) but not present in other languages(shown in column) . For example, phoneme /ɳ/ is present in Pashto but absent in Balochi (see cell 3,2).

Following observations are made;

- Complete set of Balochi's consonants is present in all languages.
- Implosive sounds are present in only Sindhi and Saraiki, while they are absent in Punjabi, Balochi, Pashto and Urdu.
- Aspiration: Voiced and voiceless aspirates are absent in Pashto, while aspirated stops exist in Urdu, Saraiki and Sindhi. As a rule, plosives and affricates are unaspirated in Western and Southern Balochi dialects, But, in Eastern Balochi dialects, voiceless stops are aspirated at word initial position, due to the consonant shift [114].

  Voiced aspirated consonants are absent in Balochi and Punjabi, although this series exists in Urdu, Saraiki and Sindhi. In Punjabi, tone is used instead of voiced aspirated consonants. Aspirated sounds are orthographically present in Punjabi text, but during articulation at word initial they are replaced with voiceless stop and high tone. While, at word medial and word final voicing remains along the tone [47].
- Saraiki and Sindhi have similar sounds except /ɳ^{h}/ is not present in Saraiki, voiced implosive are common in both languages.
- Punjabi is the only tonal language of targeted language set. There are three levels of tone: high-falling, low-rising and neutral.

	Balochi	Pashto	Punjabi	Saraiki	Sindhi	Urdu
Balochi	--	-	-	-	-	-
Pashto	ɳ	--				ɳ
Punjabi	ŋ ɳ	ŋ pʰ ʈʰ tʰ ʧʰ kʰ	--	-	-	ŋ ɳ
Saraiki	bʰ ɖʰ ɖ̱ʰ dʰ ʤʰ gʰ mʰ nʰ rʰ lʰ ɭʰ ɽʰ ɓ ɗ ʃ ɠ ɲ ɳ	pʰ bʰ ʈʰ ɖʰ ɖ̱ʰ dʰ ʧʰ kʰ gʰ mʰ rʰ lʰ ɭʰ ɽʰ ɓ ɗ ʃ ɠ ɲ	ɓ ɗ ʃ ɠ ɲ ɾʱ lʰ ɭʰ	--	-	ɓ ɗ ʃ ɠ ɲ ŋ ŋ
Sindhi	bʰ ɖʰ ɖ̱ʰ dʰ ʤʰ gʰ mʰ nʰ rʰ lʰ ɳʰ ɭʰ ɓ ɗ ʃ ɠ ɲ ɳ	pʰ bʰ ʈʰ ɖʰ ɖ̱ʰ dʰ ʧʰ kʰ gʰ mʰ nʰ rʰ lʰ ɭʰ ɓ ɗ ʃ ɠ	ɓ ɗ ʃ ɠ ɾʱ lʰ ɳʰ ɭʰ	ɳʰ	--	ɓ ɗ ʃ ɠ ɲ ŋ ɳʰ
Urdu	bʰ ɖʰ ɖ̱ʰ dʰ ʤʰ gʰ mʰ nʰ rʰ lʰ ɭʰ	pʰ bʰ ʈʰ ɖʰ ɖ̱ʰ dʰ ʧʰ kʰ gʰ mʰ nʰ rʰ lʰ ɭʰ	bʰ ɖ̱ʰ ʤʰ gʰ rʱ lʰ ɭʰ	-	-	--

FIGURE 2.15: Phonetic inventory comparison

In case of vowels, an acoustic phonetic study[26] has been carried out and corner vowels of these languages (Pashto, Saraiki, Punjabi, Urdu, Balochi and Sindhi) are compared. It is found that /i:/, /æ/, /a:/ and /u:/ vowels exhibit distinctive characteristics. In addition, it is observed that phonetic characteristics of Urdu and Punjabi vowels are similar but are different from Balochi, Pashto, Saraiki and Sindhi.

Gemination is the articulation of a consonant for a comparatively longer period of time than for a single instance of consonant. Gemination phenomenon of targeted languages is discuss as follows:

- In Urdu, in case of aspirated sounds, twinning of non-aspirated followed by aspiration occurs.
- In Punjabi, doubling of consonants occur in medial and final positions [11]. Double consonants are preceded by short vowels. Following Punjabi consonants can occur with length /k/, /g/, /tʃ͡ʃ/, /dʒ͡ʒ/, /ʃ͡ʃ/, /ɕ͡ɕ/, /p/, /b/, /n/, /m/, /l/, /v/ and /s/. Whereas, consonants /n/, /l/, /ɕ̃/, /ɕ͡ʑ/, /h/ and /j/ do not occur as geminates.
- Gemination in Pashto takes place only when consonant appears in the middle of two short vowels [51].
- In Balochi, all consonants except /y/, /h/ and peripheral phonemes may be geminated [114]. Doubling of consonants occur under certain conditions and gemination phenomenon mostly found in loanwords.
- In Saraiki, long consonants at initial position are impossible [98]. All consonants can be geminated, except /h/, /y/, /ɳ/ and /ɽ/. Gemination takes place only after stressed centralized vowels [98].

2.8 Summary

This chapter provides the phonetic overview of six Pakistani languages i.e. Pashto, Saraiki, Punjabi, Balochi, Urdu and Sindhi. In phonetic overview of language,

population size, major dialects and phonemic inventory (both consonantal and vocalic) is presented. Moreover, phonological differences among languages are also discussed.

Chapter 3

Automatic Language Identification

3.1 Introduction

A language can differ from other languages at different levels including acoustic, phonotactic, prosodic, lexical and syntactic levels. These differences can be used as features or cues to discriminate a language from other languages. Spoken level speech information is used to extract acoustic, phonotactic and prosodic features. Whereas, textual level information is required to extract lexical and syntactic features. Various levels of these features are depicted in Figure 3.1.

Acoustic: Acoustic level information shows the physical characteristics of human speech signal [58]. Usually, short term spectral features are extracted to describe the acoustic characteristics of speech signal. In speech processing studies, perceptual linear prediction (PLP), gammatone frequency cepstral coefficients (GFCC) and Mel-frequency cepstral coefficients (MFCC) are widely used acoustic features. MFCC features are widely employed for language identification task [101][94].

Phonotactic: Phonotactic information of a language deals with permissible combinations of phones and their frequency. Though, many languages can have same phones but the statistics of their occurrences and their combination may differ.

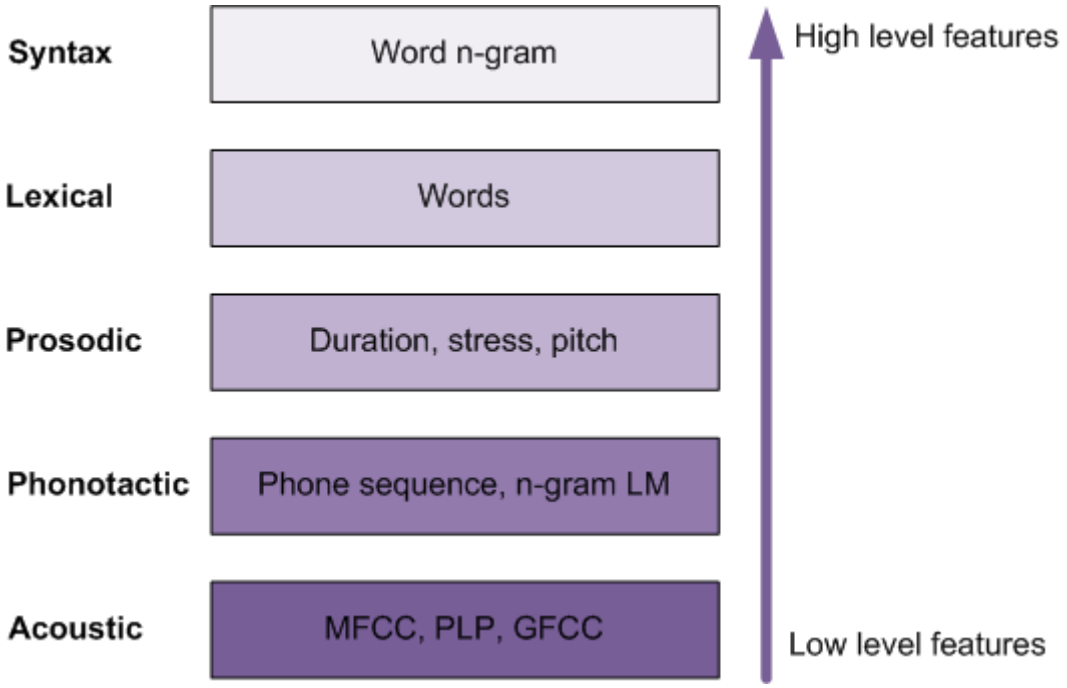


FIGURE 3.1: Levels of language identification features

Phonotactic information of a speech utterance can be extracted using a phone recognizer which changes a speech utterance into a phones sequence. Phonotactic information is extensively utilized for speaker and language recognition [74][59][118].

Prosodic: Prosody of a language is concerned with the properties of larger speech units instead of individual phonemes such as intonation, stress and rhythm. Prosodic features show longer time-span variations across the frames, whereas, acoustic features represent frame level characteristics of a signal.

Lexical: Vocabulary of a language can play a vital role to discriminate a language from other, as each language has a set of unique words.

Syntax: Syntax of a language is concerned with the set of rules for connecting words. Languages may share set of words but context of words can be different.

Lexical and syntactic information from a speech utterance can be extracted using large vocabulary continuous speech recognition (LVCSR). However, development of LVCSR for each language is onerous task, especially for resource scarce languages. This thesis focus on the study of acoustic features and techniques for language identification.

3.2 Language Identification in General

A language identification system is comprised of three main components: (1) signal pre-processing (2) feature extraction, (3) a classifier, as illustrated in Figure 3.2. Training stage is required to make classifier functional before the identification. In training stage, speech signal is pre-processed (non-speech segment removal) and converted into feature vectors sequence $X = x_1, x_2, ...X_n$ where, n is the number of frames. Then, speech characteristics of each language are statically captured and a model λ is created. During the testing/identification, speech utterance is similarly pre-processed and feature vector is extracted. Extracted feature set is then compared to a model set, $\lambda_L(l = 1, 2, , L)$, where, L denotes the number of possible languages. Finally, most likely model is selected using the following

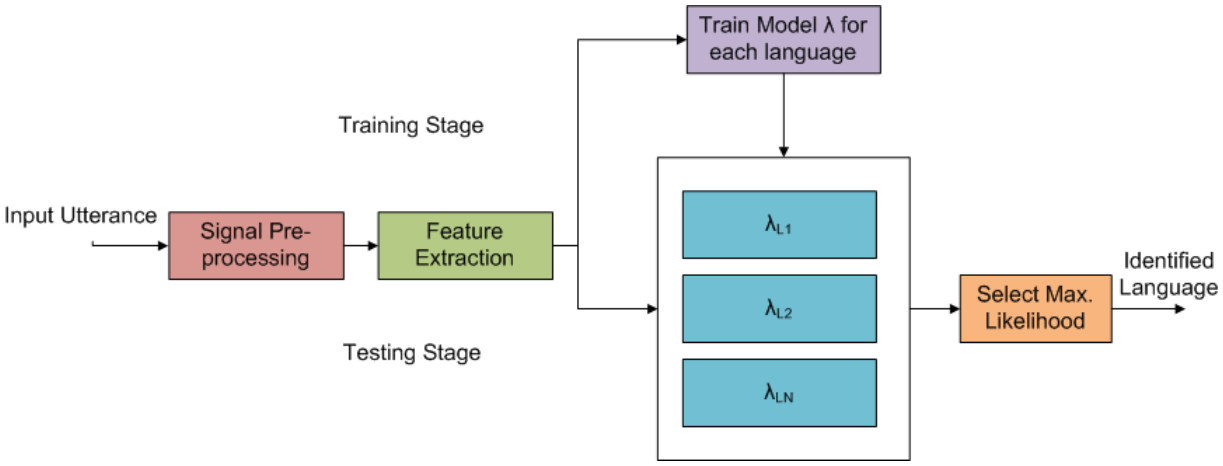


FIGURE 3.2: A general structure of language identification system

equation:

$$L = \underset{1 \leq l \leq L}{\operatorname{argmax}} P(\lambda_l | X) \tag{3.1}$$

According to this general structure of LID, there are two main steps (1) how and what speech information is extracted i.e. features and (2) how this speech information is modeled i.e. classification approach.

3.3 Language Identification Approaches

Language Identification from speech utterances has been conducted over the last 25 years [39]. In the first two decades, language identification (LID) research

proceeded slowly due to the lack of publicly available speech corpora. But, with the growing need on multilingual communication and the start of national institute of standard and technologies (NIST) language recognition (LRE) challenge, much significant progress has been made in recent years. In literature, state-of-the-art language identification systems are based on the exploitation of acoustic [70][21][8], phonotactic [74] and prosodic [72][82] features. Acoustic feature-based approaches explore how a given language sounds; phonotactic feature-based approaches use possible phone combinations of each language to infer the language from a speech utterance; and prosodic feature-based techniques focus on intonation patterns.

In subsequent sections, we present brief overview of phonotactic, prosodic and acoustic features based language identification methods.

3.3.1 Phonotactic Features based LID

In phonotatic features based LID, phone sequences of each language are used to discriminate one language from other. Typical phonotactic language identification system comprised of mainly two building blocks: the phoneme decoder and the n-gram statistical language modeling, as shown in the Figure 3.3. Phoneme decoder is required to convert input speech segment into sequence of phones and statistical language model captures the phones frequencies and also sequences for each particular language. Usually, phoneme recognizer are established on null grammars and hidden markov model (HMM). Development of phoneme decoder/recognizer requires phonetically transcribed speech data, one of the limitation of this approach.

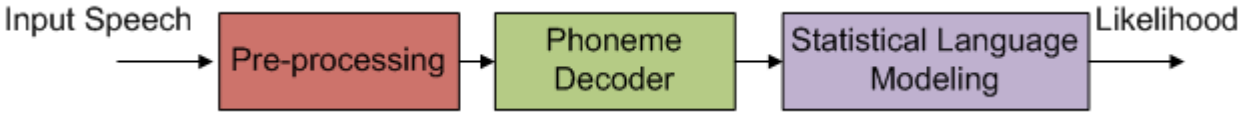


FIGURE 3.3: Scheme of Phonotactic features based LID

Zissman and Singer [121] made very first try to utilize phonotactic information for a LID system. They proposed parallel-phone recognizer followed by language model (PPRLM), in which language specific phone recognizers along with a language model is used to decode the test utterance. For language identification, a language

model is developed using uni-gram and bi-gram counts of phones. PPRLM model is evaluated on OGI-TS dataset [77] and pair-wise comparison between English, Japanese and Spanish is carried out. Average accuracy of 85.8% is achieved for 10-seconds test utterances.

Universal phoneme recognizer are also utilized for LID task by Li et al. [59]. Firstly, universal phoneme recognizer was trained to decode 258 phones of six languages i.e. Hindi, Mandarin, English, Spanish, German and Japanese. Secondly, both vector space model (VSM) and N-gram model are used for langauge modeling, to make pair-wise result. Finally, proposed system was evaluated on 30-sec test utterances of 2003 NIST LRE and 1996 NIST LRE datasets. Sim and Li [100] used the acoustic diversification to improve phonotactic features based LID. Traditional PPRLM systems use language specific phoneme set for phoneme decoding, whereas, proposed method uses multiple acoustic models trained on same phoneme set and same speech corpus, but with different training paradigms and model structure. This resulted into improved acoustic diversification among the parallel sub-systems. Proposed method is evaluated on the 2003 and 2005 NIST LRE data sets and an equal error rate (EER) of 4.71% and 8.61% is achieved.

An alternative to the traditional parallel phoneme recognition (PPR) systema, Jayram et al. proposed a parallel sub-word recognition (PSWR) LID system [45]. The sub-word recognizer (SWR) is developed using automatic segmentation followed by segment clustering. The SWR system does not require sophisticated phonetic labeling in any of the languages, whereas, precise phonetic labeling is required in PPR system. Proposed model is evaluated on 45-sec test utterances of OGI-TS corpus [55]. An overall accuracy of 70% is achieved for identification of six languages.

Instead of language model, Cordoba et al. [17] did ranking with the most frequent n-grams. Distance between each language ranking and input utterance is computed depending on relative positions for each n-gram. Such n-gram positioning resulted in more reliable longer span than traditional PPRLM, 5-gram instead of trigram. This technique resulted in 6% relative improvement over trigram. Navartil [81] proposed an alternative method for phone sequence modeling

by using acoustic pronunciation model and binary tree structure, rather than traditional N-gram models. Wang et al. [112] employed random forest language modeling (RFLM) as the back-end with parallel phoneme recognizer for LID task. Shared back-off smoothing algorithm is applied to solve the sparseness problem. The RFLM obtained 15.7% relative improvement over standard n-gram baseline. Soufifar et al. [102] proposed the idea of i-vector for representation and processing of n-grams. While, preserving the vector's discriminative power, proposed i-vector model represents large vector of the n-gram counts into low-dimensional vector. Proposed model is applied on NIST LRE 2009 dataset, and better results are achieved as compared to baseline method. Wang et al. [111] proposed the idea of phonotactic language branch variability (PLBV) for LID task. The PLBV method concentrates on the discriminative information among languages and language branches. PLBV model is evaluated on 2011 NIST LRE dataset and it is observed that PLBV method outperforms the i-vector system.

Phone level recognition is carried out with the help of clean and phonetically annotated speech data, which is onerous task for resource scarce languages.

3.3.2 Prosodic Features based LID

Prosodic information of language also plays an important role to differentiate one language from other. In prosodic-based LID, properties including pitch contour, intensity and duration are widely used as features. In literature, researchers modeled prosodic characteristics through unsupervised techniques for LID.

Obuchi and Sato [85] used prosodic hidden markov model (HMM) for LID task. Prosodic HMM are trained with un-annotated speech corpus. During feature extraction, power and f0 is estimated for each frame and HMMs are trained. System is evaluated on three languages i.e. English, Japanese and Mandarin and an overall accuracy of 60% is achieved. Prosodic HMMs are combined with phonetic HMMs and combined system outperformed with an accuracy of 85%. Lin and Wang [62] used pitch contour information for language identification. In addition, duration of pitch contour is also considered for language identification. Pitch contour information is formed by using the Gaussian mixture model (GMM). Mary and

Yegnanarayana [72] used intonation, rhythm and stress for language identification. Intonation is represented by change in F_0, distance of F_0 peak with respect to vowel onset point, duration and amplitude tilt. Voiced region duration and syllable is used to represent the rhythm. Prosodic features are modeled through multi-layer feed-forward neural network. Proposed system performance is evaluated on NIST 2003 LRE dataset and EER of 32% is obtained. System performed well for languages that can be discriminated using the rhythm/tonal characteristics such as English vs Tamil, Hindi vs Japanese.

Martnez [71] used duration, energy contour and pitch information for LID task. Prosodic features are modeled using a generative classifier based on i-vectors. System is evaluated by using NIST LRE 2009 dataset. It is observed that fusion with acoustic i-vector based system resulted into significant improvement in performance in the LID system. Language identification system established on prosodic features has been introduced [83]. A total of 87 prosodic features related to F_0, duration and intensity are extracted from syllable-level contour. Extracted features are normalized to reduce the undesirable biasness. Features are modeled by the support vector machine (SVM). Different feature selection algorithms are employed to get optimal subset of features. Model is evaluated on NIST LRE 2007 dataset and EER of 20.18% is achieved. In addition, fusion of prosodic and phonotactic model resulted in accuracy improvement.

Research showed that prosodic information is more effective for tonal languages and fusion of prosodic-phonotatic system improve system performance.

3.3.3 Acoustic Features based LID

Acoustic approaches usually use short term spectral features from un-annotated speech signal to infer language of speech utterance. In acoustic information based LID, different speech parameterization techniques are used to extract acoustic features. Speech parameterization basically collect the most important information from the speech signal while suppressing irrelevant information. Extracted acoustic features vector are modeled using different techniques.

In early years, LID systems were developed using the vector matching and distance measuring techniques [105][28]. Usually, vector quantization and K-means cluster techniques were used to represent feature vector template. Similarity of input feature vector and language template is measured using distance measuring algorithm such as Euclidean and Mahalanobis.

