

Speech processing is the technique to process and analyze the speech spoken by human beings. The different speech processing techniques are speech coding, speech recognition, speaker verification, and speech identification. ASR is a method to recognize the content of speech uttered by a speaker. Speech identification is a technique to recognize the utterance of the speech that belongs to which language or dialect. Dialect identification is a subdomain in Language identification used to identify the dialects of speech of a particular language spoken by an unknown person. A dialect of a particular language is one form of language spoken in a particular region or environment of human beings where they live. Dialects are different from accents, grammar and pronunciation of the same language. Like other spoken languages, Telugu language (TL) is multiform of different dialects viz., Telangana, Costa Andhra, and Rayalaseema. To identify any language dialects, the standard database is very important. It is a very difficult task to implement a dialect identification system as there is no standard database for dialects and many variations in language.



Dr. S. Shivaprasad is working as Professor and HoD in the Department of CSE at Malla Reddy Engineering College, India. He obtained his Ph.D. degree from Kakatiya University, Warangal in 2021, M.Tech from SIT-JNTUH campus in 2013 and B.Tech from Kakatiya University, Warangal in 2010. He was more than 30 research papers and 12 years of experience.



FOR AUTHOR USE ONLY

Shivaprasad Satla, Sadanandam Manchala

Shivaprasad Satla
Sadanandam Manchala

Telugu Dialects Identification

Using Different Speech Processing Models



Shivaprasad Satla
Sadanandam Manchala

Telugu Dialects Identification

FOR AUTHOR USE ONLY

FOR AUTHOR USE ONLY

**Shivaprasad Satla
Sadanandam Manchala**

Telugu Dialects Identification

Using Different Speech Processing Models

FOR AUTHOR USE ONLY

LAP LAMBERT Academic Publishing

Imprint

Any brand names and product names mentioned in this book are subject to trademark, brand or patent protection and are trademarks or registered trademarks of their respective holders. The use of brand names, product names, common names, trade names, product descriptions etc. even without a particular marking in this work is in no way to be construed to mean that such names may be regarded as unrestricted in respect of trademark and brand protection legislation and could thus be used by anyone.

Cover image: www.ingimage.com

Publisher:

LAP LAMBERT Academic Publishing

is a trademark of

Dodo Books Indian Ocean Ltd. and OmniScriptum S.R.L Publishing group

Str. Armeneasca 28/1, office 1, Chisinau-2012, Republic of Moldova, Europe

Printed at: see last page

ISBN: 978-620-5-49663-3

Copyright © Shivaprasad Satla, Sadanandam Manchala

Copyright © 2022 Dodo Books Indian Ocean Ltd. and OmniScriptum S.R.L
Publishing group

FOR AUTHOR USE ONLY

FOR AUTHOR USE ONLY

ACKNOWLEDGEMENT

First of all I would like to express my sincere and greatest appreciation to my research supervisor **Dr. M SADANANDAM**, Assistant Professor, Department of CSE, KU College of Engineering and Technology, (Kakatiya University), Warangal for his supervision, motivation, encouragement, support and valuable suggestions throughout my Ph.D work. It is because of him I have understood the true meaning of the research and attained knowledge in paper writing. His mission and commitment towards high-quality work will always help me to grow and explore my innovative thinking. His concern for my well-being and his confidence in me and my work led to complete this research work. As a learner, I have gained so much knowledge and awareness with his regular interactions and technical expertise. I could not have imagined having a better advisor and mentor for my Ph.D work. I sincerely wish that my association with him shall continue in the future and expect remarkable results to come in future collaborations.

It gives me immense pleasure in expressing heartfelt thanks to the **Prof.T. Srinivasulu**, Professor & Dean, Faculty of Engineering & Technology, Kakatiya University, for his valuable suggestions, technical assistance, and support for the completion of my research work.

I am very much grateful for being a part of KU CE&T family and I owe a special debt gratitude to **Prof. P. Malla Reddy**, Principal, KU College of Engineering & Technology, Kakatiya University, for providing all the required facilities at the college level and permitting me to carry out my research work.

I am grateful to **Mrs. K Sravanthi**, Chairperson, Board of Studies for CSE & IT, Faculty of Engineering & Technology, Kakatiya University, for her valuable suggestions, co-operation and encouragement in completion of my research.

I would like to thank **Dr. N Ramana** Principle, University College of Engineering (KU), Kothagudem and **Mr. K Kishore Kumar** former Chairperson, Board of Studies for CSE & IT, Faculty of Engineering & Technology, Kakatiya University for their precious advice to complete this research work.

I would like to express my external appreciation towards my parents late **Mr. S. Yadagiri** & **Mrs. S. Subadra** for their immense trust in me and ever so understanding. It is impossible to describe the pain and the compromises they had gone through to see me as a doctorate. Heartfelt thanks to my beloved wife **Mrs. S. ANUSHA** for her dedicated support through balancing the things in the right way. I am much indebted to her for patience, devotion, and bringing up our children's **S. Kruthika** & **S. Nehansh**. Without her loving support, I could never have imagined getting this work done. I also thank my brother late **Mr. S. Raju**, sister **B. Aruna**, and brother in law **Mr. S. Mahesh** for their love, affection, support, care and encouraging words that certainly acted as a paddle and propel led to have a smooth sail in my academics.

My acknowledgement to all my colleagues especially Dr. M. Nirupama Bhat, Dr. D Radharani, Dr. L Jayakumar, Dr. S.V Phani Kumar, Mrs. Keerthi, Mr. Narendra, Chinna gopi, Anusha, Prameela and Sajida and all my students especially Sharanyu, Amareshwar, Sri Hasa, Jahnvi, Aravind, Mahesh, Anil, N. Tejaswi, Kinnera who lend me a helping hand for the completion of database creation and Ph.D work.

It will be endless to record and express to each and every one who has directly or indirectly extended their cooperation during my Ph.D work.

Kakatiya University, Warangal.

(**SATLA SHIVAPRASAD**)

2021

ABSTRACT

Speech processing is the technique to process and analyze the speech spoken by human beings. The different speech processing techniques are speech coding, speech recognition, speaker verification, and speech identification. Automatic Speech Recognition (ASR) is a method to recognize the content of speech uttered by a speaker. Speech identification is a technique to recognize the utterance of the speech that belongs to which language or dialect. Dialect identification is a subdomain in Language identification used to identify the dialects of speech of a particular language spoken by an unknown person. A dialect of a particular language is one of the form of language which is spoken in a particular region or environments of human beings where they live. Dialects are different from accents, grammar and pronunciation of same language.