With the successful use of hidden markove model (HMM) and Gaussian mixture model (GMM) in speech processing area, GMM and HMM are also used in the LID task. Zissman [120] and Nakagawa [80] applied GMM and HMM for language identification and decision was being made on the basis of maximum likelihood. Kohler and Kennedy [54] used shifted delta cepstra as feature set along GMM. Proposed model outperforms PPRLM model and higher performance is achieved. GMM-universal background model (GMM-UBM) was used for language identification by Wong and Sridharan [115]. Perceptual linear prediction (PLP) coefficients were collected as feature set, and modeled through GMM-UBM. Vocal tract length normalization (VTLN) is applied to the system to minimize variation effect of speaker. Proposed system is evaluated on OGI-TS database and error rate of 27% is achieved for 10-seconds test utterances. GMM-UBM system output is merged together with PRLM system and system performance is enhanced with 18.4% error rate.

Efforts are being made to minimize the issues of inter-session channel variability and speaker variability. Hubeika et al. [40] proposed the use of Eigenchannel adaptation in model and feature domain for channel compensation in GMM-UBM based LID. Proposed method is used on NIST LRE 2007 database, and average performance cost (C_{avg}) of 14.55% is achieved for 3-seconds test utterances. Dehak et al.[21] used a total variability subspace approach for language identification. MFCC features along their shifted delta cepstral (SDC) coefficients are extracted with the configuration of 7-1-3-7, resulting in 56 dimensional feature vector. Proposed approach is further evaluated on NIST LRE 2009 dataset, and competitive outcomes are achieved with the EER of 2.2%.

Advancements of deep learning-based techniques have revived the application of neural networks for speech processing. Deep neural networks (DNN) yield state

of the art performance in classification tasks. DNNs have been used for language identification [67][65] and evaluated on utterances of duration 3 seconds. DNNs outperformed i-vector framework and a substantial improvement in accuracy is observed. Jiang et al. [5] proposed the idea of deep bottleneck feature (DBF) for language identification. DBF is low-dimensional compact representation of short utterance input speech. Deep neural networks are trained using the extracted DBF features. Effectiveness of proposed features is evaluated on NIST 2009 LRE dataset and EER of 7.01% is achieved for 3-seconds test utterance. Proposed method showed significant improvement over state of art systems for short test utterances.

Recurrent neural networks (RNN) comprised of long short-term memory (LSTM) cells outperformed i-vector in the task of language identification [116] for short utterances. Comparative analysis between the i-vector framework and LSTM has been carried out. The study showed that for 3-second long utterances, LSTM outperformed the i-vector system by up to 26%. In addition, the test utterance duration effect is also analyzed on the limited duration test data (from 0.1 seconds to 2.5 seconds). The system's accuracy deteriorates as the data duration decreases and an overall accuracy of 50% is achieved for 0.5 second long utterances. Usually, a combination of different features or approaches tends to provide better accuracy of the system [92]. Irtza et al. [44] proposed the method to organize languages into cluster on the basis of similarities and disparities among languages. Cluster specific vector representation is obtained using the i-vectors and DNN, while, Gaussian probabilistic linear discriminant analysis (GPLDA) is performed for classification at each node. Proposed hierarchical LID method is evaluated on NIST 2015 dataset and average cost C_avg of 13.3% is achieved.

3.4 Speech Corpora for Language Identification

Normally, LID techniques extract distinguishable information as features of a language from the speech corpora to differentiate one language from other. A number of speech corpora are publicly available for speech processing. Efforts are being made to develop standard benchmark corpora for the evaluation of LID algorithms

and techniques. Comprehensive details of corpora for LID task are provided in subsequent sections.

3.4.1 Oregon Graduate Institute Telephone Speech Corpus (OGI-TS)

OGI-TS is a multilingual speech corpus designed to conduct research for multilingual speech recognition and language identification (LID) [77]. Speech corpus consists of spontaneous responses recorded over telephonic channel. Corpus was recorded mainly in two stages. In the first stage, utterances are recorded from the native speakers of French, Persian, English, Spanish, Tamil, Japanese, Vietnamese, German, Korean and Mandarin Chinese. Data is recorded from 90 native speakers with 3:1 male to female speaker ratio, whereas, speaker ratio varied for each language. In second phase, more data of each language is added, in addition, 200 Hindi calls are also included. Data is transcribed at phonetic level for six languages including German, Mandarin, English, Japanese, Spanish and Hindi.

3.4.2 OGI 22 Speech Corpus

OGI-22 [56] is another multilingual speech corpus designed for language identification. This speech corpus contains speech data from twenty two languages namely: Czech, Farsi, Hindi, Arabic (Eastern), French, Hungarian, Cantonese, English, Swedish, German, Italian, Japanese, Mandarin, Swahili, Korean, Vietnamese, Russian, Malay, Polish, Tamil, Portuguese and Spanish. The corpus consists of fixed vocabulary utterances, recorded from at least 200 speakers per language. Corpus is manually verified by the native speakers. Word level orthographic and phonetic transcription of the corpus is also provided.

3.4.3 CALLFRIEND Speech Corpus

The CALLFRIEND speech corpus contains speech data of twelve languages named as: French, Mandarin Chinese, Arabic, Farsi, Vietnamese, Hindi, German, Japanese, English, Spanish, Korean and Tamil. This speech corpus was released by Linguistic Data Consortium (LDC)[1]. Different dialects of Arabic, English, French and

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

Chinese are considered during data collection. Each language's data consists of 60 telephonic conversations, equally splited into testing sets, training and validation. Transcription of this speech corpus is not available.

3.4.4 KALAKA-3

KALAKA-3 [91] is a speech corpus collected from Youtube audios for language recognition task. The speech corpus is developed to support the Albayzin 2012 language recognition challenge. Database contains TV broadcast speech of ten languages i.e. Basque, Catalan, English, Galician, German, Portuguese, French, Spanish, Greek and Italian. To evaluate the out of set language identification, it contains additional data of 11 more European languages. Audios are automatically downloaded and converted to 16 KHz 16-bit encoded WAV files.

3.5 NIST Language Recognition Evaluation

National Institute of Standard and Technologies (NIST) organizes spoken language recognition/identification evaluation challenge for the development of LID techniques. This evaluation let participant from around the world to compare and evaluate the LID techniques on standard benchmark data and share their results and findings.

First challenge named as 1993 NIST Language Identification Evaluation challenge was focused on the evaluation of LID algorithms using the OGI-TS (10 languages version) database. For LID evaluation 45-seconds and 10-seconds long utterances were used. NIST LRE 1996 challenge was focused on the identification of target language from the input speech utterances within a set of 12 languages. Participants were allowed to use LDC CALLFRIEND and other publicly available datasets for system training. Systems were evaluated using test utterance of 30s, 10s and 3s duration. Test dataset was comprised of 80 speech segments of each category.

In 2003 another challenge was organized named as NIST LRE 2003, focusing on language identification from the set of 12 languages. In addition, Russian language data is utilized for the assessment of the system performance on out of set language.

CALLFRIEND speech corpus was used for evaluation of systems. NIST LRE 2007 challenge was focused on LID of 14 languages. LDC CALLFRIEND speech corpus was used for evaluation.

In NIST LRE 2011 challenge, 24 languages are used as target languages and systems performance was evaluated on different duration of speech i.e. 3-seconds, 10-seconds and 30-seconds. Training and evaluation corpora consist of telephonic and broadcast speech data of 24 languages including seven South Asian languages namely Bengali, Dari, Hindi, Pashto, Punjabi, Tamil and Urdu.

3.6 Summary

In this chapter, general structure of the current language identification (LID)system is discussed in detail. Related LID research with respect to different spoken level features: acoustic, phonotactic and prosodic has been addressed. Moreover, a brief overview of fused LID systems based on several different features is also provided. Furthermore, some of the available multilingual speech corpora have been reviewed.

In the next chapter, we will concentrate on design and development of multilingual speech corpora of Pakistani languages.

Chapter 4

Speech Corpus Design and Collection for Language Identification

In Chapter 3, it is observed that LID techniques extract distinguishable information as features of a language from the speech corpora to differentiate one language from other. A number of speech corpora are publically available for LID technologies [55][104]. Efforts are being made to develop standard benchmark corpora for the evaluation of LID algorithms and techniques. National Institute of Standard and Technologies (NIST) organizes spoken language recognition evaluation challenge for the development of LID techniques and distributes training and testing corpora through Linguistic Data Consortium (LDC) [69]. These corpora consist of telephonic and broadcast speech data of about 24 languages [104], including only seven South Asian languages namely Bengali, Dari, Hindi, Pashto, Punjabi, Tamil, and Urdu. In order to study language identification for Pakistani languages, it is crucial to have publicly available speech corpora of these languages.

For LID from very short utterances, single word utterances speech corpus [89] is used. This speech corpus consists of Pakistan's district names (139 district names) recorded from more than 300 speakers. It is comprised of about 15 hours of speech, sampled at 8 KHz, collected in different background noises, with varying mobile

TABLE 4.1: Dataset-1: Table showing the number of utterance per language used for training, validation and testing. Each utterance being 1-10 seconds long

Language	**Training utterances**	**Validation utterances**	**Testing utterances**
Balochi(bal)	1507	274	274
Pashto(pus)	1507	274	275
Punjabi(pan)	1499	274	275
Saraiki(skr)	1504	274	277
Sindhi(snd)	1505	274	273
Urdu(urd)	1481	274	270

phones and network operators. Selected corpus is comprised of 12,286 speech utterances (from 0.27 sec to 2 sec) recorded from speakers having L1 as Balochi (bal), Pashto (pus), Punjabi (pan), Saraiki (skr), Sindhi (snd) and Urdu (urd), on average each utterance is 0.8 second long. Selected data is divided into three sets i.e. training (train), validation (val) and testing (test). Training set contains 8998 utterances and validation set comprises 1644 speech samples, whereas test data consists of 1644 utterances. Language wise data distribution is described in Table 4.1.

Dataset-1 is very challenging dataset as it consists of very short utterances i.e. proper names recorded by multiple speakers which may or may not belong to their native language. So, there is need of corpus with longer speech utterances and vocabulary. To overcome this corpus barrier, speech corpus for five languages namely Pashto, Punjabi, Saraiki, Sindhi, and Urdu is developed. The dataset is a read speech data collected using a telephone network (mobile and land-line) from different regions of Pakistan. Details of corpus design, speaker selection, corpus recording and annotation and statistics of the dataset are provided in subsequent sections.

4.1 Corpus Design

A complete analysis has been carried out to develop a speech corpus addressing the need to develop a LID system. After analysis, following parameters are defined to collect the corpus.

- Speaking style: continuous read speech

- Text source: books and online newspaper
- Sentence selection: automatic sentence selection from text corpus using greedy algorithm (explained in Section 4.2.2)
- Recording channel: narrow band speech recorded through telephonic channel
- Recording setup: interactive voice response (IVR) dialog system
- Corpus annotation scheme: sentence-level manual annotation using X-SAMPA

In addition, accents of target languages are also considered. Speakers are selected based on the criteria including speaker's language, birthplace, accent, gender, and education level, as per the details given below.

4.1.1 Language Selection

More than 60 languages are spoken in Pakistan [13], according to the 1998 census of Pakistan [1], Balochi, Pashto, Punjabi, Saraiki, Sindhi, and Urdu languages have more population, i.e., 90%, as compared to the other languages. In this study, telephonic speech corpus for five languages, namely Pashto, Punjabi, Saraiki, Sindhi, and Urdu is collected, and LID is carried out between these languages. Acoustic diversity and variation exist in languages across all regions, and usually one accent is considered as standard. In this study, each language datum is recorded from a particular region to cover a standard accent of that language.

Yousafzai accent is used for Pashto data recording, and speakers from Peshawar, Dir and Swat are selected. Punjabi language has six major accents, i.e. Doabi, Lahndi, Majhi, Malwai, Multani, Pothohari, and Powadhi. Majhi is prestige dialect of the Punjabi [31] and spoken in many major cities of Pakistan's Punjab and Indian's Punjab as well. So, Punjabi data are collected from the speakers mostly of Majhi accent. Speakers from Lahore, Sheikhupura, Sailkot, Faislabad, Gujrat, and Okara districts are selected for Punjabi data recording. From nine Sindhi dialects [25], Vicholi is the selected accent. The majority of Sindhi data is collected from the Hyderabad, Sukkhur, Ghotki, and Badin areas. Urdu speech corpus is collected from the Lahore and Sheikhupura speakers. Saraiki speech corpus is recorded from the Multani accent speakers, from Multan district.

4.1.2 Speaker Selection

Speech corpus is recorded from native speakers of each language. It is tried to have equal number of male and female speakers. Literate speakers with a minimum of 12th grade certificate are selected, to ensure a reasonable reading ability. Selected speaker's age ranged from 18 to 50 years. Each speaker is requested to record 20 sentences using telephonic channel. In addition, information related to the speaker's name, mother tongue, and birthplace district is also recorded and documented. Details of speech material and recording sessions are explained in the next sections.

4.2 Corpus Collection

The intent of this work was to develop a multilingual read speech corpus that can be used for language identification task. Addressing the need, the development of the speech corpus has four steps: (1) selection of text corpus, (2) extraction of sentences, (3) recording procedure to develop speech corpus, and (4) verification and annotation of speech corpus. The description of each step is provided in the subsequent sections.

4.2.1 Text Corpus Collection

Text corpus of each language is used for extraction of the phonetically rich sentences. Limited content of Pashto, Punjabi, Sindhi, and Urdu is available online. The majority of the content is available in the form of online news. Therefore, as a first step text corpus of each language is collected from different sources. CLE Urdu digest corpus [109] of one million words is used for Urdu. Sindhi text corpus is collected from online Sindhi books, which is publicly available by Sindhi Adabi Board[1]. Pashto text corpus is collected from online newspapers including BBC Pashto[2], Rohi[3] and Khybernews[4]. Following steps are applied for automatic cleaning of the content collected through crawling:

[1] http://www.sindhiadabiboard.org/
[2] http://www.bbc.com/pashto
[3] http://www.rohi.af/
[4] http://www.khybernews.tv/pashto/index.php

1. HTML tags are removed from the content using regular expressions.
2. Space is inserted between Pashto digits and text e.g. د۲۰۱۷کال
3. Space is inserted at start and end of Latin character sequence to avoid the issue of space omission e.g.لاتینBritanniaاز

The majority of the Punjabi content is also available in the form of online news written in Gurmukhi (used in India) script, whereas limited content in the form of books and online news is available in Shahmukhi script. Punjabi content in Shahmukhi script is extracted from Punjabi portal Wichaar[5] and Punjabi books and magazines such as Punjabi Adab, Sanjh, Swer and Trinjan. Wichaar content is crawled and cleaned by using the steps mentioned above. Saraiki content is not widely available in digital form, as there is only one Saraiki newspaper in Pakistan. Text corpus of Saraiki is collected from different Saraiki books specified below.

- سرائکی ادب دی تاریخ
- سرائکی اجرک
- سرائکی وسیب
- ساري ہتھ ہتھکڑي ہِن
- سرائکی ادب دی تاریخ
- سلسلہ چشتیہ کي بي تاج بادشاہ سئں حضرت خوا جہ غلام فرید
- جامع سرائکی قواعد

The collected text corpora are used for the extraction of phonetically balanced sentences of each language.

[5]http://www.wichaar.com/

4.2.2 Sentence Selection

After cleaning of text corpus of each language, phonetically rich sentences are extracted. Greedy algorithm [32] is used to ensure maximal coverage of triphones in the selected sentences. This algorithm processes text corpus, lists of high-frequent unigram, bigram, trigram, and phonetic lexicon of a language and generates corpus having minimum sentences but giving maximum coverage of triphones. Text corpus of each language has been collected as mentioned in the previous section. N-grams for Urdu are publically available for research [2], whereas N-grams for Pashto, Punjabi, Saraiki and Sindhi are extracted from the collected text corpora. The phonetic lexicon of Urdu [3] transcribed in X-SAMPA is used for the selection of Urdu sentences. This lexicon is comprised of 91,281 high-frequent Urdu words. Pashto phonetic lexicon [76] available in IPA format is converted to X-SAMPA format. Available phonetic lexicon of Pashto is further extended, resulting in a total 22,305 lexicon entries. Development of Sindhi phonetic lexicon is carried out using online Sindhi dictionary [75]. Sindhi word list along with IPA transcription is extracted, and IPA to X-SAMPA conversion is also carried out. Sindhi phonetic lexicon comprises 17,239 words.

The phonetic lexicon of Punjabi in Shahmukhi script is not available. Hence, an automatic system has been developed to build this resource. Punjabi content written in Gurmukhi is selected for this purpose because it has inherent feature of explicit vowels writing that is usually skipped in Shahmukhi script. By using Gurmukhi text, the following procedure is applied and a lexicon of 86,398 words is developed.

1. Gurmukhi text corpus was collected from different websites
2. Unique words list was extracted from the collected corpus and automatic conversion of Gurmukhi to IPA carried
3. IPA transcription is converted to X-SAMPA
4. Finally, Punjabi word list extracted in Step 2 is transliterated to desired Shahmukhi script using automatic conversion

Extraction of phonetic rich sentences become a challenge for resource-scarce languages such as Saraiki, (particularly unavailability of phonetic lexicon). To tackle this problem, an intelligent approach is devised and Saraiki words are automatically transcribed. High frequent uni-grams are extracted from the collected text corpus and automatic transcription of these words is generated by using the letter to X-SAMPA mapping. This word list along transcription is used as Saraiki phonetic lexicon in greedy algorithm.

The greedy algorithm ranks sentences having maximum tri-phone coverage by processing above mentioned resources including text corpus, N-grams and phonetic lexicon of each language. Sentences are selected with a minimum of 10 and maximum of 21 words to handle the effect of too short and too long sentences. A total of 381, 894, 1079, 354 and 845 sentences are selected for recording of Pashto, Punjabi, Saraiki, Sindhi and Urdu, respectively.

4.2.3 Recording Process

The recording process focuses on read speech recording from the native speakers. An interactive voice response (IVR) system is designed for recording of speech data from the speakers over a telephone line. The telephone line is connected to a computer system (CentOS 6.0) using Cisco SPA 3102 VOIP gateway device. A dial-plan is designed and deployed for recording of speech data from speakers, which ensures automatic separate wave file recording for each sentence, unique speaker id and sentence id for file naming.

Printed sentences are given to the speaker and each speaker is requested to rehearse 4-5 times before the recording session. Each speaker is requested to utter 20 sentences in a single phone call. Recorded speech is sampled at 8,000 samples and stored as UNIX WAV file with a bit rate of 16-bit. Recording process flow is shown in Figure 4.1 . Recording is carried out in different environments varied from office, class room and drawing room, etc. Each recording session (from call start till call termination) took about 5 minutes.

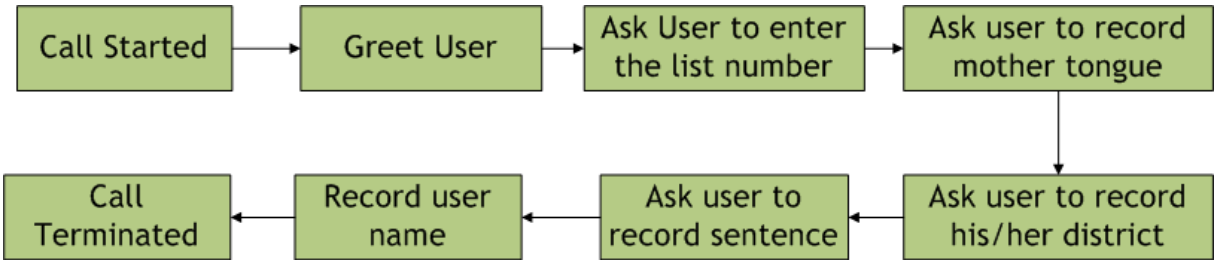


FIGURE 4.1: Call recording flow

4.2.4 Data Annotation and Verification

The recorded speech corpus is manually verified and annotated using PRAAT [12], a speech processing software. Audio files verification process ensured that audio file complied with the following criteria. The audio file is discarded if it does not fulfill the following criteria

- Audio file is not empty
- Signal to noise ratio (SNR) is not less than the defined threshold i.e. 10dB. In case of SNR less than 10dB file will not be processed further
- Any other sound is not mixed in the audio file e.g. voices of the other people, door creak, etc.
- Desired sentence is completely uttered in the audio file
- Desired speech is enclosed by a certain amount of silence, to avoid cutting words

After the verification process, data is phonetically transcribed at sentence level. Each utterance is manually transcribed in X-SAMPA format by carefully listening the sentence and aligned with the spectrogram of the audio file.

After the annotation of data, speaker and language information is added in the file name; to improve the data organization. So, post-processing is performed on the verified and annotated data which involves audio data renaming. An intelligent naming scheme is devised to organize the data. The sample naming scheme is illustrated in Figure 4.2. This naming convention gives the information related

TABLE 4.2: File name scheme description

	Description	In file name
Speaker ID	Unique speaker id assigned to each user on the basis of call order	sp001, sp002, ...
Speaker's District	Speaker district name from which he/she belongs	dt087
Language	Language spoken in utterance i.e. Pashto, Punjabi, Saraiki, Sindhi, Urdu	pus, pnb, skr, snd, urd
Gender	Male or Female	M or F
Sentence ID	Unique sentence Id	0001, 0002, 0003

to language of spoken sentence which is denoted using International code[6] of the language. In addition, sentence id, speaker id, speaker district, and speaker gender are also part of the file name. Detailed description of each part and its usage in file name is provided in Table 4.2.

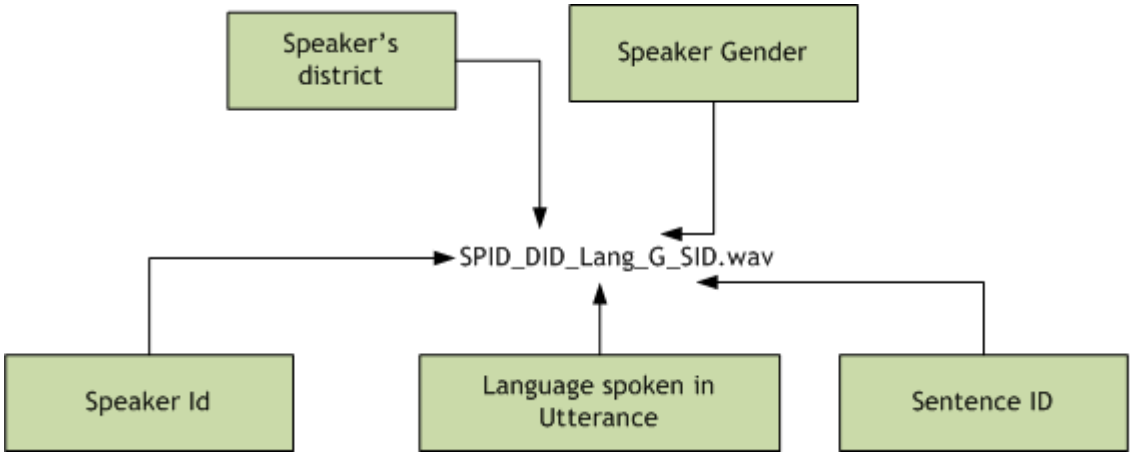


FIGURE 4.2: File naming scheme

4.2.5 Corpus Statistics

Collected corpus consists of read speech of Pashto, Punjabi, Saraiki, Sindhi and Urdu languages. The corpus contains audio data of about 10.43 hours recorded from 316 speakers. Each language contains speech data of 89 minutes on average. Table 4.3 shows the amount of collected speech data and number of speakers against each language.

Dataset-2 statistics related to number of sentences, words, phones coverage is summarized in Table 4.4. It is important to mention that the number of unique

[6]http://www.loc.gov/standards/iso639-2/php/english_list.php

TABLE 4.3: Number of speakers and duration of speech data of the target languages

Language	**Speakers**	**Data duration (Minutes)**
Pashto	55	89
Punjabi	57	126
Saraiki	69	233
Sindhi	71	89
Urdu	64	89

sentences are less than the number of total sentences because some sentences are recorded from more than one speaker. Due to unavailability of Saraiki phonetic lexicon and transcription, phonemic coverage of Saraiki data is not computed. Punjabi speech data has more unique sentences as compared to the other languages. On average each language data is comprised of 3,473 unique words.

Table 4.4: Dataset-2 statistics

	Pashto	**Punjabi**	**Saraiki**	**Sindhi**	**Urdu**
Number of unique sentences	381	894	1079	354	845
Number of total sentences	849	1076	1325	1295	1179
Average duration per utterance (in seconds)	6.3	6.9	5.5	4.0	4.4
Number of unique words	2442	5294	5553	2560	3598
Total words in transcription files	12822	17298	16406	12442	12938
Total Duration(hrs)	1.48	2.06	3.89	1.46	1.46
Number of distinct phonemes	45	66	-	67	65
Number of unique Diphones	1035	2154	-	1563	1956
Number of unique Triphones	5778	14553	-	8294	11293

Table 4.5: Number of utterances and total duration of training and testing data for target languages

	Pashto	**Punjabi**	**Saraiki**	**Sindhi**	**Urdu**	**All**
Number of training utterances	638	501	779	1085	968	3971
Number of testing utterances of 3-seconds	111	111	111	111	111	555
Number of testing utterances of 10-seconds	100	100	100	100	100	500

After the verification and annotation of speech corpus, annotated speech data is then divided into the training (80%) and test data (20 %). The distribution of training and testing data of each language is shown in Table 4.5. The training dataset consisting of, total of 3,971 utterances, 638 utterances spoken in Pashto, 501 in Punjabi, 779 in Saraiki, 1085 in Sindhi and 968 in Urdu.

Test dataset is automatically divided into two parts; (1) 3-seconds and (2) 10-seconds, on the basis of utterance duration. Three seconds test data contains only those utterances which has duration between 1 to 3 seconds whereas utterances of 3 to 10 seconds are selected for 10-seconds data set.

4.3 Summary

This chapter discuss the design and collection of speech corpus of Pakistani languages i.e. Balochi, Pashto, Punjabi, Saraiki, Sindhi and Urdu. Speech corpus is collected from native speakers of each language and transcription of each utterance in X-SAMPA format and in orthographic form is also prepared. The developed speech corpus is about 10.43 hours telephonic channel read speech, collected from 316 native speakers differing in gender, age and educational background. The collected database is divided into training and evaluation sets. This corpus minimizes the barrier of data availability for the development of speech processing applications e.g. speaker recognition and speech recognition for these language.

In Chapter 5, collected data is used to investigate the generalizability of existing LID systems on very short and long utterances. Thereafter, in Chapter 6, a hybrid features based LID system is developed that is robust to the length variation of collected data.

Chapter 5

Language Identification using Acoustic Features

The major motivation behind this study is to thoroughly investigate different acoustic features and identification approaches, which has not been well investigated for Pakistani languages. Additionally, performance of selected approaches is validated with number of acoustic features. Optimization for the use of both features and approaches is further leveraged.

5.1 Features

The raw speech signal is complex and may not be suitable as input to the LID system. Acoustic characteristics of speech segments are used as input to the LID systems. In this study, spectrogram-based features i.e. Mel-Frequency cepstral coefficients (MFCC), cochleagram-based features i.e. Gammatone frequency cepstral coefficients (GFCC) and perceptual linear prediction (PLP) are used to represent the acoustic characteristics of speech. The key difference between MFCC and GFCC features is scale; MFCC features are based on Mel-scale, whereas, equivalent rectangular bandwidth (ERB) scale is used during GFCC features extraction. ERB scale has better resolution at low frequencies.

5.1.1 Mel-Frequency Cepstral Coefficients (MFCC)

Mel-frequency cepstral coefficients (MFCC) are among widely used filterbank based features in the area of speech proceesing. The major advantage of mel-scale filtering is that it approximates the human hearing by emphasizing more on the lower frequencies than the higher frequencies. Frequency scale is converted into mel-scale f_m, using following equation [99]:

$$f_m = 2595 \ln(1 + f/700) \tag{5.1}$$

where f is the frequency on linear scale. Triangular filterbanks are applied to compute the energy from periodgram power spectrum of the speech window. Discrete cosine transformation (DCT) is used to calculate the MFCC from the outputs of the filterbanks by using the following equation [18]:

$$MFCC_j = \sqrt{\frac{2}{N}} \sum_{k=1}^{L} m_k \cos(\frac{\pi j}{N}(j - 0.5)) \tag{5.2}$$

Where L is the number of filter banks and m_k is the log of the k^{th} filter bank's output amplitude.

5.1.2 Gammatone Frequency Cepstral Coefficients (GFCC)

The Gammatone frequency cepstral coefficients (GFCC) features are auditory features based on a set of Gammatone filters which imitate the frequency response of human ears. Gammatone filter bank outputs frequency-time representation of a signal which is called a cochleagram. This cochleagram representation is required for computation of GFCC features. Gammatone filter $g(t)$ [88] can be computed using the Equation 5.3.

$$g(t) = at^{n-1}e^{-2bt}cos(2\pi f_c + \varphi) \tag{5.3}$$

Where φ is the phase that is mostly set to zero, n represents the order of filter, a denotes value of amplitude, f_c is the central frequency (in KHz) and b is the

bandwidth or rate of decay. Patterson and Moore [86] showed that gammatone function of order four n is excellent fit to the human auditory filter shape, so usually n is set to be equal or less than 4. The factor b is defined as [27]:

$$b = 1.019 * 24.7 * (\frac{(4.37 f_c)}{1000} + 1) \tag{5.4}$$

In experiments, we used filter of order (n) 4 and calculated the Gammatone filter with 64 channels. For the GFCC feature vector, the first 10 channels of Gammatone filters which correspond to frequency range less than 200 Hz are excluded.

5.1.3 Perceptual Linear Prediction (PLP) Coefficients

Perceptual linear prediction (PLP) features [35] are inspired from the findings of psychophysics of hearing. PLP features are extracted using the following steps:

1. Speech signal is segmented using the Hamming window and fast Fourier transformation (FFT) is applied on speech signal to convert it into frequency domain. Power spectrum of windowed speech signal is calculated using the following equation

$$P(\omega) = Re[S(\omega)]^2 + Im[S(\omega)]^2 \tag{5.5}$$

   Where Re and Im are the real and imaginary components of short-term speech signal, respectively.

2. Frequency of power spectrum $P(\omega)$ is converted into Bark-scale Ω using:

$$\Omega(\omega) = 6 \ln\{\omega/1200\pi + [(\omega/1200\pi)^2 + 1]^{0.5}\} \tag{5.6}$$

   Where ω is the angular frequency in radian(s).

3. Wrapped power spectrum $\Omega(\omega)$ is convolved with power spectrum of critical-band masking curve $\Psi(\Omega)$, where $\Psi(\Omega)$ is defined as:

$$\Psi(\Omega) = \begin{cases} 0 & \text{for } \Omega < -1.3, \\ 10^{2.5(\Omega+0.5)} & for -1.3 \leq \Omega \leq -0.5, \\ 1 & \text{for } -0.5 < \Omega < 0.5 , \\ 10^{1.0(\Omega-0.5)} & for 0.5 \leq \Omega \leq 2.5 \\ 0 & \text{for } \Omega > 2.5 \end{cases} \tag{5.7}$$

Convolution of $\Psi(\Omega)$ with $P(\omega)$ produce critical-band power spectrum using following equation:

$$\theta(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} P(\Omega - \Omega_i)\Psi(\Omega) \tag{5.8}$$

This $\theta(\Omega_i)$ is further down sampled at approximately 1-Bark interval.

4. Sensitivity of hearing i.e. Equal-loudness is simulated using the pre-emphasis of sampled $\Theta[\Omega(\omega)]$ using

$$\Xi[\Omega(\omega)] = E(\omega)\Theta[\Omega(\omega)] \tag{5.9}$$

Where function $E(\omega)$ approximates the sensitivity of hearing at different frequencies.

5. Spectral amplitude variation of the critical-band power spectrum is reduced by approximating the power law of hearing using Equation 5.10.

$$\Phi(\Omega) = \Xi(\Omega)^0.33 \tag{5.10}$$

This law of hearing simulates the relationship between sound's intensity and perceived loudness.

6. Linear prediction (LP) is applied to the power spectrum signal

7. Inverse Fourier transformation is applied to obtain PLP coefficients from the predictor coefficients

Visual comparison of speech waveform, MFCC, GFCC and PLP features is shown in Figure 5.1. Computation steps of these features are shown in Figure 5.2, to show a comparative scheme of MFCC, GFCC and PLP computation.

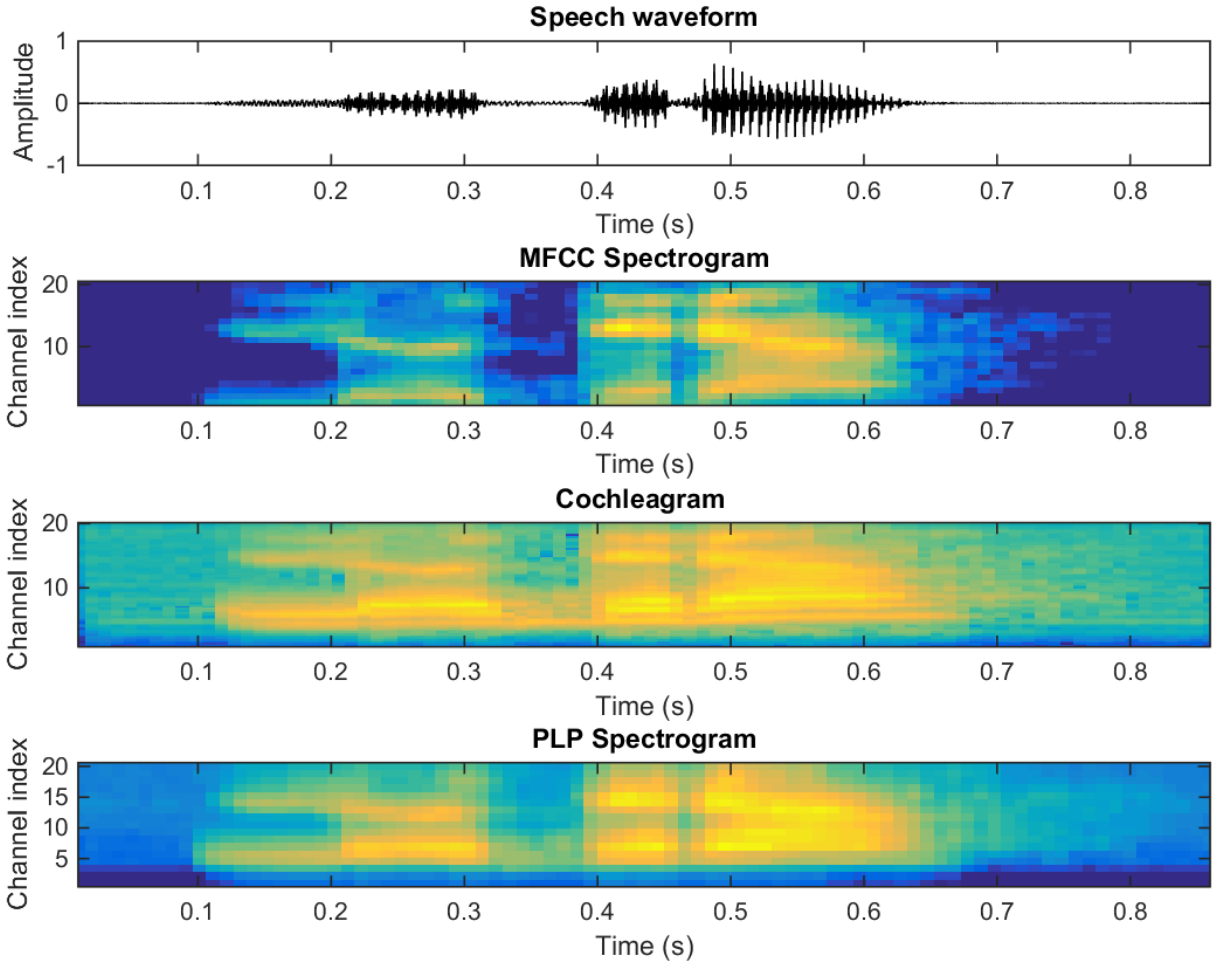


FIGURE 5.1: Visual comparison of MFCC, GFCC and PLP features

5.2 Frameworks

For LID, two well-known and widely used approaches are considered: (1) Gaussian mixtures model with universal background model i.e. GMM-UBM; and (2) i-vector approach. Details of these approaches are discussed in subsequent sections.

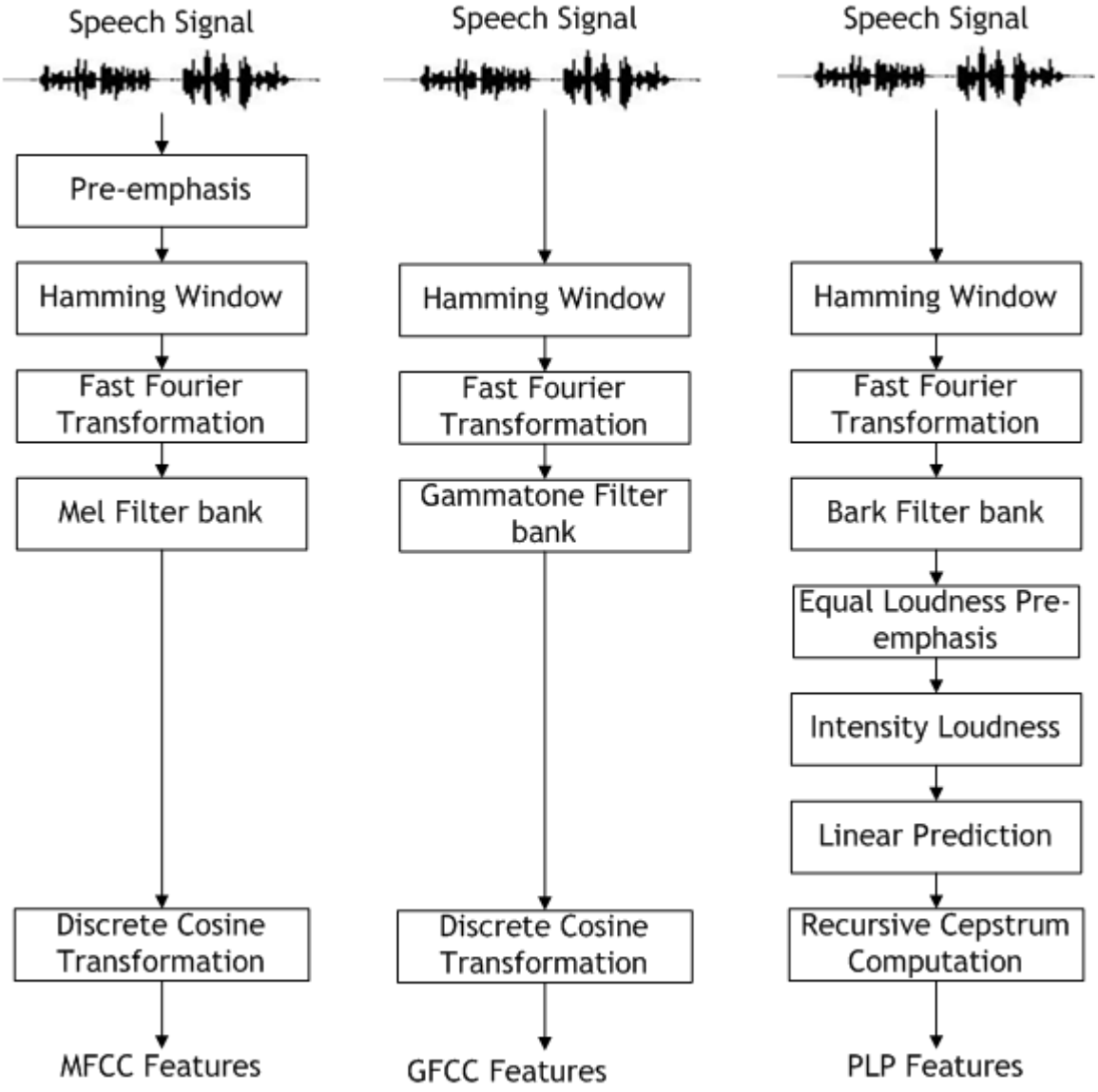


FIGURE 5.2: The computation steps of MFCC (left), GFCC(center) and PLP (right)

5.2.1 GMM-UBM

Gaussian mixture model (GMM)-universal background model (UBM) is an enhanced version of traditional Gaussian mixture modeling technique. Training of GMM-UBM is carried out in two steps: First, frame-level acoustic features from combined data of all languages are extracted. These features are utilized in training of universal background model (UBM). Moreover, expectation maximization (EM) algorithm is used for training of the model. Second, separate data of each language is used and language specific GMM is adapted from UBM. More details

of GMM-UBM are available in [90].