Like other spoken languages, Telugu Language (TL) is multiform of different dialects viz., Telangana, Costa Andhra, and Rayalaseema. To identify any language dialects, the standard database is very important. It is a very difficult task to implement a dialect identification system as there is no standard database for dialects and many variations in language. This is the reason where the research in Telugu dialects is very less, even though it was the same scope of research work along with language identification. In this work, a system for identifying the dialect from Telugu speech utterances is proposed. As like, any pattern

recognition system Dialect recognition system consists of three phases namely feature extraction, Training, and Testing phases. Features of speech utterances play a prominent role in the recognition of dialect from the speech utterance. In this work, Spectral features of speech signals like Mel frequency Cepstral Coefficients (MFCC), Delta MFCC, Delta Delta MFCC, and prosodic features like Pitch, Intensity, Energy, formants, and Loudness are extracted and their role is established.

In order to improve the performance of system, Hybrid feature vectors and optimized feature vectors are derived from traditional features like MFCC and prosodic features. The different training models like the Hidden Markov model (HMM), Gaussian Mixture Model (GMM), Deep Neural Network (DNN), and K nearest neighbor (KNN) are used to implement the training phase of the system in order to analyze the behavior of features of speech signals for the classification.

The experiments are carried out with different feature vectors with different training models to achieve the better performance. The performance of proposed system to identify the dialects from Telugu speech is impressive.

List of Figures

1.1	Working of ASR	3
1.2	Pronunciation of vowels and consonants	9
1.3	Different forms of Consonant letter “  ”	10
1.4	Block diagram of database creation	14
1.5	Working of an average filter	18
3.1	Basic working of general model	43
3.2	MFCC feature extraction	46
3.3	<i>Mel</i> filter bank	49
3.4	Basic structure of HMM model	54
3.5	Basic structure of DNN model	60
3.6	Basic DNN Structure contain N-hidden layer.	60
3.7	Neuron structure	61
3.8	ReLU activation function	62
3.9	Training phase of HMM based dialect identification	64
3.10	Testing phase of HMM based dialect identification	65
3.11	Training phase of GMM based dialect identification	66
3.12	Testing phase of GMM based dialect identification	67
3.13	DNN based dialect identification	68
3.14	Performance of Dialect Identification with different models using MFCC	72
3.15	Performance of Dialect Identification with different models using MFCC + Δ MFCC + $\Delta\Delta$ MFCC	72
4.1	High and Low frequency	77
4.2	Different prosodic features of Rayalaseema dialect	78
4.3	Different prosodic features of Andhra dialect	79

4.4	Different prosodic features of Telangana dialect	79
4.5	Amplitude of speech signal	83
4.6	$K - NN$ model when $K = 1$ nearest neighbour	87
4.7	Working of $K - NN$ model with $K = 3$ nearest neighbours	88
4.8	Methodology used in Training Phase	89
4.9	Methodology used in Testing Phase	90
4.10	Performance of KNN model with different prosodic features	94
5.1	Basic diagram to extract new features	100
5.2	Block diagram of the proposed method for dialect identification	103
5.3	Principle components	106
5.4	Methodology to calculate PCA.	106
5.5	Optimized feature extraction	108
5.6	Training phase	109
5.7	Testing Phase	110
5.8	Performance of GMM with new feature vectors	112
5.9	Performance of HMM model with new feature vectors	113
5.10	Performance of DNN model with new feature vectors	114
5.11	Performance of HMM model with optimized features.	116
5.12	Performance of GMM model with optimized features	117
5.13	Performance of DNN model using Optimized features	118
6.1	Performance of Dialect identification system with different models using MFCC + Δ MFCC + $\Delta\Delta$ MFCC	123
6.2	GMM based Dialect Identification with MFCC and New feature Vectors	124
6.3	HMM based Dialect Identification with MFCC and New feature Vectors	125
6.4	DNN based Dialect Identification with MFCC and New feature Vectors	126
6.5	GMM based Dialect Identification with New and optimized feature Vectors	128

6.6	HMM based Dialect Identification with New and optimized feature Vectors	129
6.7	DNN based Dialect Identification with New and optimized feature Vectors	130
6.8	Analysis of DNN model identification time with different duration of test samples	132
6.9	The performance of HMM, GMM and DNN using proposed new features	134
6.10	The performance of HMM, GMM and DNN using optimized features	134
6.11	Comparison of proposed model with published work	136

FOR AUTHOR USE ONLY

List of Tables

1.1	Datasets of different dialects of Telugu Language	15
1.2	Different parameters are used in database creation	16
3.1	Parameters used in Training DNN	62
3.2	Telugu dialect database details	70
3.3	Performance of Dialect Identification using MFCC + Δ MFCC + $\Delta\Delta$ MFCC	71
4.1	Different pronunciations of same word corresponding to each dialect	77
4.2	Sample of words which have similar way of pronunciation	82
4.3	Different prosodic values of Costa Andhra Speech samples	84
4.4	Different prosodic values of Telangana Speech samples . .	85
4.5	Different prosodic values of Rayalaseema Speech samples .	85
4.6	Mean values of different prosodic features related to each dialect	91
4.7	Overall accuracies produced by different prosodic features with the K-NN model	93
5.1	Performance of GMM with new feature vectors	112
5.2	Performance of HMM with new feature vectors	113
5.3	Performance of system with DNN with new feature vectors	114
5.4	The performance of HMM-based dialect identification using optimized feature vectors	116
5.5	The performance of GMM-based dialect identification using optimized feature vectors	117
5.6	The performance of DNN-based dialect identification using optimized feature vectors	118

6.1 Performance of Dialect Identification system of Telugu Language using different models with MFCC + Δ MFCC + $\Delta\Delta$ MFCC	22
6.2 Performance of Dialect Identification System for GMM with new feature vectors	124
6.3 Performance of Dialect Identification System for HMM with new feature vectors	125
6.4 The performance of Dialect Identification System for DNN with new features	126
6.5 The performance of GMM based Dialect Identification System using optimized feature vectors	128
6.6 The performance of HMM based Dialect Identification System using optimized features	129
6.7 The performance of DNN based Dialect Identification System using optimized features	130
6.8 Analysis of time taken for identify the dialect of test utterance using DNN	131
6.9 Average Performance of Different models in dialect identification with different features	133
6.10 Comparison of proposed model with published work	136

Abbreviations

ASR	Automatic Speech Recognition
TL	Telugu Language
DI	Dialect Identification
SGD	Stochastic Gradient Decent
HMM	Hidden Markov Model
GMM	Gaussian Mixture Model
DNN	Deep Neural Network
K-NN	K-Nearest Neighbour
PCA	Principle Component Analysis
MFCC	Mel Frequency Cepstral Coefficients
FFT	Fast Fourier Transform
DCT	Discrete Cosine Transform
Δ MFCC	Delta MFCC Mel Frequency Cepstral Coefficients
$\Delta\Delta$ MFCC	Delta Delta Mel Frequency Cepstral Coefficients
SVM	Support Vector Machine
TEO	Teager Energy Optimization
XGB	eXtreme Gradient Boosting
ZCR	Zero Crossing Rate
STE	Short Term Energy
SDC	Shift Delta Coefficients
PRLM	Phone Recognition and Language Modelling
CAP	Credit Assignment Path

Contents

ABSTRACT	i
List of Figures	iii
List of Tables	vi
Abbreviations	viii
1 INTRODUCTION	1
1.1 Automatic Speech Recognition	1
1.1.1 Working of ASR	2
1.1.2 Advantages of ASR	3
1.1.3 Challenges of ASR	4
1.2 Dialect Identification	4
1.2.1 Advantages of Dialect Identification	6
1.2.2 Challenges of Dialect Identification	7
1.3 Telugu Language	8
1.3.1 History of Telugu	8
1.4 Dialects in Telugu Language	11
1.5 Telugu Dialect Database Creation	13
1.6 Tools Used in Database Creation	15
1.6.1 PRAAT Tool	15
1.6.2 Sony ICD-PX470 Digital Voice Recorder	16
1.6.3 Streaming Audio Recorder	17
1.6.4 Average Filter	18
1.7 Research Objectives	19
1.8 Major Contributions of The Research Work	19

1.9 Motivation for The Present Work	20
1.10 Organization of Thesis	21
2 LITERATURE SURVEY	24
2.1 Introduction	24
2.2 Different Features for Speech Processing and Dialect Identification	25
2.3 Different Approaches for Dialect Identification	32
2.4 Difficulties in Dialect Identification	38
2.5 Need for Different Approaches of Research in Dialects of Telugu Language	40
2.6 Conclusion	41
3 SPECTRAL FEATURES BASED TELUGU DIALECTS IDENTIFICATION FROM SPEECH	42
3.1 Introduction	42
3.2 Feature Extraction	45
3.2.1 Mel Frequency Cepstral Coefficients (MFCC)	45
3.2.1.1 Delta MFCC	51
3.2.1.2 Delta Delta MFCC	51
3.3 Different Statistical Models	53
3.3.1 Hidden Markov Model (HMM)	53
3.3.1.1 Gaussian Mixture Model:	57
3.3.1.2 Deep Neural Networks (DNN)	59
3.4 Proposed Methodology to Identify the Dialects of Telugu Language	63
3.4.1 Hidden Markov Model (HMM) based Dialect Identification System	63
3.4.1.1 Feature Extraction	63
3.4.1.2 Training of HMM for Dialect Identification	63
3.4.1.3 Testing phase of HMM for Dialect Identification	64
3.4.2 GMM based Dialect Identification System	65
3.4.2.1 Training phase of GMM:	66

3.4.2.2	Testing phase of GMM:	66
3.4.3	DNN based Dialect Identification System	67
3.4.3.1	Training phase of DNN	67
3.4.3.2	Testing phase of DNN	69
3.5	Results of HMM, GMM and DNN for Dialect Identification	69
3.6	Conclusion	73
4	PROSODIC FEATURE EXTRACTION TECHNIQUES TO IDENTIFY DIALECTS OF TELUGU LANGUAGE	74
4.1	Introduction	74
4.2	Prosodic Features for Dialect Identification	76
4.2.1	Pitch	76
4.2.2	Intensity	80
4.2.3	Energy	81
4.2.4	Formants	81
4.2.5	Loudness	82
4.3	Statistical Models used for Dialect Identification	86
4.3.1	K-Nearest Neighbor (K-NN) Algorithm	86
4.3.1.1	Working of K-NN algorithm	88
4.4	Dialect Identification Using Prosodic Feature	89
4.4.1	Training Phase	89
4.4.1.1	Testing Phase	89
4.5	Results	91
4.6	Conclusion	94
5	NEW FEATURES FOR TELUGU DIALECT IDENTIFICATION USING STATISTICAL APPROACHES	96
5.1	Introduction	96
5.2	New Feature Vectors for Dialect Identification	99
5.3	Methodology Used for Dialect Identification	101
5.3.1	Training Phase	101
5.3.2	Testing Phase	102
5.4	Optimized features for Dialect Identification	102

5.4.1	Principal Component Analysis (PCA)	103
5.5	Dialect Identification using Optimized Feature Vectors . . .	107
5.5.1	Training Phase	109
5.5.2	Testing Phase	109
5.6	Results of Dialect Identification with New Feature Vectors and Optimized Feature Vectors	110
5.7	Conclusion	119
6	PERFORMANCE EVALUATION OF DIALECT IDENTIFICATION SYSTEMS WITH DIFFERENT MODELLING TECHNIQUES	120
6.1	Introduction	120
6.2	Experimental Setup	121
6.3	Performance Evaluation of Dialect Identification using MFCC with Different Models	122
6.4	The Performance Evaluation of Dialect Identification with MFCC and Prosodic Features using HMM, GMM and DNN	123
6.5	Performance Evaluation of Optimized Feature Vectors	127
6.6	Comparison of different Dialect Identification System	132
6.7	Comparison study with reputed published work	135
7	SUMMARY AND CONCLUSION	137
7.1	Summary	137
7.2	Scope for Future Work	138
	BIBLIOGRAPHY	139
	LIST OF PUBLICATIONS	148

Chapter 1

INTRODUCTION

Speech processing is a technique to study the behaviour of speech signals and process the signal to identify the language and speaker from speech utterance [1]. Speech processing technique consists of different domains like Speech identification, Speaker identification, Speech verification, Language identification etc.[2]. It is used in several applications including Automatic Speech Recognition system, Interactive Voice Response System (IVRS), Background application for speech processing etc.