During testing phase, acoustic feature of test utterance are used to compute the log likelihood score on the basis of mean difference of the GMM and UBM likelihood ratio.

In this study, initially a baseline GMM-UBM with 1024 mixtures is developed using training data of all languages. Later on set of experiments are performed to optimize the number of mixtures in GMM-UBM.

5.2.2 i-vector

Motivated by the use of joint analysis factor (JFA) [19] as feature extractor for various acoustic applications such as speaker recognition, Dehak et al. [20] devised a new approach for front-end factor analysis, named i-vectors. JFA is based on separate channel and speaker dependent subspace, whereas i-vector is single spaced, referred as the "total variability space". This new space includes both speaker and channel variabilities simultaneously, without making any distinction between channel and speaker effects. i-vector system development is categorized into three phases; (1) i-vector extraction (2) variability compensation and (3) scoring. Each i-vector based system differs in terms of variability compensation used to perform identification.

In i-vector approach, session and speaker dependent GMM super-vector μ is extracted. This super-vector μ [20] is calculated using equation given below.

$$\mu = m + Tw \tag{5.11}$$

Here m is a speaker and session independent UBM super-vector, T is a rectangular low rank total variability matrix which represents the variation across a large collection of training data and w is an independent vector based on normally distributed random vector $N(0, I)$. The w vector is represented by the BaumWelch (BW) statistics N and F for a given utterance u , which are calculated using the UBM. Given K frames and an UBM m with C mixture components defined in any feature space of dimension D; the BW statistics for a given acoustic utterance

u [20] can be obtained by following equations.

$$N_c(u) = \sum_{n=1}^{K} P(c|y_n, m) \tag{5.12}$$

$$F_c(u) = \sum_{n=1}^{K} P(c|y_n, m)(y_n) \tag{5.13}$$

Here $c = 1, .., C$ is the Gaussian mixture index, m_c is the mean mixture of component c and $P(cy_n, m)$ is the posterior probability of mixture component c given the observation y_n at time n. The i-vector w[20] is obtained by using the following equation

$$w = (I + T^t \sum^{-1} N(u)T)^{-1} . T^t \sum^{-1} \widetilde{F}(u) \tag{5.14}$$

$N(u)$ is a $CF \times CF$ dimension diagonal matrix with diagonal blocks $N_c I(c = 1, C)$. $\sum$ is a $CF \times CF$ dimensional diagonal covariance matrix obtained during factor analysis training (see [49]). Whereas, $\widetilde{F}(u)$ is a $CF \times 1$ dimension vector acquired by concatenating all $\widetilde{F}_c$ statistics for a given utterance u. Derivation details of these parameters can be found in [50]

During the enrollment phase, background model (UBM) is trained using the available training data; total variability (TV) matrix i.e. T-matrix is also trained using the same data. BW statistics are computed and i-vector of 400 dimension is extracted. The numbers of Gaussian mixtures are optimized and 1024 Gaussian components are used for final evaluation.

In i-vector based LID system, channel compensation is done through total variability space instead of GMM super vector and low dimensional vectors are extracted. Number of channel compensation techniques exist e.g. within-class covariance normalization (WCCN) [34], probabilistic linear discriminant analysis (PLDA), nuisance attribute projection (NAP) [14] etc. In this study, PLDA technique is used for channel compensation. The motivation for using this technique is that,

TABLE 5.1: Dataset-2 training corpus

Language	**Number of training utterances**	**Number of speakers**
Pashto	638	54
Punjabi	501	53
Saraiki	779	52
Sindhi	1085	69
Urdu	968	58

PLDA reduce the dimensionality of i-vector and attempts to find out new orthogonal axes to maximize variance between different classes. Consine distance scoring specified in [20] is used for the scoring of test utterances.

5.3 Data set and Performance Measure

During the experiments, i-vector and GMM-UBM are trained using the train set of Dataset-1 and Dataset-2. For Dataset-2 Eighty-nine minutes data of each individual language is being used, whereas effectiveness of LID system is measured individually for 3-seconds test utterances for both identification approaches. The distribution of training data of each language in Dataset-2 is shown in the Table 5.1.

The performance metric of percentage of correctly classified speech utterances E_{cc}[8] is used for the evaluation of LID achieved through i-vector and GMM-UBM, which is measured as follows:

$$E_{cc} = \frac{U_{cc}}{U_T} \tag{5.15}$$

Where U_{cc} and U_T represent the number of correctly recognized utterances and the total number of utterances of test data, respectively.

In addition to E_{cc}, Un-weighted average recall (UAR) and equal error rate (EER) are also calculated for the system having maximum E_{cc}.

Un-weighted Average Recall (UAR) = mean(R1, R2, RN) where R1 is recall of class 1 , R2 is recall of class 2 etc.

EER is point where false acceptance rate(FAR) becomes equal to false rejection rate (FRR).

$$FalseAcceptanceRate(FAR) = \frac{TotalFalseAcceptance}{TotalFalseAttempts} \tag{5.16}$$

$$FalseRejectionRate(FRR) = \frac{TotalFalseRejection}{TotalTrueAttempts} \tag{5.17}$$

5.4 Experiment Setup

Before the extraction of features, voice activity detection on the audio signal is performed and non-speech segment(s) from the utterance are removed. Voice activity detection on all of the audio files is performed using low complexity variable frame rate analysis [106], resulting into the filtration of non-speech frames. MFCC and PLP features are computed by block processing the utterance in a sliding window of 20ms with an overlap of 10ms. These features are extracted by using Kaldi toolkit [87], during the computation of these features 20 filter bank channels are used.GFCC features are extracted using 64 channels Gammaton filters.

5.5 Results

All the experiments described in this chapter are conducted on the Dataset-1 and Dataset-2. The experiments results are summarized in next section, along the details of each system performance.

5.5.1 GMM-UBM

Experiments are carried out by varying the number of mixtures of the GMM-UBM and E_{cc}% is calculated on Dataset-1 and Dataset-2. The performance of the GMM-UBM based LID system with various acoustic features and with 64 mixtures, 128 mixtures, 256 mixture, 512 mixtures ,1024 mixtures, 2048 mixtures and 4096 mixtures has been studied. Experiment results on Dataset-1 are summarized in Table 5.2, for different number of mixtures. Whereas,experiment results on Dataset-2 are summarized in Table 5.4. UAR and EER are also calculated for system having maximum E_{cc} for MFCC, GFCC and PLP feature set. UAR and EER on test data of Dataset-1 and Dataset-2 are shown in Table ?? and Table 5.5, respectively.

It is noted that E_{cc}% increase with the increase in number of mixtures from 64 to 2048, with 4096 mixtures E_{cc} drops, same trend reported in [9]. It is clear that 1024 and 2048 mixtures of GMM-UBM perform better for Dataset-1 and Dataset-2, respectively. Feature comparison shows that PLP features better discriminate the data as compared to the others. Experiments results show that number of mixture can impact the performance and should be optimized.

TABLE 5.2: GMM-UBM system E_{cc}(%) on Dataset-1

Mixtures	**MFCC**	**GFCC**	**PLP**
64	53.64	53.64	55.17
128	56.50	55.29	57.90
256	58.15	58.15	59.30
512	59.67	59.67	60.15
1024	59.30	59.30	**61.55**
2048	59.36	59.06	60.82
4096	56.93	57.84	59.12

TABLE 5.3: GMM-UBM system UAR and EER (%) on test data of Dataset-1

Features	**UAR**	**EER**
MFCC	59.83	25.18
GFCC	59.33	25.06
PLP	61.00	23.11

TABLE 5.4: GMM-UBM system E_{cc}(%) on 3s test data of Dataset-2

Mixtures	**MFCC**	**GFCC**	**PLP**
64	63.78	53.69	64.86
128	66.48	55.13	67.02
256	67.56	58.73	68.46
512	69.00	64.14	69.90
1024	70.99	67.56	71.17
2048	72.25	72.07	**72.61**
4096	71.53	71.89	72.43

TABLE 5.5: GMM-UBM system UAR and EER (%) on test data of Dataset-2

Features	**UAR**	**EER**
MFCC	72.20	17.29
GFCC	71.80	16.75
PLP	72.61	16.93

5.5.2 i-vector Results

In this study, different configurations of i-vector are explored in terms of number of Gaussian components and i-vector length and performance of LID system is evaluated in terms of accuracy. Table 5.6 summarizes i-vector based LID system performance in terms of accuracy with different configurations of Gaussian components and i-vector dimensions on Dataset-1.

TABLE 5.6: i-vector system E_{cc}(%) on Dataset-1

I-vector length	**# of Gaussian Mixtures**	**MFCC**	**GFCC**	**PLP**
400	64	56.75	57.96	60.34
400	128	56.81	60.09	60.64
400	256	57.05	59.00	60.82
400	1024	56.44	55.59	60.21
600	128	56.14	60.46	**61.61**
800	128	56.44	60.88	61.25

TABLE 5.7: GMM-UBM system UAR and EER (%) on test data of Dataset-1

Features	**UAR**	**EER**
MFCC	58.33	24.14
GFCC	61.00	22.14
PLP	61.61	22.01

TABLE 5.8: i-vector system E_{cc}(%) on 3-s test data of Dataset-2

I-vector length	**# of Gaussian Mixtures**	**MFCC**	**GFCC**	**PLP**
400	64	77.65	69.54	75.49
400	128	74.41	71.17	79.09
400	256	76.57	69.90	78.37
400	512	74.23	67.38	**79.27**
400	1024	75.31	67.20	75.85
600	128	74.95	71.53	77.11
800	128	76.21	71.35	78.19

It is clear from the Table 5.8 that 400 dimensional i-vector extracted from the posterior super-vector of UBM with 512 Gaussian components outperforms. UAR and EER on test data of Dataset-1 and Dataset-2 are shown in Table 5.7 and Table 5.9, respectively.

TABLE 5.9: GMM-UBM system UAR and EER (%) on test data of Dataset-2

Features	**UAR**	**EER**
MFCC	76.60	14.4
GFCC	71.60	17.65
PLP	79.27	13.58

Several observations can be made from the results presented in the Table 5.2, Table 5.4, Table 5.6 and Table 5.8. First, PLP feature outperforms the other features and achieve the best results, irrespective of LID approach. Unfortunately, using GFCC features with both GMM-UBM and i-vector approach degrades performance. Second, it can be observed from results that the i-vector approach yields better results than GMM-UBM approach for both Dataset-1 and Dataset-2.

TABLE 5.10: Confusion matrix of LID system using i-vector for Dataset-1

	bal	**pan**	**pus**	**skr**	**sin**	**urd**
bal	**172**	16	9	20	22	35
pan	35	**105**	26	20	19	70
pus	15	9	**215**	2	1	33
skr	29	8	1	**210**	11	18
sin	19	30	6	30	**177**	11
urd	42	21	30	32	11	**134**

TABLE 5.11: Confusion matrix of LID system using i-vector for Dataset-2

	pan	**pus**	**skr**	**sin**	**urd**
pan	**77**	3	18	3	10
pus	1	**104**	1	1	4
skr	6	4	**94**	3	4
sin	6	3	6	**91**	5
urd	12	8	5	13	**73**

Confusion matrix of best i-vector configuration against Dataset-1 and Dataset-2 are tabulated in Table 5.10 and Table 5.11, respectively. Where rows and columns correspond to the actual class and assigned class of the test data, respectively. It can be observed from the confusion matrices (See Table 5.10,5.11) that Urdu(urd) language confused with other languages irrespective of the datasets. Dataset-1 accuracy is low as compared to the Dataset-2, it can be due to following reasons:

- Dataset-1 average utterance duration(0.8 sec) is very short as compared to the average duration of Dataset-2

- Dataset-1 is more challenging as compared to the Dataset-2 because it is sharing same vocabulary among all speakers

Based on these factors, it is being observed that confusion between Urdu and Punjabi is more in Dataset-1 as compared to the Dataset-2. This may be due to the common words between Urdu and other languages. In addition, Pashto language identification has higher accuracy as compared to the other languages, irrespective of utterance duration.

Following observations can be made from the confusion matrix of GMM-UBM approach:

- Urdu has lower recognition as compared to the Pashto, Punjabi and Sindhi. More than 24% of the Urdu testing utterances have been misrecognized and confused to the Punjabi compared to only 13% and 6% confusion with Sindhi and Pashto, respectively. This may be due to the Urdu as lingua franca of Pakistan and its words are common among other languages.

- 91% testing data of Pashto language has been correctly classified, while majority of the remaining utterances are being misrecognized as Urdu.

- Punjabi has about 73% recognition rate while majority misclassified utterances are recognized as Urdu. These languages are closely related /acoustically similar to each other [26]. Therefore, there is more confusion between Urdu and Punjabi.

5.6 Summary

This chapter discusses the use of GMM-UBM and i-vector to model the acoustic characteristic of Pakistani languages. A set of different acoustic features i.e. MFCC, GFCC and PLP are employed for language identification. GMM-UBM and i-vector system are trained using the Dataset-1 and Dataset-2. Experiments were performed to check the impact of number of mixtures and i-vector length on LID accuracy. Results showed that PLP features performs better irrespective of

dataset and LID technique. Moreover, it is also observed that i-vector technique has better ability to discriminate between different languages.

Chapter 6

Merge Bidirectional Long Short Term Memory Network (BLSTM) for Language Identification

A resurgence of deep learning-based techniques has revived the use of neural networks for speech processing. Deep neural networks (DNN) yield state-of-the-art performance in classification tasks. DNNs have been used for language identification[66][64] and evaluated on utterances of duration 3 sec. DNNs outperformed i-vector framework and a substantial improvement in accuracy was observed. Recurrent neural networks (RNN) with long short-term memory (LSTM) memory cells outperformed i-vector in the task of language identification[116] for short utterances. Comparative analysis between the i-vector framework and LSTM has been carried out. The study showed that for 3-sec long utterances, LSTM outperformed the i-vector system by up to 26%. In addition, the effect of the test utterance duration is also analyzed on the limited duration test data (from 0.1 seconds to 2.5 seconds). The system's accuracy deteriorates as the data duration decreases and an overall accuracy of 50% is achieved for 0.5 second long utterances. Usually, a combination of different features or approaches tends to provide better accuracy of the system [92]. In a study by [46], in addition to DNNs, RNNs

are also explored for accent identification and a fusion of DNNs and RNNs is experimented with. Fusion of networks is evaluated using the NLSC corpus on test data with 45-second utterances. It is observed that a combination of networks performed better as compared to individual networks.

Motivated by the outstanding performance of LSTM-RNN in related fields, in this study we use the Bidirectional LSTM (BLSTM) model for LID task of very short utterances (0.27s to 2s). Performance of LID using two different types of acoustic features i.e. spectrogram and cochleagram is investigated. The training procedure of LSTM networks, particularly BLSTM, takes more time and tuning than feed-forward networks. In this study, many aspects of BLSTM training are explored, such as number of hidden layers, size of hidden layers and regularization methods. Moreover, spectrogram feature-based BLSTM-RNN and cochleagram feature-based models are merged together to utilize the strengths of both features.

6.1 Features

Acoustic characteristics of speech segments are used as input to the LID systems. In this study, spectrogram-based features i.e. Mel frequency cepstral coefficients (MFCC) and cochleagram-based features i.e. Gammatone frequency cepstral coefficients (GFCC) are used to represent the acoustic characteristics of speech. These features are computed along with the shifted delta cepstral (SDC) coefficients. They are an extension of delta-cepstral coefficients i.e. stacked version of delta coefficients over several frames. SDCs are used to enhance the accuracy of speaker recognition, language recognition and native language detection systems [97][108][96].

SDC features are typically written as $N - d - P - k$ where:

- N: number of cepstral coefficients in each frame
- d: time advance and delay for delta computation
- P: time shift between consecutive frames
- K: number of frames to be concatenated to form the final vector

For a given N-dimensional cepstral feature vector i.e. $c_0, c_1, c_2, ...C_{N-1}$ at a given time t, we obtain [15] $c(t,i)$ by using Equation 6.1

$$c(t,i) = c(t+iP+d) - c(t+iP-d) \tag{6.1}$$

SDC are stacked version of delta coefficients over several frames i.e. K, as shown in Equation 6.1

$$SDC(t) = [c(t,0)^t c(t,1)^t ....c(t,K-1)^t]^t \tag{6.2}$$

The detailed shifted delta computation is reported by Campbell et al.[15]. MFCC features are extracted with a hamming window of 20ms with 10ms frame shift, filtered through a Mel-scale filter bank. Seven MFCC features are used and SDC parameters are computed with the configuration of 7-1-3-7 and concatenated with static coefficients, resulting in a 56-dimensional input vector.

For the GFCC feature vector, the first 10 channels of Gammatone filters which correspond to frequency range less than 200 Hz are excluded. Seven GFCC features are used and SDC parameters with configuration of 7-1-3-7 are computed which results in a 56-dimensional vector of GFCC-SDC.

6.2 Bidirectional LSTM RNN Model

Long short term memory (LSTM) network [38] is a special type of recurrent neural network (RNN) with the capability of learning long-term dependencies. Each LSTM cell has inputs, outputs and a system of gating units to control the information flow. Internal state unit c^t is the key component of the cell which is regulated by the multiplicative units called gates i.e. input gate i^t, output gate o^t and forget gate f^t. A block diagram of an LSTM cell is shown in Figure 6.1.

Equations of LSTM inputs, outputs, state unit and gates are provided in Equation 6.4, Equation 6.7, Equation 6.6 and Equation 6.5, respectively, more details of which can be found in [30].

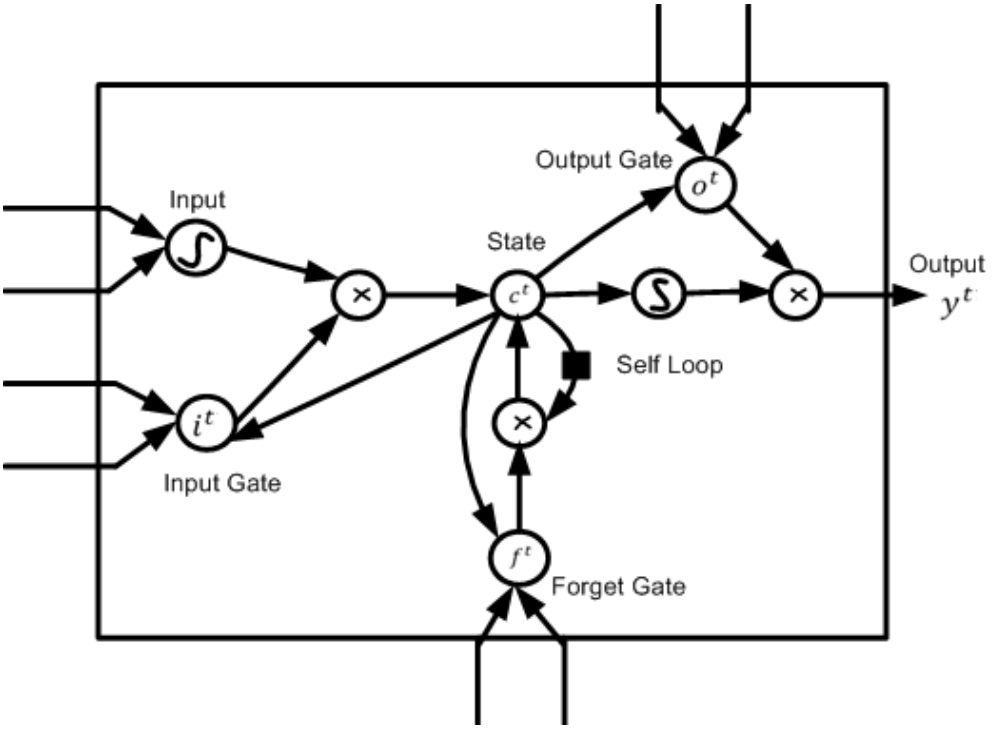


FIGURE 6.1: Schematic diagram of a long short-term memory cell (LSTM)

$$Z^t = tanh(W_z x^t + R_z y^{t-1} + b_z) \tag{6.3}$$

$$i^t = \sigma(W_i x^t + R_i y^{t-1} + P_i \odot c^{t-1} + b_i) \tag{6.4}$$

$$f^t = \sigma(W_f x^t + R_f y^{t-1} + P_f \odot c^{t-1} + b_f) \tag{6.5}$$

$$c^t = i_t \odot z^t + f^t \odot c^{t-1} \tag{6.6}$$

$$o^t = \sigma(W_o x^t + R_o y^{t-1} + P_o \odot c^t + b_o) \tag{6.7}$$

$$y^t = o_t \odot tanh(c^t) \tag{6.8}$$

Where σ is the logistic sigmoid function,i^t, o^t, f^t and c^t are the input, output, forget gate and cell internal state vectors, respectively. Here all b are bias vectors, the P are peephole weight vectors and R are recurrent weight matrices. x^t is the input vector of size 56 (MFCC-SDC or GFCC-SDC) at time t, W denotes input weight matrices, *tanh* denotes the activation function and $\odot$ represents element-wise product of the vectors. In this study, a multi-layer bidirectional LSTM recurrent neural network is being used in order to utilize previous and

future context of a speech frame. Each layer of the bidirectional LSTM network consists of two separate hidden layers i.e. forward layer and backward layer. The first one takes input sequences as-is and the second one the reversed copy of the sequence. Output values from these separate layers are concatenated to generate the output y^t as illustrated in Figure 6.2.