1.1 Automatic Speech Recognition

Speech is a wonderful medium through which human beings are able to express their thoughts, views, feelings and emotions. But now we are in an era of dealing with the computers where basically we will communicate with the computers using the peripheral devices. Through our continuous effort deals with artificial intelligence where we will make the machine act like human beings. With the advent of this technology, we communicate with devices in the same way as we communicate with human beings i.e. via speech.

Automatic Speech Recognition (ASR) is a branch of Artificial Intelligence (AI) and pattern recognition where humans interact with device's interface through their voices as they do with other human beings [3].

The primary endeavour to create procedures for speech recognition is based on the coordinate transformation of the speech signal into an arrangement of phoneme-like units. ASR has attracted much consideration on the last three decades and has seen the sensational change within the final decade. Nowadays it has diverse regions of applications like dictation, the program controlling, automatic phone calls, weather report data framework, travel data, IVRS system and the applications designed in which human being is not required for communication like IoT[4].

The ultimate aim of ASR technology is to process the speech and identify the language dialects etc., with best accuracy irrespective of some important agents such as the source person who had provided the speech input, surroundings around the speaker, size of the speech as well as the accent of the speaker and the quality of the thing that was used to record the speech.

1.1.1 Working of ASR

The working of automatic speech recognition is shown in Figure 1.1 and it will be follows like:

1. The process starts with the human speech that is given as input through microphone or other device.
2. The device generates a wave signal of speech utterances.
3. This wave signal might be mixed with noise due to surroundings of the speaker as well as it may also contain more pauses when the

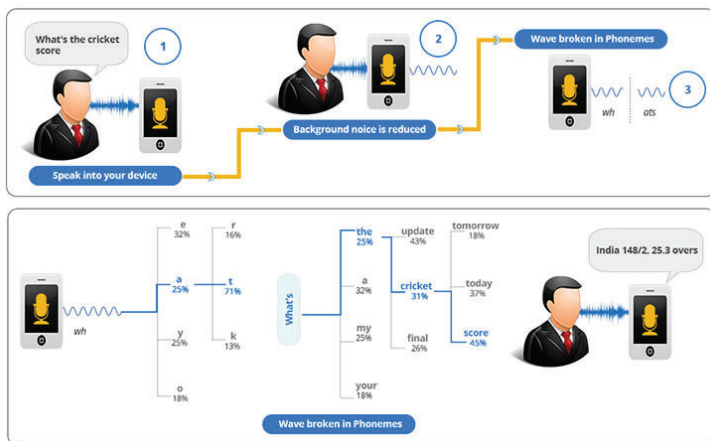


Figure 1.1: Working of ASR

speaker is uttering which can't be eliminated.

- In order to get more accurate results, those kinds of noises need to be removed. So the wave signal is filtered in the next stage.
- Then obtained filtered wave signal is splitted into phonemes that are the building sounds of language as well as words.
- Starting from the first phoneme, ASR will use statistical probability analysis in order to find out the next possible phoneme thus it will derive the words from them to sentences.
- Thus ASR understands human beings speech and responds to the speech in more meaningful manner.

1.1.2 Advantages of ASR

- ASR technology is faster because it takes less time to provide the required documents rather than providing the documents based on

queries.

2. With the help of ASR, we can avoid spelling mistakes which will quite happen if we type a query.
3. Through ASR we can pay our attention towards our regular work but it is quite different when we type our query.
4. It is very much helpful for disabilities like visually impaired people, who can't type with their hands, to send queries or inputs through speech.

1.1.3 Challenges of ASR

1. Accuracy of system is very important and it may not be changed with noise and any other factor.
2. ASR performance is not to be degraded when people talk quickly with different accents.
3. As there are a number of variation of dialects in a standard language, identification of dialects is difficult task.
4. It may face difficulties with limited vocabulary.
5. It requires vast periods of training.

1.2 Dialect Identification

Dialect identification is a technique to identify the dialects of particular region of spoken language from the speech utterance of unknown speaker [5]. Dialect can be termed as different variations within the

speech of a particular language. We can observe some variations such as selection of words, pronunciation, idioms and grammar in the speech given by people belonging to different regions. These variations are not only due to their differences in their geographical regions and environments [6].

There exists a small difference between the terms dialect and accent though they sound same. Accent will mainly focus upon how a speaker will pronounce a word whereas dialect is focus not only on pronunciation but also on speaker's selection of words, stress and intonation of speakers. We can identify the dialect based upon different factors like loudness, nasality and tonal. Due to dialect identification we can find out some interesting factors such as the speaker's age, their gender, some sort of their health status. This dialect identification helps us to improve some services like telephonic-services, e-health, e-education etc. for rural people also.

The dialect identification is used in automatic system like e-health, e-market system people calls directly are connected to the recognition specific system without participation of tele call receiver etc. As it is very useful in different applications which are easy to rural people, it is very essential to implement dialect identification system with good performance. In this, Dialect identification system for Telugu language is implemented.

1.2.1 Advantages of Dialect Identification

1. By identifying dialect from a speech, we can identify the region of the speaker.
2. Dialect identification helps us to develop some sort of e-services like e-health, e-market, e-education, telephonic services etc. Which helps for older and homebound people also.
3. We can even provide a better education services by identifying the dialect because if teaching happens in same dialect then it is very very easy to understand.
4. Automatic speech recognition system is improved
5. Improving Human-Computer interaction and enhancing its applications.
6. Improving the security for remote access communication.
7. Communication is very fast through speech rather than typing the text.
8. Dialect identification impacts ASR performance which is based on speaker's utterances in different dialects.
9. Along with the above, Dialect Identification is useful for
 - ◇ Formers
 - ◇ News broad casts
 - ◇ Voice search(automatic speech recognition)

- ◇ Twitter sentiment analysis
- ◇ Forensic operations
- ◇ Advertisements
- ◇ Charismatic speakers
- ◇ Regional tourism
- ◇ Dialect to Dialect translation

1.2.2 Challenges of Dialect Identification

1. In order to identify dialects, we need to prepare our own data sets if data sets are not readily available for some languages.
2. Identification and Extraction of feature vectors in order to design classifiers.
3. If the Standard language contains different dialects and there are many differences between the standard language and dialects.
4. For dialect identification, we need to train the system as part of implementation which may take more time to train huge speech.
5. The database of dialect may be dynamic as some people may add new regional words.
6. Design a dialect identification system using optimized features in order to improve the performance in terms of fast response and good accuracy.

7. Identifying dialects in a language is difficult task compared to language identification because a lot of similarities exist among dialects of same language.
8. It is very difficult to discriminate the features among the dialects of same language.

1.3 Telugu Language

The Telugu Language is a popular language which consists of Sanskrit elegance, Tamil sweetness and Kannada essence. Being one of the early languages in India, the Telugu language has been recognized “Ancient Language” in 2008 by the government of India. The majority Telugu speaking people belong to Telangana, Andhra Pradesh followed by Yanam, Karnataka and Maharashtra [7]. If we are to say that there are Telugu people in every nook and corner of the world it won't be an exaggeration because according to the Census 2018, the Language stood in 15th position in the world with a population of around 85 million speaking it, 3rd language in highest number of native speakers in India, with 6.93% and also it is the most popular spoken language of the Dravidian Language family.

1.3.1 History of Telugu

Telugu is a language that belongs to languages of Dravidian, spoken by people who especially live in the Indian states like Telangana, Andhra Pradesh, some part of Karnataka and Yanam district of Puducherry [8]. In the Telugu language, alphabets play a very prominent role in any

recognition system. The alphabets/letters are known as Telugu aksharralu. Telugu is richest of alphabets compare to any other Indian language. The Telugu language consists of Fifty six letters [Eighteen vowels and Thirty eight consonants], out of which Two vowels & Two consonants are deleted. Total, 52 core letters are present in the Telugu language. The vowels are known as “Achchulu” and consonants are known as “Hallulu”. In Telugu, vowels add short /o/ and /e/ along with /o:/ and /e:/ of Indo-Aryan languages [8]. The pronunciation of words is shown in Fig.1.2.

Vowel	Transliteration	IPA pronunciation
అ	a	/a/
ఆ	aa or A	/a:/
ఇ	i	/i/
ఈ	I or ee	/i:/
ఉ	u	/u/
ఊ	U or oo	/u:/
ఋ	R	/ru:/
ౠ	Ru	/ru:/
ఱ	- lu	/lu/
ఱ	- lu	/lu:/
ఎ	e	/e/
ఐ	E	/e:/
఑	ai	/ai/
ఔ	o	/o/
ౌ	O	/o:/
ౡ	ou	/au/
	(depends on transliterator)	
అం	am or aM	/o:/
	(depends on transliterator)	
అః	-	/aha/

Consonant	IAST character
క ఖ గ ఘ జ	k kh g gh n
చ ఛ జ ఝ ఞ	c ch j jh n
ట ఠ డ ఢ ణ	t th d dh n
త థ ద ధ న	t th d dh n
ప ఫ బ భ మ	p ph b bh m
య ర ల వ	y r l w (v)
శ ష స హ	ś ṣ ṣ h
ళ క్ష ఱ	ḷ a kṣ a ṛ a

(a) Pronunciation of vowels

(b) Pronunciation of consonants

Figure 1.2: Pronunciation of vowels and consonants

From the above table the consonants pronounced [7] like

- In Fig.1.2(b), the consonants ఙ (n) and ఞ (n) are cannot occurred alone. These consonants combined with other consonants.
- ట ఠ డ ఢ ణ (t th d dh n) which are in third row of Fig.1.2 are pronounced by

retroflex. In utterance of these sounds, wrap the tongue back and touch the end to the roof of the mouth.

- తధదధన (*t th d dh n*) are called as dental consonants which are in fourth row of Fig.1.2. During utterance of these consonants tongue touches the back of upper teeth.

A consonant vowel syllables are formed by adding diacritic form to consonants. For example

ఖ + ఈ (ీ) → ఖీ /kʰi/ + /i:/ → /kʰi:/

జ + ఉ (ు) → జు /dʒa/ + /u:/ → /dʒu:/

And also to make different forms of consonants, the VOTTULU plays an important role. Let's consider the letter బ and its different forms to make the different sounds shown in Fig.1.3.

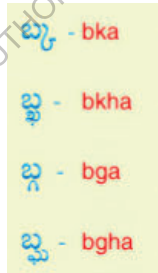


Figure 1.3: Different forms of Consonant letter “బ”

To pronounce the words of the Telugu language or identify the alphabets, prosody is a very important feature. Prosody features are supra segmental which are not concerned with phonic segments individually i.e., with vowels or consonants. But these are associated with the properties of syllables including stress, rhythm and intonation [25]. To identify the pronunciation in Telugu, it is required to represent the vibration

of vocal track which includes the mouth, nasal passages, throat and sinuses because Vocal folds cuts the overflow with vibration and produce a buzzing sound which is not clear as like its utterance.

These buzzing sounds involves the vocal track and their frequencies. The vocal tract frequencies are called as resonance. Resonance represents the shape and amplitude of speech wave of vocal tone.

The amplitude and shape of the speech signal depend on the vocal tract shape and length as well as cavities and structure of soundwaves.

1.4 Dialects in Telugu Language

A geographic region is referred to as a "Mandal". A dialect of a language is speech utterances spoken by a particular region of people [9]. Different regions of people, who speak the same language, with different dialects. Like other spoken languages, the Telugu language is multiform of dialects and it consists of three dialects namely Telangana, Costa Andhra, and Rayalaseema. The reasons for the formation of different dialects in Telugu Language

1. The emperors who rules the particular regions
2. Occupation of human being in the region
3. Accent of language which is habituated to more people in the region.

In this scenario, Kannada and Tamil influence Telangana dialect strongly and most of Urdu language is also mixed in Telangana dialect. The

distinct speciality of Rayalaseema is mixed of Kannada and Tamil accent due to geographical and historical. English and Sanskrit languages strongly influence the coastal Andhra. In some regions, Odia accent also influences the Andhra dialect.

The dialects of Telugu spoken language are three namely:

1. Telangana dialect: Which is popularly spoken in the almost all the districts of Telangana states (10 districts in combined Andhra Pradesh)
2. Coastal Andhra dialect: It is one of the popular dialect of Telugu Language which is spoken by people who live in nine districts of Andhra Pradesh state. (East Godavari, West Godavari, Krishna Guntur, Prakasam, Nellore, Srikakulam, Vizianagaram, and Visakhapatnam).
3. Rayalaseema dialect- The four districts (Chittoor, Anantapuram, Kurnool, and Kadapa) of Andhra Pradesh state people speak this dialect popularly.