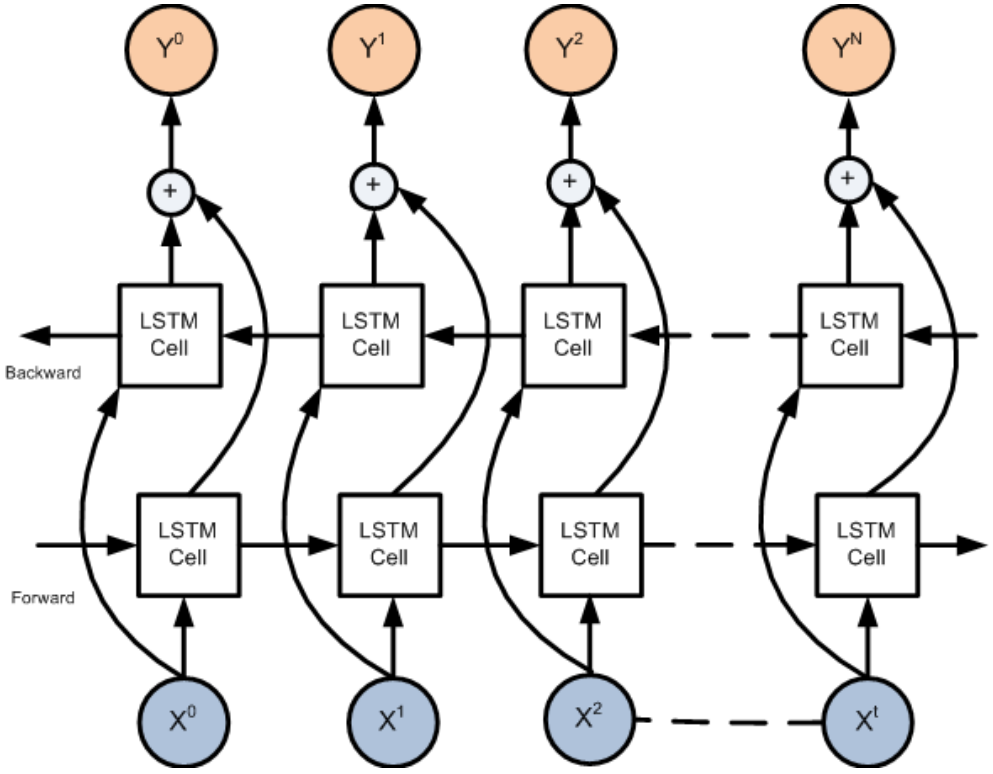


FIGURE 6.2: General architecture of bidirectional LSTM RNN

Multiple layers of LSTM RNN are stacked on top of each other resulting in a deep network architecture. Output sequence of one layer is used as the input sequence of the next layer, ensuring that the next layer receives input from both backward and forward layers of the level below, as illustrated in Figure 6.3.

Output layer of the network is configured as Softmax, to map input x_j to a class probability P_j defined as

$$P_j = exp(x_j) / \sum_i exp(x_i) \tag{6.9}$$

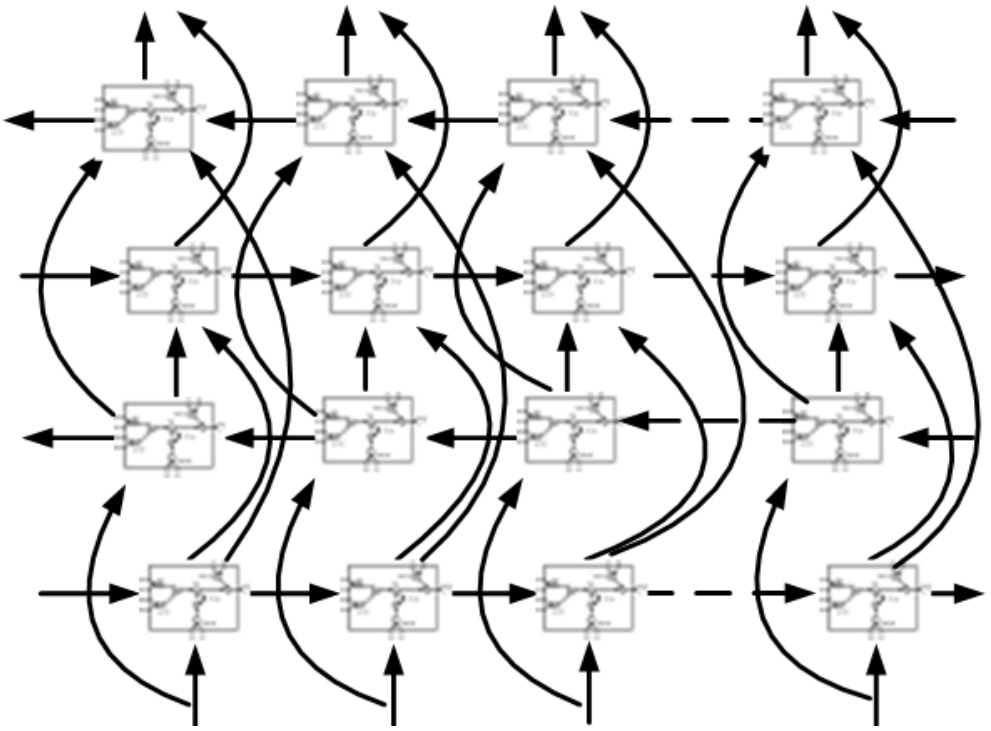


FIGURE 6.3: General architecture of deep bidirectional LSTM RNN

where i is an index over all the classes. Output layer of size six is used (one for each language). cross entropy (CE) function is used as a cost function for back-propagating gradients in the training stage, defined as

$$C = -\sum_{j} t_j \log P_j \tag{6.10}$$

Where t_j denotes the target probability of the class j against the current utterance. Several experiments are conducted to optimize different aspects of BLSTM model in terms of number of layers, hidden units and regularization methods. Details of these experiments are provided in subsequent sections.

6.2.1 Number of Hidden Layers

Theoretically, an adequately wide neural network with only one hidden layer can approximate any function when trained on a sufficient amount of data [22]. But, an extensively wide network may end up memorizing the corresponding output value which is not useful for practical applications because every input value may not be part of the training data. Multiple hidden layers are better because they can

learn hierarchical, more complex internal representations. Therefore, a number of experiments are conducted to find out the optimal depth i.e. number of hidden layers. During these experiments, hidden units are fixed to 256 for each forward and backward direction LSTM. Dropout of 0.2 and L2 input weight regularization of 0.01 is applied at each layer.

6.2.2 Size of Hidden Layers

The number of hidden units in neural network can influence its performance significantly. Fewer hidden units can cause under-fitting result in high errors on training and validation data. On the other hand, a large number of units may cause over-fitting and can result in higher testing error. Optimal numbers of units are required to minimize the effect of under-fitting and over-fitting. Different studies have been carried out to find the various rules for the determination of optimal number of units of the different layers of a neural network [53]. In most of the cases, researchers optimized the neural network with different configuration in terms of number of hidden units and layers. Nagendra and Kher [78] suggest to obtain the best number by iteratively adjusting the number of neurons while considering the error during neural network testing. We used iterative adjustment of number of hidden units, started from the lowest number i.e. 128 and gradually increased upwards in power of two up to 600 due to memory limitations.

6.2.3 Regularization Methods

In deep learning, regularization methods are helpful to decrease model complexity by minimizing weights values, since small values result in smoother hypothesis functions. In this study, two regularization methods are used: dropout [37] and weight regularization i.e. standard L2 regularization. L2 regularization is applied to input connection of each LSTM unit at each layer of the network. Different dropout and L2 regularization combinations are further investigated using grid search to find the optimal configuration. Dropout values are considered in the range of 0.0-0.5 with increment of 0.1, whereas, L2 regularization values in the range of 0.00-0.02 with increment of 0.01 are used in experiments. Different combinations are tried and model performance is evaluated on validation data.

6.3 Merged BLSTM RNN Model

Wide variety of acoustic features is available with different strengths and weaknesses. A combination of several different features as an input to the single system may result in more accurate results. Feature combination can be done in several ways at several levels. A study by[61] combined MFCC and perceptual linear prediction (PLP) features with HMM-GMM ASR system, resulting in significant reduction of word error rate (WER). Similarly, a study by [107] showed that combination of MFCC and GFCC features using convolution neural networks (CNN) provides significant advantages over a CNN with one feature set. Zhang et al. [117] used a combination of different features including MFCC, PLP, linear frequency cepstral coefficients (LFCC), GFCC for language recognition and observed that feature combination results in lower equal error rate (EER).

In this study, instead of low level feature combination, two BLSTM models trained on MFCC-SDC and GFCC-SDC are combined. The two best model architectures from the Section 3 experiments are used and merged together. These two independent BLSTMs models return their final output sequence, thus dropping the temporal dimension (i.e. converting the input sequence into a single vector). These two vectors are concatenated and forwarded to fully connected layers, to learn mappings from both high level feature vectors to the output classes, as shown in Figure 6.4. The last layer of this network is a softmax layer which outputs probabilities for each class.

Similar to BLSTM models, a numbers of experiments are carried out to find the optimal number of fully connected layers and size of layers for merged BLSTM RNN model.

6.4 Experimental Setup

Each BLSTM-RNN network is optimized using Adam optimizer [52], a widely used optimization method with an initial learning rate of 10^3. Several experiments are carried out to find a best architecture in terms of number of layers, size of hidden layers and regularization methods. For all GFCC feature-based experiments, 60

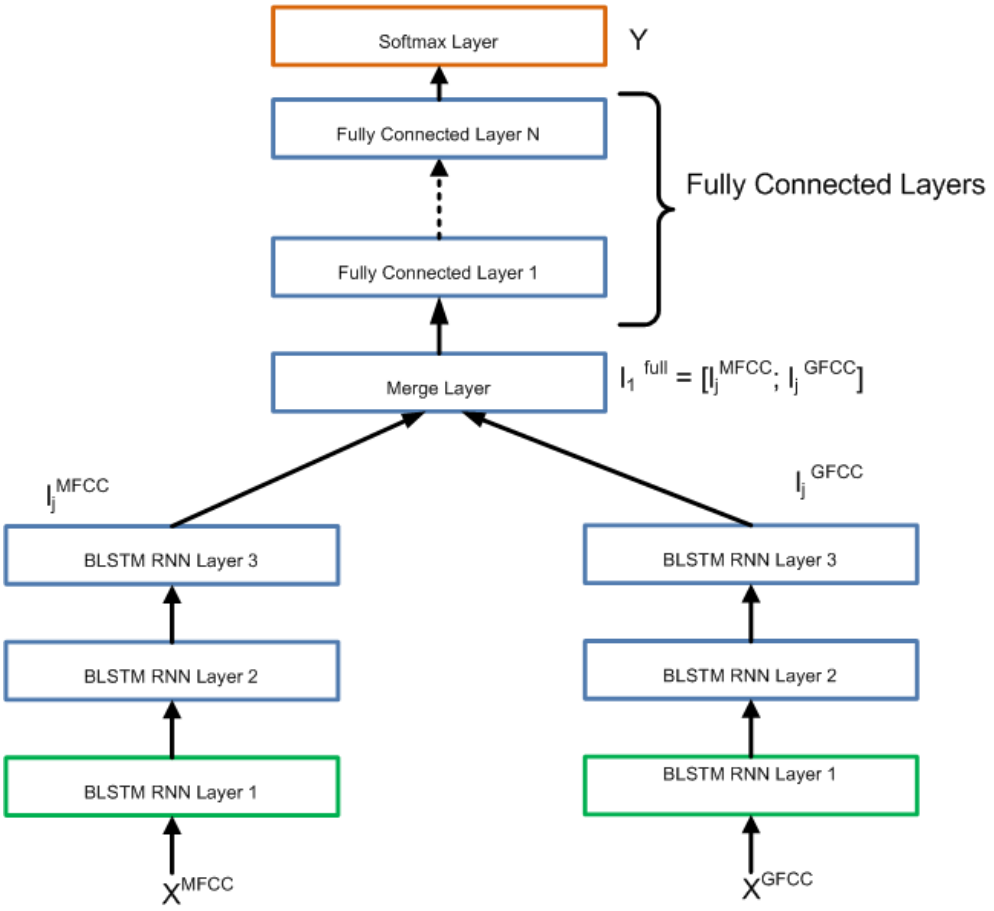


FIGURE 6.4: Architecture of merged BLSTM RNN models

epochs are used for training of the network, whereas, for MFCC feature-based experiments 30 epochs are used. Each network configuration is evaluated using the validation set and the model with the best validation accuracy is used for later experiments. In merged BLSTM models approach, rectified linear unit (ReLU) [79] is used for activation of fully connected layers and dropout of 0.4 is applied at each layer.

BLSTM-RNN models are implemented in Python, using the Keras [16] neural network library, running on top of Tensorflow. Experiments are conducted on a machine with NVIDIA GeForce GTX 1080 GPU with 8GB of memory and 32GB RAM.

BLSTM networks are trained using the train set of dataset-1. The languages include Balochi, Pashto, Punjabi, Saraiki, Sindhi and Urdu. val set of the dataset-1 is used for the parameter optimization. Optimized network is evaluated using

the test set.

6.5 Results and Discussions

A number of experiments are carried out for BLSTM-RNN model using MFCC and GFCC as input features and model accuracy on validation data is calculated. In addition, network architecture search is also performed for merged BLSTM models. Details of experiments are given in subsequent sections.

6.5.1 BLSTM RNN Network Architecture Search

6.5.1.1 Number of Hidden Layers

Results of experiments for number of hidden layers are shown in Table 6.1. During experiments, the effect of number of hidden layers in terms of training accuracy (train Acc), validation accuracy (val Acc), training loss (train loss) and validation loss (val loss) is also examined. Validation loss shows the value obtained from same epoch as the validation accuracy, whereas training loss is obtained from the last epoch.

It is observed that training and validation accuracy of MFCC feature-based BLSTM network increases with the use of more than one hidden layers and highest accuracy is achieved with two hidden layers. Highest validation accuracy of GFCC feature-based BLSTM network is achieved with four hidden layers. Therefore, for later experiments two hidden layers will be used for MFCC feature-based BLSTM and four hidden layers will be used for GFCC feature-based BLSTM network.

TABLE 6.1: Effect of number of hidden layers on training and validation accuracy using MFCC and GFCC features, BLSTM network with 256 neurons/layer and dropout of 0.2, L2 of 0.01

		Feature Set							
#Layers	#Params[M]	MFCC				GFCC			
		Val Acc(%)	train Acc(%)	train loss	val loss	val Acc(%)	train Acc(%)	train loss	val loss
1	0.6	52.92	99.33	0.03	1.14	56.02	62.96	0.98	0.63
2	2.2	53.89	98.36	0.05	1.34	60.03	66.12	0.92	1.11
3	3.7	53.89	97.92	0.07	1.36	59.12	70.25	0.80	1.12
4	5.3	52.98	97.80	0.08	1.35	60.40	71.28	0.78	1.10

TABLE 6.2: Validation data accuracy with different number of hidden units, 2 layers are used in MFCC feature-based BLSTM network and 4 layers are used in GFCC feature-based BLSTM network; dropout of 0.2 and L2 of 0.01 is applied at each layer of both models.

# Hidden Units	# Params [M]		val Acc(%)	
	MFCC	GFCC	MFCC	GFCC
128	0.5	1.3	54.20	59.12
256	2.2	5.3	53.89	60.40
512	8.6	21.22	54.01	60.03
600	11.80	29.09	53.35	50.61

6.5.1.2 Size of Hidden Layers

After finding out the number of hidden layers, experiments are carried out to find out the optimal number of neurons in each layer. During experiments equal numbers of neurons are used at each layer. Two layers are used in MFCC feature-based BLSTM and four layers are used in GFCC feature-based BLSTM. Dropout of 0.2 and L2 input weight regularization of 0.01 is applied at each layer. Table 6.2 shows the results of comparison of hidden layer size. It is observed that with larger number of neurons, the number of trainable parameters rises and the validation accuracy start deteriorating. It is observed that two layers BLSTM model with 128 neurons performs better for MFCC features set and BLSTM model of four layers with 256 neurons yields higher validation accuracy for GFCC feature set. This network configuration is further improved by using the dropout and L2 regularization.

6.5.1.3 Regularization Methods

Table 6.3 shows validation accuracy against different configurations of dropout and L2. It is observed that dropout of 0.3 with L2 of 0.01 yields the best result with MFCC as feature vector, whereas for GFCC feature set, dropout of 0.2 with L2 of 0.01 yields best validation accuracy.

The LID accuracy attained by optimized MFCC feature-based BLSTM model on the training and validation data at different training epochs is shown in Figure 6.5. Similarly, optimized GFCC feature-based BLSTM model accuracy on the training and validation data at different training epochs is illustrated in Figure

TABLE 6.3: Validation accuracy (%) with different combinations of dropout and L2, MFCC feature-based BLSTM network with 2 layers and 128 neurons/layer, GFCC feature-based BLSTM network with 4 layers and 256 neurons/layer

Dropout	L2					
	0.00		0.01		0.02	
	MFCC	GFCC	MFCC	GFCC	MFCC	GFCC
0	52.06	61.31	54.07	59.85	53.83	59.48
0.1	53.89	62.04	55.71	60.88	55.53	59.30
0.2	56.44	62.28	57.36	61.00	57.66	60.00
0.3	57.96	**62.53**	**58.45**	61.00	57.54	60.15
0.4	56.93	62.34	57.96	59.12	57.96	60.09
0.5	53.95	59.79	55.23	58.94	55.53	57.42

6.6. A comparison between the accuracy curves suggests that MFCC feature-based network is faster to train and yields highest validation accuracy than GFCC feature-based network. Loss curves of MFCC and GFCC feature-based models are shown in Figure 6.7 and Figure 6.8, respectively. It suggests that there is slightly more over-fitting in MFCC feature-based model than the GFCC feature-based network even with higher dropout and L2 regularization values.