Example of three slangs for the sentence “he came”

Telangana slang: “aaduachindu” ఆడుఅచ్చిండు

Rayalaseema slang: “vaaduvchinaadu” వాడువచ్చినాడు

Coastal Andhra slang: “vaduvachadu” వాడువచ్చాడు

Three dialects (Coastal Andhra, Telangana and Rayalaseema) of Telugu Language vary at semantic, morphological and phonological levels with respect to each other. At phonological level, these

dialects vary in terms of rhythm and intonation show observable variations. Therefore, in this research, the dialects of Telugu Language as Coastal Andhra, Telangana and Rayalaseema are identified from Telugu speech utterances of unknown speaker.

1.5 Telugu Dialect Database Creation

A huge Telugu Corpus has been created by collecting the speech samples of different speakers using various recorders which are suitable for different environments and situations. In this case, speech samples were collected in online and offline and edited by using device called streaming audio recorder.

The Chosen speakers ages are between 9 and 50 and different places are like working environments of office, schools, colleges and public place like parks, roadside, vendors etc.

The speakers who uttered speeches have been given freedom to speak on their own topic like own interests, habits, politics, self-description about family or home town etc.

The speech samples are recorded from different speakers including literates, illiterates and employees who are working in different occupations in different workplaces. The speech samples of different speaker were recorded in mono sound with the frequency of 44,100Hz. These speech samples are pre-processing using average filter to get rid of noise from the speech signal and also English words if any, in the speech the

utterances are removed.

Total Seven hours Five minutes duration of speech Corpus is created from three dialects of Telangana, Costa Andhra and Rayalaseema speakers out of which 2h 35min Telangana, 2h 47min Costa Andhra and 1h 43min Rayalaseema and stored in separate dialect folders.

The speech Corpus of each dialect is divided into two categories of samples which are used for training and testing randomly. The test sample lengths of three dialects are 3s, 5s and 8s. In case of Telangana dialect, speech Corpus of 2h 10min length is used for training and 40 min length is used for testing. In Rayalaseema dialect, 1h 10min length is used for training and 33min is for testing and 2h 07min is used for training and 40min length is used for testing in Costa Andhra. Fig. 1.4 represents the basic steps to create the Telugu dialogue Database from speakers speech utterances.

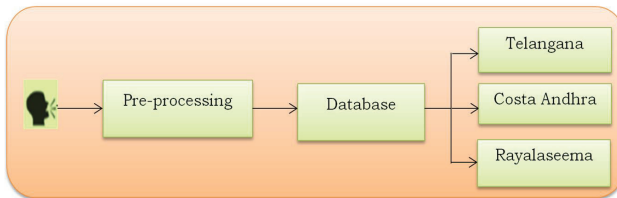


Figure 1.4: Block diagram of database creation

The following criterion is considered during recording speech samples:

- * Speakers involved in the utterance of speeches are different with respect to age, occupation and education background.

- * The speech was recorded in various places like public places of temples, roads, working places like offices, schools, colleges etc.
- * Speakers have given freedom to speak on their own topic.
- * Huge number of speech samples collected from three dialects of Telugu spoken language.

The details of database which was created for experiments are depicted in Table.1.1.

Table 1.1: Datasets of different dialects of Telugu Language

S.No.	Dialect	Total time of speech data	Speakers for each dialect	Period of each test sample	Age of speakers	Sampling Frequency
1	Telangana	2h 35 min	80	3-8s	9-50	44,100Hz
2	Costa Andhra	2h 47 min	75	3-8s	9-50	44,100Hz
3	Rayala-seema	1h 43 min	75	3-8s	9-50	44,100Hz

The different parameters are used to create the database are shown in Table.1.2.

1.6 Tools Used in Database Creation

1.6.1 PRAAT Tool

PRAAT is one of the famous and oldest speech analyzing tool, where we can record or edit the speech signal. The PRAAT tool can extract the prosodic feature of the speech signal like pitch, intensity, formats,

Table 1.2: Different parameters are used in database creation

Tools used for recording and editing	PRAAT tool, Sony ICD-PX470 4GB Digital Voice Recorder, Streaming Audio Recorder
Speakers age	9 to 50
Number of dialects	3
Number of speakers in each dialect	Telangana (80), Costa Andhra (75) Rayalaseema (75)
Sample rate	44kHz
Channels	MONO

etc. PRAAT could be an exceptionally adaptable device which examines the speech. It is useful for the investigation of spectrography, the synthesis of articulation, and neural systems. It is freeware and freely downloadable from “<http://www.fon.hum.uva.nl/praat/>” [10]. PRAAT tool provides different applications for generating waveforms, spectrogram (wide and narrow), playing recorded sound in reverse, track the pitch. It facilitates for filtering the speech signal with different filters like high-pass filter, low-pass filter, band-stop filter, and band-pass filters. It also useful, to refine the speech signal by enhancing regions, label words, syllables, segments, or individual phonemes. Animated plots and outline models can be designed for vocal tract which makes particular sounds by PRAAT tool [11].

1.6.2 Sony ICD-PX470 Digital Voice Recorder

The PX Series of Sony digital voice recorder is a useful device for business applications in which voice recordings are useful [12]. This PX series of Sony recorder has 4GB internal memory and can be extendable up to 32GB internal memory with memory cards.

It is an intelligent device which reduces noise using its cut function of noise. It also navigates individual recordings with its cues of track mark based system. Sony voice recorder smoothenes the voice to listen even the voices are recorded in public places which adds noise influences of wind, air, and other sounds. It eliminates noise and smooths the voice using different filters which are available in it. It also capable to filter out low frequency voices when voice is recorded using low cut filter. This recorder stores the voice file in formats viz., PCM and MP3 files.

The PCM file occupies more memory and contains larger sound details whereas MP3 files are compressed and takes less memory. The MP3 files are popular to store large recordings and huge speech utterances. It consist of auto recording facility and ICD-PX 470 which is capable of recording voice automatically without noise and sets optimisations setting in order to access local track frequencies. This recorder has S-microphone which is useful to capture loud or quiet voices or sounds. It can also capture low frequency voices.

1.6.3 Streaming Audio Recorder

Streaming audio recorder is a powerful and popular tool for enriching the voices in music entertainment. This recorder is useful to capture and record the voices or sound from music websites, various video platforms, radio stations, video streaming or voice charts effectively [13].

This device can record voices or sounds from different devices si-

multaneously including computer and microphone and stores recorded audio in different output formats like MP3, FLAC, and WMA etc. As the sounds are stored in high quality, these recorded sounds can be used in several applications including gadgets, speech processing applications. It is also help full to organise the audio file in batch format in turn editing is simplified. It provides the facility to match the audio files and split the large audio files into small files.

1.6.4 Average Filter

While recording the speech signal, there is some background noise or noise from headphones are also added. In order to delete the noise from the speech signal before extracting the features, the Average filter is used[14]. By using the average filter, replace all the features with their averages, thereby eliminating the noise. It can be used in images or signals to smoothen the blur or noise signal with preserving the boundaries [15]. The Fig.1.5 shows the before and after applying the average filter.

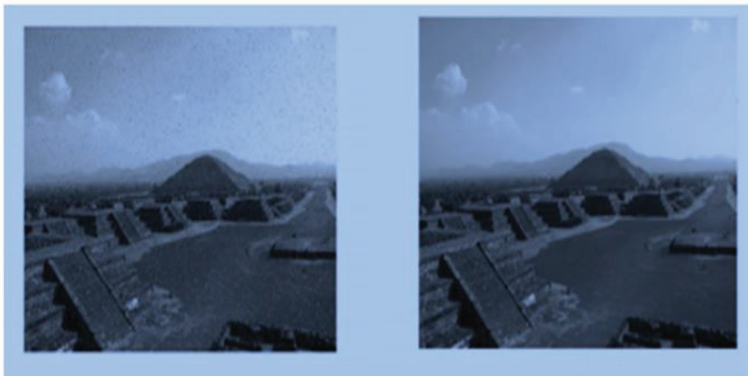


Figure 1.5: Working of an average filter

1.7 Research Objectives

- ❖ Find out an accurate model to identify the dialects of the Telugu Language from raw speech signal.
- ❖ As there is no standard database for Telugu language to do research in dialects, create a database for Telugu Language corpus for dialect identification, thus research work may be improved.
- ❖ As the performance of dialect identification system increases the accuracy of ASR, the performance of dialect identification should be improved in turn to provide new services in different fields like telemedicine, formers, health care which are helpful for older and home bound people.
- ❖ Implement the accurate model with different feature extractions to identify the dialects in different scenarios, using various features like spectral features and prosodic features of speech utterances and different statistical models.
- ❖ Deriving new feature vectors and optimized feature vectors from existing conventional features of speech utterances.

1.8 Major Contributions of The Research Work

- ❖ Text independent dialect identification task has been discussed.
- ❖ Database is created for the dialect of Telugu language.
- ❖ The significance of acoustic-prosodic features for text- independent dialect identification tasks has been demonstrated.

- ❖ New feature vectors are derived from different traditional prosodic features and spectral features and implemented dialect identification system using Hidden Markov Model (HMM), Gaussian Mixture Model, Deep Neural Network (DNN) and the K-Nearest Neighbor (KNN) model.
- ❖ The significance of hybrid features by concatenation of Spectral feature vectors (i.e. MFCC + Δ MFCC + $\Delta\Delta$ MFCC) and prosodic feature (i.e., PITCH and LOUDNESS) has been illustrated with different statistical models like HMM, GMM and DNN models.
- ❖ The prominent role of different features has been established in identification of dialects from raw speech signals.
- ❖ Investigated to determine the optimal features by reducing the dimensionality of feature vectors to speed up in identification of dialects and improve the Dialect Identification system performance.
- ❖ Derived the new features by modifying the frame level and utterance level features and created system using the Deep Neural Network (DNN) model to identify the dialects. The performance of system is compared with existing works.

1.9 Motivation for The Present Work

We believe in the fact that “A dialect belongs to one of the form of a particular language which is spoken by human beings belong to specific region or social group. The dialects in the language differences in pronunciation, grammar, syntax or vocabulary”, which is the primary

motivation for our present work. The Dialect identification helps us to develop some sort of e-services like e-health, e-market, e-education, telephonic services etc. which helps for older and homebound people also.

It can even provide a better education services by identifying the dialect because if teaching happens in same dialect then it is very easy to understand. Dialect Identification improving the performance of ASR, which are available in every electronic gadgets. The usage of hand held devices is increased rapidly in a wide range of different applications and grown in popularity in which speech is a one input. This motivated us to design an automatic Telugu dialect identification system using shortest utterance of speech. Due to the popularity and usage of speech based systems, the research area of dialect identification has attracted more researchers of speech processing to design a system for identifying regional dialects in Telugu.

Thus, the aim of the research is to improve the performance of dialect identification system and reduce the complexity of the identification system with respect to training and testing phases. This is achieved in our work by deriving new features from the spectral and prosodic features of speech frames and also with optimized features which have reduced in dimensionality and no redundant data.

1.10 Organization of Thesis

The thesis is organized into six chapters. A brief overview of the chapters and their contents are as follows:

Chapter 1 is introduction which explains the background of dialect identification and Telugu language clearly. It presents objectives of work and also explains challenges and motivation for Dialect identification.

Chapter 2 presents research work in the literature survey in the field of dialect identification and speech processing. This literature survey chapter starts with different features for speech processing and dialect identification. Further. It explains the existing approaches for dialect identification and their related works. Finally, the need of work is explained.

Chapter 3 deals the identification of dialects using spectral features, extracted from the speech signals. The chapter starts with the feature extraction of MFCC, Delta MFCC, and Delta Delta MFCC from speech utterance. This chapter explains the different statistical models like the Gaussian mixture model (GMM), Hidden Markov model (HMM), and Deep Neural Network (DNN) based Dialect Identification with the extracted features.

Chapter 4 describes the prosodic features based dialect identification system. The prosodic features also called supra segmental features, extracted from the speech signal are pitch, intensity, formants, energy, etc. This chapter establishes the role of prosodic features in identifying the dialects of Telugu Language from shortest duration speech utterances.

Chapter 5 deals with deriving new features from the traditional Spectral and Prosodic features. This chapter presents the procedure to derive new features from spectral features and prosodic features. This chapter also deals with the reduction of dimensionality of feature vectors using a Principal Component Analysis (PCA). The performance of models based dialect identification system is evaluated towards the end of the chapter.

Chapter 6 presents the overall performance evaluation of different approaches for dialect identification tasks of the Telugu language.

Finally, in Chapter 7, summary of the research work and contributions for implementing dialect identification system has been presented. The future scope and directions for further research work in dialect identification are discussed.