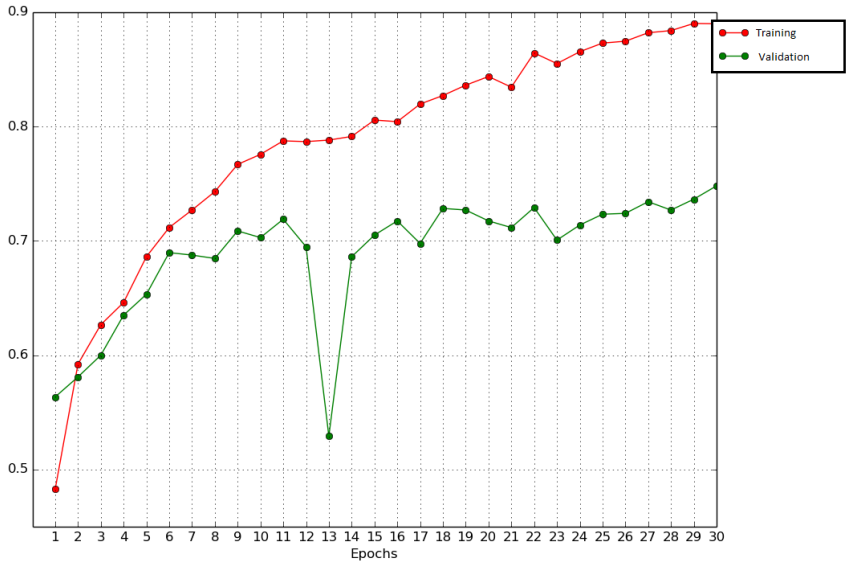


FIGURE 6.5: MFCC feature-based BLSTM-RNN model accuracy on the training and validation data over 30 Epochs, 3 layers BLSTM model with 256 neurons/layer, dropout of 0.4 and L2 of 0.00

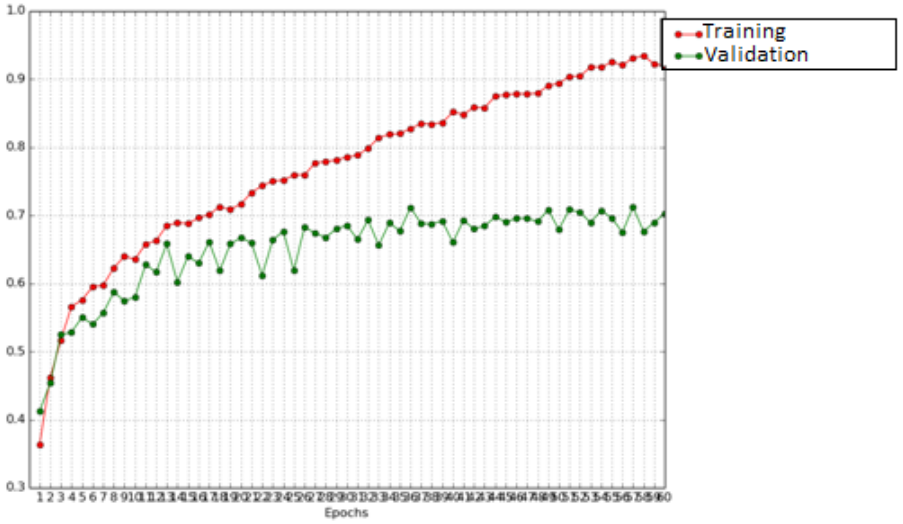


FIGURE 6.6: GFCC feature-based BLSTM-RNN model accuracy on the training and validation data over 60 Epochs, 2 layers BLSTM model with 512 neurons/layer, dropout of 0.2 and L2 of 0.01

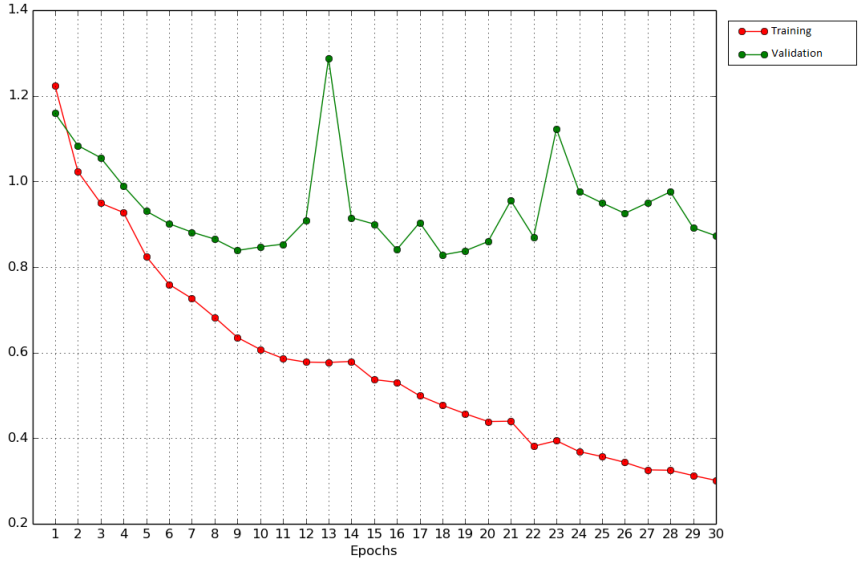


FIGURE 6.7: MFCC feature-based BLSTM model loss over 30 epochs

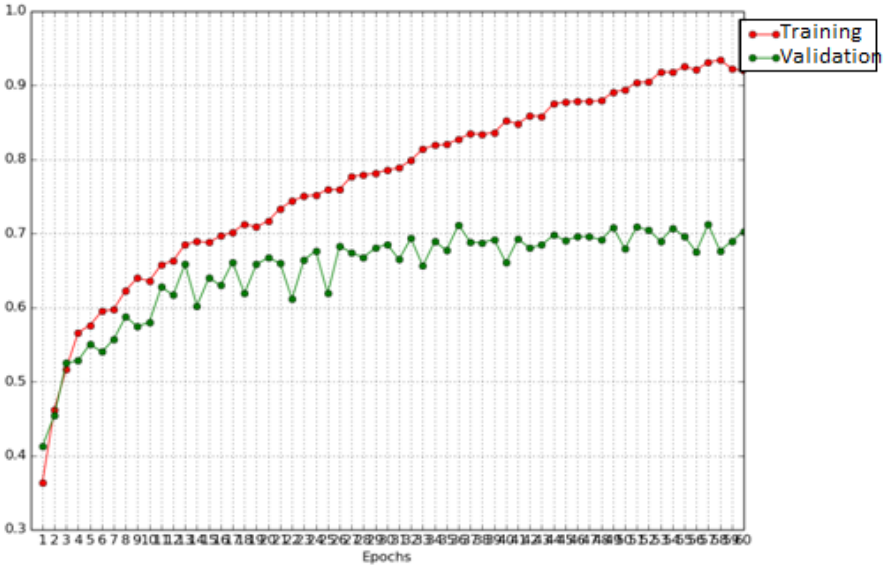


FIGURE 6.8: GFCC feature-based BLSTM model loss over 60 epochs

6.5.2 Merged BLSTM RNN Models Architecture Search

After the network optimization of MFCC feature-based BLSTM model and GFCC feature-based BLSTM model, their outputs are combined. Outputs of BLSTM models trained on MFCC and GFCC features are merged and forwarded to the fully connected (FC) layers as shown in Figure 6.4. Experiments are conducted to find out the optimal number of fully connected layers and size of layers. To find out the optimal number of FC layers, experiments are performed with 256 neurons in each layer and dropout of 0.3 is applied at each layer. It is observed that two FC layers are optimal number of layers.

After finalization of number of fully connected layers, experiments with two FC layers and varying layers size are conducted and validation accuracy is computed. During experiments, equal numbers of neurons are used at each layer. It is observed that merged BLSTM models network with two fully connected layers of size 512 yields the highest validation accuracy.

This model is later evaluated on the test set and an accuracy of 62.59 % is achieved. Moreover, UAR of 62.50% and EER of 18.36% is achieved on the test set. Confusions among languages are shown as a confusion matrix in Table 6.4 and recalls

in Table 6.5. It's evident from Table 6.4 that Pashto (pus) language is identified more correctly than the rest of the languages and confusions are most frequent among Sindhi and Balochi.

TABLE 6.4: Confusion matrix for the test set using the merged BLSTM models, rows are reference and columns are hypothesis

	bal	pan	pus	skr	snd	urd
bal	**231**	7	2	12	15	7
pan	43	**164**	7	10	21	29
pus	29	35	**135**	3	8	64
skr	10	1	0	**258**	4	1
snd	68	20	2	10	**158**	16
urd	30	134	8	8	11	**83**

TABLE 6.5: Recall for each language and the un-weighted average recall (UAR) on the test set (%)

bal	pan	pus	skr	snd	urd	UAR
84	60	49	94	58	30	62.5

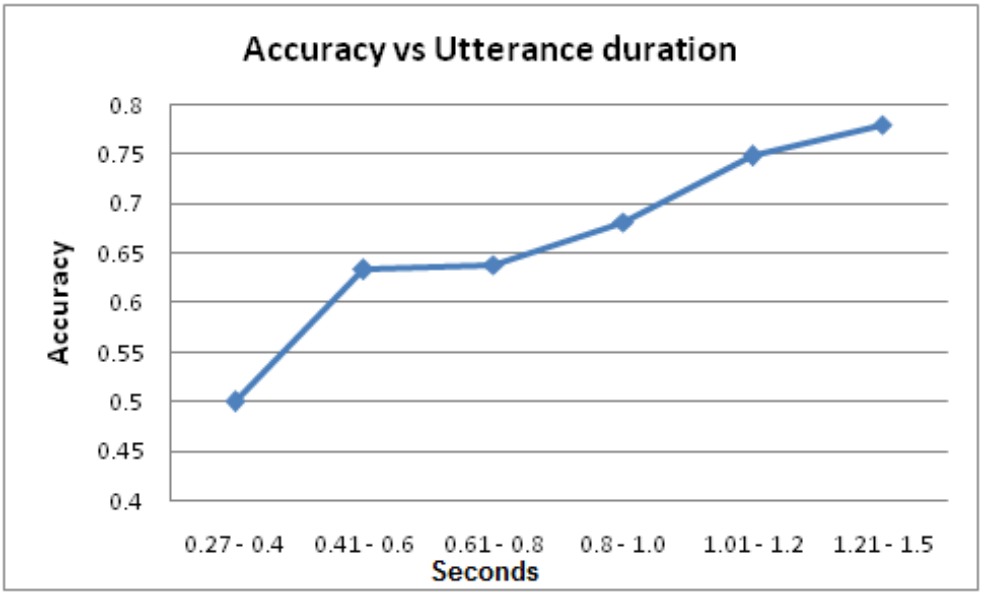


FIGURE 6.9: Merged BLSTM model performance on very short utterances

Figure 6.9 shows the relationship between system accuracy and duration of test utterances. Test data set is divided into six groups on the basis of utterance duration and accuracy of each group is computed. It is evident from Figure 6.9 that system accuracy increases with the duration of test utterances so a reliable system can be developed when longer test utterances are available.

6.6 Summary

This chapter discusses the use of spectrogram and cochleagram features for language identification from very short speech utterances (0.8s on average). Bidirectional long short-term memory (BLSTM) models are adopted to solve this complex problem of LID for limited duration speech data. Several configurations of BLSTM models are explored and compared. This study indicates that MFCC features are more robust than GFCC features for speech data recorded in various acoustical environments, with various quality mobile phones and network operators.

In addition, BLSTM models trained using MFCC and GFCC features are also merged together to take advantage of both feature sets. It is observed that the merged models approach outperforms the individual models. However, by looking at the confusion matrix given in Table 6.4, it is observed that the system confuses among languages which are acoustically and geographically close, such as Punjabi and Pashto and Sindhi and Balochi.

Chapter 7

A Capsule Network Based Approach for Language Identification

A resurgence of deep learning-based techniques has revived the use of neural networks for acoustic modeling. In particular, convolutional neural networks (CNN) and RNN are widely used for audio signals detection and classification [95]. Recurrent neural networks can capture long time dependency and CNN show strength in detecting local features. A variant of RNN i.e. BLSTM is used for language identification in Chapter 6 and it is observed that BLSTM system outperforms the i-vector system. However, training of RNNs based network is computationally more expensive than CNN based network and takes a longer time than CNNs. Therefore, for tasks that have short time dependency such as keyword spotting or phoneme-level recognition, CNN based systems are still taking an active part of research and show competitive performance.

However, convolution neural network are not good in capturing spatial relationships of low level features. To solve this issue usually deep network or method of max-pooling is used. Whereas, max-pooling leads to the loss of valuable information by ignoring the neurons having minimum activation value. In speech processing, spatial relationship between speech features play an important role.

In traditional speech processing, different information e.g. pitch and formant frequencies are extracted from the spectrogram. The spatial information of these features in frequency and time domain decide what spectrogram represents. Existing CNNs based speech processing system were missing the spatial information of these features, which can be overcome by applying the capsule network.

Ispired by this, in this chapter we propose a capsule network based framework for language identification, as shown in Figure 7.1. This is motivated by capsule networks (CapsNet) [36], which have shown promising results in the area of image classification, text classification [119], sound event detection [63] and speech recognition [7]. A set of experiments are conducted while varying factors such as number of convolution layers for feature detection, the kernel size, channel size to find the best capsule network model. Details of architecture and experiments are provided in subsequent sections.

7.1 Capsule Network

Proposed network is a variant of network proposed by Sabour et al. [93]. It consists of two parts: (1) encoder and (2) decoder. Details of these parts are provided in subsequent sections.

7.1.1 Encoder

Encoder part of network takes $X \in \mathbb{R}^{F \times T}$ as input and learns to decode it into a 16 dimensional vector of instantiation parameters. Encoder is comprised of three parts, (1) feature detection: group of convolutional layers, (2) primary capsule: replace the scalar output of feature detector to vector output and (3) language capsule layers: build parent-child relationship.

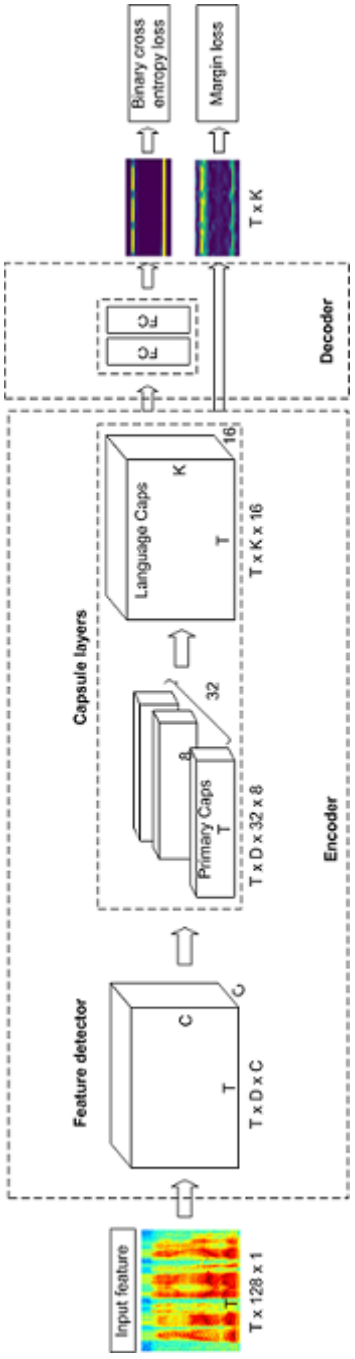


FIGURE 7.1: Proposed approach architecture which comprised of two parts: (1) encoder (2) decoder

7.1.1.1 Feature Detection Layers

Multiple convolutional layers are used to detect local features from the input features. Feature vector $X \in \mathbb{R}^{F \times T}$ with zero padding is fed to the feature selection layers, where F represents number of frequency bins and T is number of frames in an utterance. Output of feature detection stage is vector $H \in \mathbb{R}^{M \times F^1 \times T}$, where M

is number of feature map of last convolution layer, F^1 is the number of frequency bands after pooling. Number of convolution layers, kernel size and number of filters are optimized for each data-set.

7.1.1.2 Primary Capsule (PrimaryCaps) Layer

Primary capsule layer is used to convert scalar output of feature detector's convolution layer to vector-output. Primary capsule layer consists of P convolutional capsule layers with C channels. Each channel is comprised of 8-dimensional capsules. These capsules work as low level capsules, which are forwarded to the LangCaps layer. During experimentation, values of P and C are optimized for each dataset.

7.1.1.3 Language Capsule (LangCaps) Layers

In LangCaps layers, outputs of low level capsules u_i are used to calculate prediction vectors of high level capsules $\hat{u_{j|i}}$, by multiplying with a weight matrix W_{ij}, as shown in Equation 7.1

$$\hat{u_{j|i}} = W_{ij} u_i \tag{7.1}$$

$$W_{ij} = [M \times N] \tag{7.2}$$

Dynamic routing process proposed by Sabour et al. [93] is applied for selection of the these prediction vectors on the basis of similarity between each high level capsule's output and its prediction vectors. Connection weight between prediction vector and high level capsule's depends on the similarity between them, more similarity results in larger connection weight. Contribution of prediction vector to its corresponding high level capsule can be further increased by weight gain.

Let v_j denotes the output of high level capsule and u_i represents the output of low level capsule, then v_j can be computed using Equation 7.3.

$$v_j = \frac{\|s_j\|^2}{1 + \|s_j\|^2} \frac{s_j}{\|s_j\|^2} \tag{7.3}$$

Where s_j represents its total input, which is a weighted sum over all prediction vectors $u_{j|i}$ and can be calculated using Equation 7.4

$$s_j = \sum_i c_{ij} \hat{u_{j|i}} \tag{7.4}$$

Where c_{ij} denotes the coupling coefficients, which are determined by the dynamic routing process [93], which fulfills the idea of assigning parts to wholes. These coefficients represent amount of agreement between low level capsule i and high level capsule j. c_{ij} calculates how likely low level capsule i can activate high level capsule j, strong agreement between the properties of i and j results in higher c_{ij}. Sum of the coupling coefficients between low level capsule i and all the capsules in the layer above i is equal to 1. Coupling coefficients are determined using Equation 7.5.

$$c_{ij} = \frac{exp(b_{ij})}{\sum_k exp(b_{ik})} \tag{7.5}$$

Where b_{ij} denotes the initial logits i.e. log prior probabilities that low level capsule i should couple with high level capsule j. Initially, b_{ij} are set to be 0 and re-computed in each iteration, by the similarity between prediction vector and high level capsule $j$output v_j.

$$b_{ij} \leftarrow b_{ij} + \hat{u_{j|i}}.v_j \tag{7.6}$$

Margin loss of each language capsule k is calculated using Equation 7.7

$$L_k = T_k max(0, m^+ - \|v_k\|)^2 + \lambda(1 - T_k)max(0, \|v_k\| - m^-)^2 \tag{7.7}$$

Where T_k is 1 iff class k exists, m^-, m^+ and λ are hyper-parameters and set to 0.9, 0.1 and 0.5, respectively.

7.1.2 Decoder

Decoder part of network takes a sixteen dimensional vector from the correct LangCaps and learns to decode it into spectrogram of a language. During network training, decoder only utilize the correct LangCaps vector whereas, incorrect vectors are ignored. In network, decoder serve as a regularizer, it takes output vector

of the correct LangCaps as input and learns to recreate an $F \times T$ spectrogram. Euclidean distance between the re-generated spectrogram and input spectrogram is used as loss function. Decoder part force capsules to learn features that are useful for re-generating the original spectrogram.

Two feed-forward fully connected (FC) layers are used in decoder part. Each output of the LangCaps gets weighted and directed into each neuron of the first FC layer as input. First FC layer takes $D \times L$ inputs that are all directed to 512 neurons. Therefor, there are $D \times L \times 512$ trainable parameters, where D is the dimension of LangCaps's vector and L is the number of languages. Output of the first FC layer is fed to the second FC layer having 256 neurons.

7.2 Feature Extraction

Spectrogram extracted from raw signals sampled at 8 kHz are used as input to this network. Spectrogram of speech recordings consisting of 128 frequency bins for each 20ms frame, with a 10ms overlapping Hamming window are extracted, resulting feature matrix $X \in \mathbb{R}^{S \times F \times T}$. Where S is number of samples, F is number of frequency bins and T is number of frames. Extracted feature matrix is fed to the network as input.

7.3 Experiment Setup

The network hyper-parameters optimization was obtained by means of a random search strategy for each dataset. The number and the shape of convolutional layers in feature detector, size of primary capsule, LangCaps dimensions and the maximum number of routing iterations have been varied. Hyperparameters used in the proposed approach for dataset-1 and dataset-2 are provided in Table 7.1 and Table 7.1, respectively. Each capsule network is optimized using Adam optimizer [52], a widely used optimization method with an initial learning rate of 10^3. Rectified linear unit (ReLU) [79] is used for activation of convolution and fully connected (FC) layers. The dropout rate of 0.25 is set to the each FC layer. To minimize the overfitting, batch normalization [43] is also applied. For dynamic capsule routing, three iterations are used.

Table 7.1: Hyperparameters used in network for dataset-1

	Feature detector		**Capsule layers**			**Feed-forward layers**	
	Conv1	**Conv2**	**PrimayCaps**	**LangCaps**	**LangCaps**	**FC1**	**FC2**
kernel	$64 \times 3 \times 3$	$32 \times 3 \times 3$	$16 \times 3 \times 3$	-	-	-	-
stride	1×1	1×1	2×2	-	-	-	-
pooling size	-	2×2	-	-	-	-	-
Dropout	-	0.25	-	-	-	0.25	0.25
activation function	ReLU	ReLU	Squashing	Squashing	Squashing	ReLU	ReLU
number of hidden units	-	-	-	-	-	512	256
Dimension of capsule	-	-	8	16	16	-	-

Table 7.2: Hyperparameters used in network for dataset-2

	Feature detector					**Capsule layers**			**Feed-forward layers**	
	Conv1	**Conv2**	**Conv3**	**Conv4**	**Conv5**	**PrimayCaps**	**LangCaps**	**LangCaps**	**FC1**	**FC2**
kernel	$64 \times 3 \times 3$	$16 \times 3 \times 3$	$16 \times 3 \times 3$	$16 \times 3 \times 3$	$16 \times 3 \times 3$	$32 \times 3 \times 3$	-	-	-	-
stride	1×1	1×1	1×1	1×1	1×1	2×2	-	-	-	-
Dropout	-	-	-	-	-	-	-	-	0.25	0.25
activation function	ReLU	ReLU	ReLU	ReLU	ReLU	Squashing	Squashing	Squashing	ReLU	ReLU
number of hidden units	-	-	-	-	-	-	-	-	512	256
Dimension of capsule	-	-	-	-	-	8	16	16	-	-

Capsule network model is implement in Python, using the Keras [16] neural network library, running on top of Tensorflow. Experiments are conducted on a machine with NVIDIA GeForce GTX 1080 GPU with 8GB of memory and 32GB RAM. Capsule networks are trained and evaluated using the dataset-1 and dataset-2, separately.

7.4 Results and Discussion

Two sets of experiments are conducted: (1) language identification for very short utterances i.e. dataset-1 and (2) language identification for relatively long utterances i.e. 3-seconds using dataset-2.

7.4.1 Dataset-1 Results

The LID accuracy attained by optimized model on the training and validation data at different training epochs is shown in Figure 7.2.

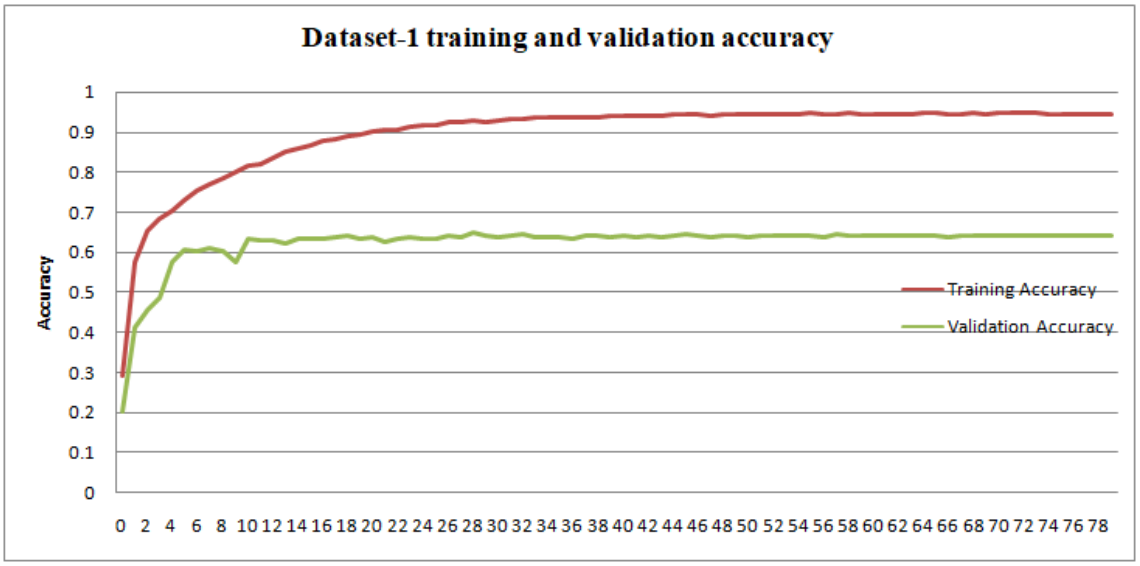


FIGURE 7.2: Capsule network based model accuracy on the training and validation data over 80 Epochs

Results reported in Table 7.3 show both the best performance achieved using capsule network and baseline system results (merged BLSTM model reported in Chapter 6). Results show recall for each language , un-weighted average recall (UAR) and EER using the baseline system and capsule network. It is evident

from Table 7.3 that capsule network based LID system outperforms, with overall gain of 7.42% .

TABLE 7.3: Recall for each language and the un-weighted average recall (UAR) on the test set of dataset-1 using capsule network

	Balochi	Pashto	Pujabi	Saraiki	Sindhi	Urdu	UAR	EER
Baseline system	84	60	49	94	58	30	62.50	18.36
Capsule network	70	83	44	77	88	59	69.92	14.42

Table 7.4 shows both the confusion and discrimination performance of capsule network based LID system considering all language pairs in the form of a confusion matrix.

TABLE 7.4: Confusion matrix dataset-1, ground truth is represented in the Y-axis while the predicted language is represented in the X-axis

	Balochi	**Pashto**	**Punjabi**	**Saraiki**	**Sindhi**	**Urdu**
Balochi	**191**	15	25	8	3	32
Pashto	9	**227**	11	3	2	23
Punjabi	19	36	**120**	15	13	72
Saraiki	19	1	4	**214**	9	30
Sindhi	6	3	8	7	**240**	9
Urdu	33	29	20	19	11	**158**

7.4.2 Dataset-2 Results

The LID accuracy attained by optimized model on the training and validation data of dataset-2 at different training epochs is shown in Figure 7.3.

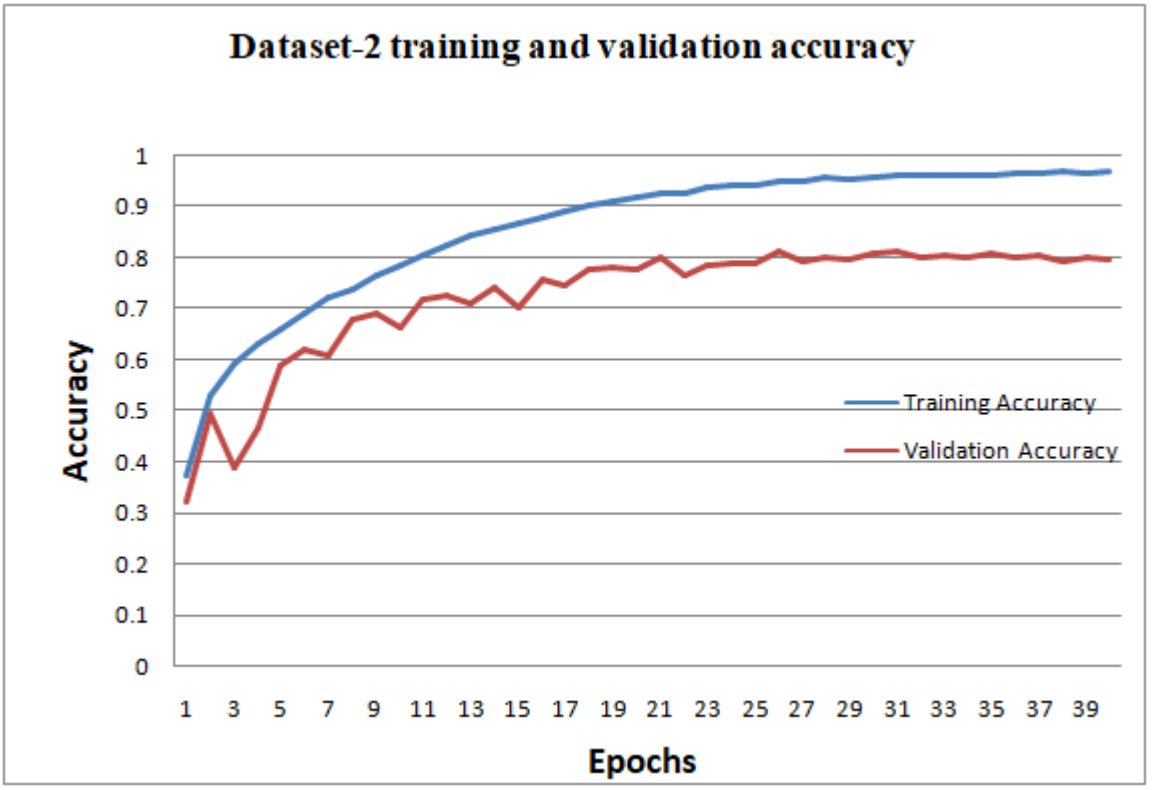


FIGURE 7.3: Capsule network based model accuracy on the training and validation data over 40 Epochs

Results reported in Table 7.5 show both the best performance achieved using capsule network and baseline system. Results show recall for each language and un-weighted average recall (UAR) using the baseline system and capsule network. It is evident from results that capsule network based LID system outperforms.

TABLE 7.5: Recall for each language and the un-weighted average recall (UAR) on the test set of dataset-2 using capsule network (%)

	Pashto	Pujabi	Saraiki	Sindhi	Urdu	UAR	EER
Baseline system	94	69	85	82	66	79.27	13.58
Capsule network	89	63	90	80	84	81.20	10.27

In order to analyze both the confusion and the discrimination performance of the systems considering all the languages pairs, Table 7.6 shows the confusion matrix of the capsule network based LID system. This confusion matrix shows that majority of the test utterances are predicted correctly using the LID system based on capsule network, however, Urdu -Punjabi confusion is still noticeable.

TABLE 7.6: Confusion matrix dataset-2, target language is shown in the Y-axis while the predicted language is shown in the X-axis

	Pashto	**Punjabi**	**Saraiki**	**Sindhi**	**Urdu**
Pashto	**99**	1	1	3	7
Punjabi	1	**70**	5	14	21
Saraiki	3	2	**100**	5	1
Sindhi	3	2	8	**89**	9
Urdu	1	10	2	5	**93**

7.5 Summary

This chapter focuses on the automatic spoken language identification using capsule network. The proposed approach of capsule network use convolutional neural network as feature detector. Several capsule layers are designed to effectively select representative frequency bands for each individual language. Experiment results shows that the proposed approach outperforms the previous state-of-the-art i-vector, BLSTM and merged BLSTM methods.

Language identification based on capsule network based performed better than the baseline models, irrespective of utterance duration. From this, we can conclude that capsule networks could capture the speech features very efficiently. Moreover, this approach is flexible to develop LID system for a language with minimum amount of training data.

Chapter 8

Conclusions and Future Work

This research focuses on the automatic spoken language identification of six major Pakistani languages namely Balochi, Pashto, Punjabi, Saraiki, Sindhi and Urdu. The motivation of the work presented in this thesis was to explore different speech features and existing language identification frameworks for Pakistani languages. This thesis mainly focuses on the utilization of acoustic features for automatic spoken language identification (LID) of very short utterance. This thesis has proposed the bidirectional long short term memory neural network framework for spoken language identification. This research will also be useful for retrieval and translation of multimedia content of Pakistani language as more and more local multimedia content is becoming available online. This research yields a number of original contributions in the area of LID, especially for Pakistani languages.

In this thesis, design and collection of telephonic speech corpus of Pakistani languages is discussed. Speech corpus is collected from native speakers of each language and transcription of each utterance in X-SAMPA format and in orthographic form is also prepared. The developed speech corpus is about 10.43 hours telephonic channel read speech, collected from 316 native speakers differing in gender, age and educational background. The collected database is divided into training and evaluation sets. This corpus is later on used for the development of language identification system.

Different acoustic features including Mel-frequency cepstral coefficients (MFCC),

gammatone frequency cepstral coefficients and perceptual linear prediction are employed for LID. Starting from GMM-UBM as baseline LID system, I-Vector based LID system is also developed. In order to increase the identification accuracy, different configuration of number of Gaussian and i-vector dimensions are investigated. The performance of the systems is evaluated on very short utterances and on test data of 3-seconds. Experimental results showed that I-vector based LID system outperformed GMM-UBM system. It is also evident from results that system accuracy improves with the duration of test utterance.

An end-to-end language identification system for very short utterances using bidirectional long short term memory neural networks is proposed. In this chapter, two BLSTM networks and trained using the spectrogram and cochleagram based feature vectors. Networks are optimized individually in context of number of layers and layers size. This work also proposed an approach to merge optimized BLSTM networks, to get the joint benefits of features. Merged output is forwarded to fully connected dense layers. Experimental results showed the superior ability of language identification using the cochleagram features based BLSTM. It is shown in results that merging of these BLSTM networks achieved good performance.

In addition to RNN based language identification, an end-to-end language identification system based on capsule network is also proposed. Optimal architecture of capsule network is found through random search strategy. Compared to i-vector and merged BLSTM models, the capsule network based LID system achieved much better results for very short utterances(Dataset-1) and 3-second test utterances (Dataset-2).

8.1 Future Work

During this research we have notably pushed forward the state of the art in LID but various parts of this work still have potential for more research. To cite just the ones that could be investigated in short to medium time are discussed below.

This thesis investigates various acoustic features such as MFCC, GFCC and PLP for language identification, Moreover, combined effect of these features is also

investigated. Feature extraction process can be expanded by extracting complementary information related to the phonotactic and prosodic properties of data. It may be beneficial to use complementary prosodic and phonotactic information along the acoustic features for the development of the LID system. Moreover, for phonotactic feature extraction there is a need to develop a phone decoder covering phonemes set of Pakistani languages. The available phone decoders can not be effectively used for South Asian languages that have different phonetic characteristics compared to the European languages.

The biggest challenge in the development of LID system for languages spoken in Pakistan is the availability of training speech data. Resources such as OGI or CallFriend do not exist for these under-resource languages. Initially, we have recorded the telephone speech data of six mostly spoken languages and there is need to continue work in this direction.

In this study, a linear classification technique was investigated for LID system. As an extension to this work, hierarchical classification can be adopted to initially classify languages into family groups and then make fine-grained between them. Instead of utilizing single feature set, different feature vectors can be employed at each level for better discrimination among languages.

Author Bibliography

- Farooq M. U., **Adeeba F.** and Hussain S., "X-vectors Based Urdu Speaker Identification for Short Utterances", in the Proceedings of O-COCOSDA 2019, Cebu, Philippines

- Farooq M. U., **Adeeba, F.**, Rauf S. and Hussain S., "Improving Large Vocabulary Urdu Speech Recognition System using Deep Neural Networks," in the Proceedings of Interspeech 2019, Graz, Austria.

- **Adeeba, F.** , Hussain, S. "Native Language Identification in Very Short Utterances Using Bidirectional Long Short-Term Memory Network", in IEEE Access, vol. 7, pp. 17098-17110 (2019)

- **Adeeba, F.** , Hussain, S. "Acoustic Feature Analysis and Discriminative Modeling for Language Identification of Closely Related South-Asian Languages", in Circuits Syst Signal Process (2017) (URL: https://link.springer.com/article/10.1007/s00034-017-0724-1)

- **Adeeba, F.**, Hussain, S. and Akram, Q. "Urdu Text Genre Identification", in the Proceedings of Conference on Language and Technology 2016 (CLT 16), Lahore, Pakistan. (URL: http://www.cle.org.pk/clt16/).

- **Adeeba, F.**, Hussain, S., Habib, T., Haq, E. and Shahid, K.S. "Comparison of Urdu Text to Speech Synthesis using Unit Selection and HMM based Techniques", in the Proceedings of 19th Oriental COCOSDA Conference 2016, Bali, Indonesia. (URL: http://www.ococosda2016.org/)

- Shahid, Kh. S., Habib, T., Mumtaz, B., **Adeeba, F.** and Ehsan Ul Haq. "Subjective Testing of Urdu Text-to-Speech (TTS) System", in the Proceedings of Conference on Language and Technology 2016 (CLT 16), Lahore, Pakistan. (URL: http://www.cle.org.pk/clt16/).

- Urooj, S., Shams, S., Hussain, S. and **Adeeba, F.** "Sense Tagged CLE Urdu Digest Corpus", in the Proceedings of Conference on Language and Technology 2014 (CLT 14), Karachi, Pakistan. (URL: http://cs.dsu.edu.pk/clt14/). Presentation

- Akram, Q., Hussain, S., **Adeeba, F.**, Rehman, S. and Saeed, M. "Framework of Urdu Nastalique Optical Character Recognition System", in the Proceedings of Conference on Language and Technology 2014 (CLT 14), Karachi, Pakistan. (URL: http://cs.dsu.edu.pk/clt14/).

- **Adeeba, F.**, Akram, Q., Khalid, H. and Hussain, S. "CLE Urdu Books N-grams", poster presentation in Conference on Language and Technology 2014 (CLT 14), Karachi, Pakistan. (URL: http://cs.dsu.edu.pk/clt14/).

- Habib, W. Basit, H. R., Hussain, S. and **Adeeba, F.** "Design of Speech Corpus for Open Domain Urdu Text to Speech System Using Greedy Algorithm", in the Proceedings of Conference on Language and Technology 2014 (CLT14), Karachi, Pakistan. (URL:http://cs.dsu.edu.pk/clt14).

- Akram, Q., Niazi, A., **Adeeba, F.**, Urooj, S., Hussain, S. and Shams, S. "A Comprehensive Image Dataset of Urdu Nastalique Document Images", in the Proceedings of Conference on Language and Technology 2016 (CLT 16), Lahore, Pakistan. (URL: http://www.cle.org.pk/clt16/).

- Ahmed, T., Urooj, S., Hussain, S., Mustafa, A., Parveen, R., **Adeeba, F.**, Hautli, A. and Butt, M. "The CLE Urdu POS Tagset", poster presentation in Language Resources and Evaluation Conference (LERC 14) 2014, Reykjavik, Iceland. (URL: http://lrec2014.lrec-conf.org/en/).

- **Adeeba, F.**, Hussain, S. "Experiences in Building Urdu WordNet,", in the Proceedings of 9th Workshop on Asian Language Resources (ALR9). (URL:http://www.dlsu.edu.ph/conferences/alr9/2011/)

References

[1] Population by mother tongue. http://www.pbs.gov.pk/sites/default/files//tables/POPULATION%20BY%20MOTHER%20TONGUE.pdf, Last accessed on November 27, 2017.

[2] F. Adeeba, Q. A. Akram, H. Khalid, and S. Hussain. CLE Urdu books n-grams. In *Conference on Language and Technology(CLT)*, 2014.

[3] F. Adeeba, S. Hussain, T. Habib, E. Haq, and K. S. Shahid. Comparison of Urdu text to speech synthesis using unit selection and HMM based techniques. In *Proceedings.Oriental COCOSDA*, 2016.

[4] E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu. Language identification: A tutorial. *IEEE Circuits and Systems Magazine*, 11(2):82–108, Secondquarter 2011. ISSN 1531-636X. doi: 10.1109/MCAS.2011.941081.

[5] Bing B. Jiang, Y. Song, S. Wei, J. H. Liu, I. V. McLoughlin, and L. R. Dai. Deep bottleneck features for spoken language identification. *PLOS ONE*, 9(7):1–11, 07 2014. doi: 10.1371/journal.pone.0100795. URL https://doi.org/10.1371/journal.pone.0100795.

[6] J. L. G. Baart and E. L. Baart-Bremer. *Bibliography of Languages ofNorthern Pakistan*. National Institute of Pakistan Studies, Quaid-i-Azam University, 2001.

[7] J. Bae and D. S. Kim. End-to-end speech command recognition with capsule network. In *Interspeech*, pages 776–780, September 2018.

[8] M. H. Bahari, N. Dehak, H. V. Hamme, L. Burget, A. M. Ali, and J. Glass. Non-negative factor analysis of Gaussian mixture model weight adaptation for language and dialect recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(7):1117–1129, July 2014. ISSN 2329-9290. doi: 10.1109/TASLP.2014.2319159.

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

[10] N. Bertoldi and M. Federico. *Cross-Language Spoken Document Retrieval on the TREC SDR Collection*, pages 476–481. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003. ISBN 978-3-540-45237-9. doi: 10.1007/978-3-540-45237-9_41. URL https://doi.org/10.1007/978-3-540-45237-9_41.

[11] T. K. Bhatia. *Punjabi: A Conginitive-descriptive Grammar*. Descriptive grammars. Routledge, 1993. ISBN 9780415003209.

[12] P. Boersma. Praat a system for doing phonetics by computer. *Glot International*, 5:341–345, 2001.

[13] L. Campbell and R. G. Gordon. *Language*, 84(3):636–641, 2008. ISSN 00978507, 15350665. URL http://www.jstor.org/stable/40071078.

[14] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I.

[15] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-carrasquillo. Support vector machines for speaker and language recognition. *Computer Speech and Language*, 20:210–229, 2006.

[16] F. Chollet. Keras. https://github.com/fchollet/keras, 2015.

[17] R. Cordoba, L. F. D. Haro, F. Fernandez-Martnez, J. Macias-Guarasa, and J. Ferreiros. Language identification based on n-gram frequency ranking. In *Interspeech 2007, 8th Annual Conference of the International Speech Communication Association. Proceedings. (Interspeech '07).*, volume 1, pages 354–357, August 2007.

[18] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, August 1980. ISSN 0096-3518. doi: 10.1109/TASSP.1980.1163420.

[19] N. Dehak, P. Dumouchel, and P. Kenny. Modeling prosodic features with joint factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2095–2103, 2007.

[20] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2011.

[21] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak. Language recognition via i-vectors and dimensionality reduction. In *INTERSPEECH*, 2011.

[22] R. Eldan and O. Shamir. The power of depth for feedforward neural networks. *CoRR*, abs/1512.03965, 2015.

[23] J. Elfenbein. Balochi phonology. In *Phonologies of Asia and Africa*, pages 761–776. Linguistic Society of America, 1997.

[24] J. H. Elfenbein. *The Baluchi Language*. Royal Asiatic Society Books. Taylor & Francis Group, 2002. ISBN 9780947593346. URL https://books.google.com.pk/books?id=0XH-GgAACAAJ.

[25] Ethnologue. Sindhi. https://www.ethnologue.com/language/snd.

[26] M. Farooq. An Accoustic Phonetic Study of Six Accents of Urdu in Pakistan, school = University of Management and Technology, address = Lahore,Pakistan, year = 2014,. Master's thesis.

[27] B. R. Glasberg and B. C. J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1):103 – 138, 1990. ISSN 0378-5955. doi: https://doi.org/10.1016/0378-5955(90)90170-T. URL http://www.sciencedirect.com/science/article/pii/037859559090170T.

[28] F. J. Goodman, A. F. Martin, and R. E. Wohlford. Improved automatic language identification in noisy speech. In *International Conference on Acoustics, Speech, and Signal Processing,*, pages 528–531 vol.1, May 1989. doi: 10.1109/ICASSP.1989.266480.

[29] R. G. Gordon. *Ethnologue: Languages of the World*. SIL International, 15 edition, 2005.

[30] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional LSTM networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052 vol. 4.

[31] G. A. Grierson. volume IX: Indo-Aryan family. Central group, chapter Part 1, Specimens of western Hindi and Panjabi, page 609. Calcutta: Office of the Superintendent of Government Printing, India, 1916.

[32] W. Habib, R. H. Basit, S. Hussain, and F. Adeeba. Design of speech corpus for open domain Urdu text to speech system using greedy algorithm. In *Conference on Language and Technology(CLT)*. Pakistan.

[33] T. Habibullah. and R. Barbara. *A Reference Grammar of Pashto [microform] / Habibullah Tegey and Barbara Robson.* Distributed by ERIC Clearinghouse [Washington, D.C.], 1996. URL http://www.eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED399825.

[34] A. O. Hatch, S. S. Kajarekar, and A. Stolcke. Within-class covariance normalization for SVM-based speaker recognition. In *INTERSPEECH*, 2006.

[35] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990. doi: 10.1121/1.399423.

[36] G. E. Hinton, A. Krizhevsky, and S. D. Wang. Transforming auto-encoders. In *Proceedings of the 21th International Conference on Artificial Neural Networks - Volume Part I*, ICANN'11, pages 44–51, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-21734-0. URL http://dl.acm.org/citation.cfm?id=2029556.2029562.

[37] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012.

[38] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[39] A. S. House and E. P. Neuburg. Toward automatic identification of the language of an utterance. i. preliminary methodological considerations. *The Journal of the Acoustical Society of America*, 62(3):708–713, 1977.

[40] V. Hubeika, L. Burget, P. Matejka, and P. Schwarz. Discriminative training and channel compensation for acoustic language recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 301–304, January 2008.

[41] S. Hussain. *Phonetic Correlates of Lexical Stress in Urdu*. PhD thesis, Northwestern University, 1997.

[42] M. Ijaz. Phonemic inventory of Pashto. 2003. URL http://www.cle.org.pk/Publication/Crulp_report/CR03_15E.pdf.

[43] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 448–456. JMLR.org, 2015. URL http://dl.acm.org/citation.cfm?id=3045118.3045167.

[44] S. Irtza, V. Sethu, E. Ambikairajah, and H. Li. Using language cluster models in hierarchical language identification. *Speech Communication*, 100: 30 – 40, 2018. ISSN 0167-6393. doi: https://doi.org/10.1016/j.specom.2018.04.004.

[45] A. K. V. S. Jayram, V. Ramasubramanian, and T. V. Sreenivas. Language identification using parallel sub-word recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, pages I–32, April 2003. doi: 10.1109/ICASSP.2003.1198709.

[46] Y. Jiao, M. Tu, V. Berisha, and J. M. Liss. Accent identification by combining deep neural networks and recurrent neural networks trained on long and short term features. In *INTERSPEECH*, pages 2388–2392, 09 2016.

[47] N. Karamat. Phonemic inventory of Punjabi. 2003. URL http://www.cle.org.pk/Publication/Crulp_report/CR02_21E.pdf.

[48] A. Keerio. *Acoustic analysis of Sindhi speech - a pre-curser for an ASR system*. PhD thesis, University of Sussex, February 2011. URL http://srodev.sussex.ac.uk/6325/.

[49] P. Kenny, G. Boulianne, and P. Dumouchel. Eigenvoice modeling with sparse training data. *IEEE Transactions on Speech and Audio Processing*, 13(3): 345–354, May 2005. ISSN 1063-6676. doi: 10.1109/TSA.2004.840940.

[50] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel. A study of interspeaker variability in speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(5):980–988, July 2008. ISSN 1558-7916. doi: 10.1109/TASL.2008.925147.

[51] J. Z. Khan. Syllabification rules in Pashto. 2002. URL http://www.cle.org.pk/Publication/Crulp_report/CR02_19E.pdf.

[52] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[53] S. Kiranyaz, T. Ince, A. Yildirim, and M. Gabbouj. Evolutionary artificial neural networks by multi-dimensional particle swarm optimization. *Neural Networks*, 22(10):1448–1462, 2009.

[54] M. A. Kohler and M. Kennedy. Language identification using shifted delta cepstra. In *The 2002 45th Midwest Symposium on Circuits and Systems, MWSCAS-2002.*, volume 3, pages III–69, August 2002. doi: 10.1109/MWSCAS.2002.1186972.

[55] T. Lander, R. A. Cole, B. T. Oshika, and M. Noel. The OGI 22 language telephone speech corpus. In *Fourth European Conference on Speech Communication and Technology.*

[56] T. Lander, R. A. Cole, B. T. Oshika, and M. Noel. The OGI 22 language telephone speech corpus. In *European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 817–820, September 1995.

[57] A. Latif. Phonemic inventory of Siraiki language and acoustic analysis of Voiced Implosives. 2003. URL http://www.cle.org.pk/Publication/Crulp_report/CR03_16E.pdf.

[58] John Laver. *Principles of Phonetics*. Cambridge Textbooks in Linguistics. Cambridge University Press, 1994. doi: 10.1017/CBO9781139166621.

[59] H. Li, B. Ma, and C. H. Lee. A vector space modeling approach to spoken language identification. *IEEE Transactions on Audio, Speech, and Language*

Processing, 15(1):271–284, January 2007. ISSN 1558-7916. doi: 10.1109/TASL.2006.876860.

[60] H. Li, B. Ma, and K. A. Lee. Spoken language recognition: From fundamentals to practice. *Proceedings of the IEEE*, 101(5):1136–1159, May 2013. ISSN 0018-9219. doi: 10.1109/JPROC.2012.2237151.

[61] X. Li. *Combination and generation of parallel feature streams for improved speech recognition*. PhD thesis, 2005. Chair - Richard M. Stern.

[62] C. Y. Lin and H. C. Wang. Language identification using pitch contour information. In *Proceedings. (ICASSP 2005). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, March 2005.

[63] Y. Liu, J. Tang, Y. Song, and L. Dai. A capsule based approach for polyphonic sound event detection. *CoRR*, abs/1807.07436, 2018.

[64] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno. Automatic language identification using deep neural networks, 2014.

[65] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno. Automatic language identification using deep neural networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5337–5341, May 2014. doi: 10.1109/ICASSP.2014.6854622.

[66] A. Lozano-Diez, R. Zazo, D. T. Toledano, and J. Gonzalez-Rodriguez. An analysis of the influence of deep neural network (DNN) topology in bottleneck feature based language recognition. *PLOS ONE*, 12(8):e0182580, 2017.

[67] A. Lozano-Diez, R. Zazo, D. T. Toledano, and J. Gonzalez-Rodriguez. An analysis of the influence of deep neural network (DNN) topology in bottleneck feature based language recognition. *PLOS ONE*, 12(8):e0182580, 2017.

[68] B. Ma, C. Guan, H. Li, and C. H. Lee. Multilingual speech recognition with language identification. In *7th International Conference on Spoken Language Processing, ICSLP2002*.

[69] A. Martin, A. Le, D. Graff, and J. V. Santen. 2007 NIST language recognition evaluation supplemental training set, 2017.

[70] D. Martinez-Gonzalez, O. Plchot, L. Burget, O. Glembek, and P. Matejka. Language recognition in i-vectors space. In *INTERSPEECH*, August 2011.

[71] D. Martnez, L. Burget, L. Ferrer, and N. Scheffer. i-vector based prosodic system for language identification. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4861–4864, March 2012. doi: 10.1109/ICASSP.2012.6289008.

[72] L. Mary and B. Yegnanarayana. Prosodic features for language identification. In *2008 International Conference on Signal Processing, Communications and Networking*, pages 57–62, January 2008. doi: 10.1109/ICSCN.2008.4447161.

[73] C. P. Masica. *The Indo-Aryan Languages*. Cambridge Language Surveys. Cambridge University Press, 1993. ISBN 9780521299442. URL https://books.google.com.pk/books?id=Itp2twGR6tsC.

[74] P. Matejka, P. Schwarz, J. Cernocky, and P. Chytil. Phonotactic language identification using high quality phoneme recognition. In *Eurospeech*, 2005.

[75] P. Mewaram. *A Sindhi-English dictionary*. The Sind Juvenile Co-operative Society, Hyderabad, Sind, 1910.

[76] D. Mostefa, K. Choukri, S. Brunessaux, and K. Boudahmane. New language resources for the Pashto language. In *Proceedings. Language Resource and Evaluation (LREC)*, 2012.

[77] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika. The OGI multi-language telephone speech corpus. In *International Conference on Spoken Language Processing*, volume 2, pages 895–898, 1992.

[78] S. M. S. Nagendra and M. Khare. Artificial neural network approach for modelling nitrogen dioxide dispersion from vehicular exhaust emissions. *Ecological Modelling*, 190(1):99–115, 2006.

[79] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, USA, 2010. Omnipress. ISBN 978-1-60558-907-7. URL http://dl.acm.org/citation.cfm?id=3104322.3104425.

[80] S. Nakagawa, Y. Ueda, and T. Seino. Speaker-independent, text-independent language identification by HMM. In *International Conference on Spoken Language Processing*, pages 1011–1014, 1992.

[81] J. Navratil. Spoken language recognition-a step toward multilinguality in speech processing. *IEEE Transactions on Speech and Audio Processing*, 9 (6):678–685, September 2001. ISSN 1063-6676.

[82] R. W. M. Ng, T. Lee, C. C. Leung, B. Ma, and H. Li. Analysis and selection of prosodic features for language identification. In *2009 International Conference on Asian Language Processing*, pages 123–128, December 2009. doi: 10.1109/IALP.2009.34.

[83] R. W. M. Ng, T. Lee, C. C. Leung, B. Ma, and H. Li. Spoken language recognition with prosodic features. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(9):1841–1853, September 2013. ISSN 1558-7916. doi: 10.1109/TASL.2013.2260157.

[84] P. Nihalani. Sindhi. *Journal of the International Phonetic Association*, 25 (2):9598, 1995. doi: 10.1017/S0025100300005235.

[85] Y. Obuchi and N. Sato. Language identification using phonetic and prosodic HMMs with feature normalization. In *Proceedings. (ICASSP 2005). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages 569–572, March 2005.

[86] R. D. Patterson and B. C. J. Moore. Auditory filters and excitation patterns as representations of frequency resolution. *Frequency selectivity in hearing*, pages 123–177, 1986.

[87] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, and G. Stemmer. The kaldi speech recognition toolkit. In *In IEEE 2011 workshop*, 2011.

[88] J. Qi, D. Wang, Y. Jiang, and R. S. Liu. Auditory features based on gammatone filters for robust speech recognition. In *2013 IEEE International Symposium on Circuits and Systems (ISCAS2013)*, pages 305–308, May 2013. doi: 10.1109/ISCAS.2013.6571843.

[89] S. Rauf, A. Hameed, T. Habib, and S. Hussain. District names speech corpus for pakistani languages. In *Proceedings.Oriental COCOSDA*, 2015.

[90] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1):19–41, 2000.

[91] L. J. Roderiguez-Fuentes, M. Penagarikano, A. Varona, M. Diez, and G. Bordel. Kalaka-3: A database for the assessment of spoken language recognition technology on youtube audios. *Language Resource Evaluation*, 50(2): 221–243, June 2016. ISSN 1574-020X.

[92] L. J. Rodrguez-Fuentes, M. Penagarikano, A. Varona, M. Daez, G. Bordel, D. Martaez, J. Villalba, A. Miguel, A. Ortega, E. Lleida, A. Abad, O. Koller, I. Trancoso, P. Lopez-Otero, L. Docio-Fernandez, C. Garcia-Mateo, R. Saeidi, M. Soufifar, T. Kinnunen, T. Svendsen, and P. Franti. Multi-site heterogeneous system fusions for the albayzin 2010 language recognition evaluation. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 377–382, December 20111.

[93] S. Sabour, N. Frosst, and G. E. Hinton. Dynamic routing between capsules. *CoRR*, abs/1710.09829, 2017. URL http://arxiv.org/abs/1710.09829.

[94] S. O. Sadjadi, T. Hasan, J. W. Suh, C. Zhang, M. Mehrabani, H. Boril, A. Sangwan, and J. H. L. Hansen. UTD-CRSS systems for NIST language recognition evaluation 2011. 2011.

[95] J. Salamon and J. P. Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283, March 2017. ISSN 1070-9908. doi: 10.1109/LSP.2017.2657381.

[96] M. V. Segbroeck, R. Travadi, and S. S. Narayanan. Rapid language identification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(7):1118–1129, 2015.

[97] M. Senoussaoui, P. Cardinal, N. Dehak, and A. L. Koerich. Native language detection using the i-vector framework. In *Interspeech*, September 2016.

[98] C. Shackle. *The Siraiki language of central Pakistan: a reference grammar*. School of Oriental and African Studies, University of London, 1976. ISBN 9780728600263. URL https://books.google.com.pk/books?id=xvW7AAAAIAAJ.

[99] O. D. Shaughnessy. *Speech communication: human and machine*. Addison-Wesley series in electrical engineering. Addison-Wesley Pub. Co., 1987. ISBN 9780201165203. URL https://books.google.com.pk/books?id=mHFQAAAAMAAJ.

[100] K. C. Sim and H. Li. On acoustic diversification front-end for spoken language identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(5):1029–1037, July 2008. ISSN 1558-7916.

[101] E. Singer, P. A. Torres-Carrasquillo, D. A. Reynolds, A. McCree, F. Richardson, N. Dehak, and D. E. Sturim. The MITLL NIST LRE 2011 language recognition system. In *Odyssey*, 2012.

[102] M. Soufifar, M. Kockmann, L. Burget, O. Plchot, O. Glembek, and T. Svendsen. i-vector approach to phonotactic language recognition. In *INTERSPEECH*, 2011.

[103] B. Spooner. Balochi: Towards a biography of the language. In Harold F. Schiffman, editor, *Language Policy and Language Conflict in Afghanistan*, pages 319–336. Brill, Leiden, 2011.

[104] S. Strassel, K. Walker, K. Jones, D. Graff, and C. Cieri. New resources for recognition of confusable linguistic varieties: the LRE11 corpus. In Haizhou Li, Bin Ma, and Kong-Aik Lee, editors, *Odyssey 2012: The Speaker and Language Recognition Workshop, Singapore, June 25-28, 2012*, pages 202–208. ISCA, 2012. doi: http://www.isca-speech.org/archive/odyssey_2012/od12_202.html.

[105] M. Sugiyama. Automatic language recognition using acoustic features. In *[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*, pages 813–816 vol.2, April 1991. doi: 10.1109/ICASSP.1991.150461.

[106] Z. H. Tan and B. Lindberg. Low-complexity variable frame rate analysis for speech recognition and voice activity detection. *IEEE Journal of Selected Topics in Signal Processing*, 4(5):798–807, 2010.

[107] A. Tjandra, S. Sakti, G. Neubig, T. Toda, M. Adriani, and S. Nakamura. Combination of two-dimensional cochleogram and spectrogram features for deep learning-based ASR. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4525–4529, 2015.

[108] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller. Approaches to language identification using Gaussian mixture models and shifted delta cepstral features. In *7th International Conference on Spoken Language Processing, ICSLP 2002*, pages 89–92. International Speech Communication Association, 2002.

[109] S. Urooj, S. Hussain, F. Adeeba, F. Jabeen, and R. Parveen. CLE Urdu digest corpus. In *Conference on Language and Technology(CLT)*, pages 47–53.

[110] A. Waibel, P. Geutner, L. M. Tomokiyo, T. Schultz, and M. Woszczyna. Multilinguality in speech and spoken language systems. *Proceedings of the IEEE*, 88(8):1297–1313, Aug 2000. ISSN 0018-9219. doi: 10.1109/5.880085.

[111] X. Wang, Y. Wan, L. Yang, R. Zhou, and Y. Yan. Phonotactic language recognition using dynamic pronunciation and language branch discriminative information. *Speech Communication*, 75(C):50–61, December 2015. ISSN 0167-6393. doi: 10.1016/j.specom.2015.10.001. URL https://doi.org/10.1016/j.specom.2015.10.001.

[112] X. R. Wang, S. J. Wang, J.E. Liang, and B. Xu. Improved phonotactic language identification using random forest language models. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4237–4240, March 2008. doi: 10.1109/ICASSP.2008.4518590.

[113] N. Wayn. Languages and their families. 2002. URL http://www.cle.org.pk/Publication/Crulp_report/CR02_17E.pdf.

[114] G. Windfhur. *The Iranian Languages*. Routledge Language Family Series. Routledge, 2012. ISBN 9780415622356. URL https://books.google.com.pk/books?id=HzHrtwAACAAJ.

[115] E. Wong and S. Sridharan. Methods to improve Gaussian mixture model based language identification system. In *International Conference on Spoken Language Processing (ICSLP2002)*, volume 3, September 2002.

[116] R. Zazo, A. Lozano-Diez, J. Gonzalez-Dominguez, D. T. Toledano, and J. Gonzalez-Rodriguez. Language identification in short utterances using long short-term memory (LSTM) recurrent neural networks. *PLOS ONE*, 11(1):e0146917, 2016.

[117] Q. Zhang, G. Liu, and J. H. L. Hansen. Robust language recognition based on diverse features, 2014.

[118] X. Zhang, X. Xiao, H. Wang, H. Suo, Q. Zhao, and Y. Yan. Speaker recognition using a kind of novel phonotactic information. In *2008 6th International Symposium on Chinese Spoken Language Processing*, pages 1–4, Dec 2008. doi: 10.1109/CHINSL.2008.ECP.94.

[119] W. Zhao, J. Ye, M. Yang, Z. Lei, S. Zhang, and Z. Zhao. Investigating capsule networks with dynamic routing for text classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3110–3119. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/D18-1350.

[120] M. A. Zissman. Automatic language identification using Gaussian mixture and hidden markov models. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 399–402 vol.2, April 1993. doi: 10.1109/ICASSP.1993.319323.

[121] M. A. Zissman and E. Singer. Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling. In *Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing*, volume i, pages I/305–I/308 vol.1, April 1994. doi: 10.1109/ICASSP.1994.389377.

[122] V. W. Zue and J. R. Glass. Conversational interfaces: advances and challenges. *Proceedings of the IEEE*, 88(8):1166–1180, Aug 2000. ISSN 0018-9219.