Chapter 2

LITERATURE SURVEY

2.1 Introduction

In this chapter, different existing techniques for Dialect identification systems (DIS) have been explored. The methods for identifying the dialects are divided into two categories. They are signal base dialect identification systems and text-based dialect identification systems.

In signal-based dialect identification approaches, the dialects are identified based on acoustic-phonetic and prosody information using raw speech signal. Whereas, in the text-based dialect identification system, the phonotactic, word level, and continuous speech data are considered to identify the dialects.

The main difference between these two approaches is that, the segmented and labelled speech corpus is not required in signal-based dialect identification to identify the dialects. The performance of text-based dialect identification systems is admirable than the signal-based dialect identification systems. Research in the dialect identification has been conducted for more than twenty years and many of the technology and techniques are developed.

This chapter is organized as follows: the different features for speech processing and dialect identification in section 2.2. In section 2.3, dif-

ferent approaches for dialect identification are explained. Section 2.4 represents the difficulties in dialect identification and the need of this research in dialect is mentioned in section 2.5. Section 2.6 gives the summary of conclusion.

2.2 Different Features for Speech Processing and Dialect Identification

George [16] surveyed several issues to recognize dialects of different languages.

Barly [17] used terminology pronounced by the same region human beings to identify the dialects from the speech. This was the first dialect identification system in order to check the Midland dialect.

Imène GUELLIL et al.[18] proposed a method for identify the dialects from Arabic speech in social media. This was implemented using supervised algorithm and lexicons. Authors implemented dialects recognition system to identify Algerian and French with 25086 words. It is extracted from social media messages using text based approaches [3].

Sreeraj V V Rajeev Rajan [19] identified the Malayalam dialects by considering the fusion of MFCC and Teager energy operator (TEO). The Support Vector Machine (SVM) was used for classification purposes. For identifying the dialects, the authors created the database in the studio environment, for four dialects of Malayalam with 300 speech samples for each dialect. The experiments were carried with MFCC and TEO

separately, and then combined feature vectors. The Performance of the system was 65% with MFCC, 75.3% with TEO, and 78% with combined features.

Saud Khan et al.[20] identified the Pashto Language dialects using MFCC and SVM. Authors created database by collecting the voice samples of different regions. The Cepstral features which are laid on optimal class boundaries are calculated by using different statistical parameters by SVM. These features were considered for identifying Pashto dialects. The result produced by SVM is impressive with MFCC features.

Suwon Shon, Ahmed Ali [21] recognized the dialects of Arabic Language by using different acoustic features like MFCC, Mel-Scale filter bank energies (FBANK) and spectral energies on Multi-Genre Broadcast 3 (MGB-3) data base. To identify the dialects, authors used end to end DID system and Siamese Neural network. Authors also considered the similarities and dissimilarities between the dialects for reducing feature vectors dimensionality and to increase the performance of dialect identification system. The performance of model with FBANK was 78% which was quite impressive compare to MFCC (i.e., 73%).

Mahnoosh Mehrabani et al.[22] analyzed the difference dialects of Arabic Language and also South Indian languages. Authors first calculated the spectral acoustics variance between dialects to clearly analyze the difference between dialects. To identify the dialects, authors used GMM

model with MFCC features. The statistical log likelihood was used to identify the dialects of Arabic Language. Second, the Authors proposed text-independent dialect identification by extracting the Pitch and Energy contour from continuous speech. The results produced by models are consistent. However authors did not consider the Loudness, Grammatical structure and stress etc., to identify the dialects.

Jacqueline Ibrahim [23] designed a dialect identification model to recognize the eight dialects of Indonesian speech. In order to implement Indonesian dialect identification, MFCC, spectral flux, and spectral centroid of speech utterances were used. The authors used the SVM classification method and k-means classification technique to implement this Indonesian dialect recognition system. The system reported 55% performance with SVM.

Mona Abdullah [24] reported the performance of Assamese dialect identification system with different models like GMM and GMM-UBM and MFCC feature vectors. Authors carried out experiments to identify different dialects of Assamese (Kamrupi and Goalparia) with 13 hours 30 minutes speech corpus of Assamese language. The system showed good performance with 85.7% for GMM and 98.3% for the GMM-UBM model.

Tanvira Ismail et al.[25] designed a model to recognize the Kamrupi dialects from spontaneous speech utterances. Spectral features like MFCC and GMM were used to design the dialect identification system

and reported accuracy was 89.9% with GMM.

Grabe, E et al. [26] designed a word-level-based model to classify the dialects of English using speech signals. The acoustic features were extracted from the word level of English speech. The speech Corpus of Intonational variations in English (IVIe) was used to carry out the experiments. Authors extracted features from Word Level but phonemes or syllables of the speech signal were not considered. Authors used different classification methods like SVM, XGB (Xtreme Gradient Boosting) ensemble algorithms. In this, the Authors consider only spectral features.

Mengistu, Melesew et al.[27] the Amharic language dialects were identified with independent of the text. MFCC and its variants like Δ MFCC, $\Delta\Delta$ MFCC were considered to carry out experiments on Amharic language speech Corpus which was collected from 100 speakers. Vector quantization (VQ), GMM, and hybrid approaches were designed to implement dialect identification and it was established the importance of MFCC, Δ MFCC, and $\Delta\Delta$ MFCC in the identification of dialects with 85.9% accuracy.

Trang et al. [28] used different features and reduction methods to create speech recognition system using MFCC with different features and PCA was used to reduce the feature vectors. It was observed that MFCC+PCA with HMM performed well and gave good results with 92.2%.

The phoneme of speech utterances were analyzed with different features dimensions using VBPCA [28].

Ghosal et al. [29] used different features based on occurrence pattern of Zero-Crossing Rate (ZCR) and Shot Time Energy (STE) with co-occurrence matrix. The performance was improved with these features using Support Vector Machine (SVM).

Nagaratna B et al. [30] implemented dialect recognition system to recognize the Kannada Dialects using SVM and Neural Networks. The MFCC, Δ MFCC and $\Delta\Delta$ MFCC features were used to identify the dialects. The Neural Networks was trained on sentence level features. The Hyper parameters of Neural Networks and SVM were chosen using grid search method. Overall, the Neural networks have performed well compared to SVM.

Ambareesh Prakash et al. [31] investigate the role of the different features like Spectral features and Prosodic features to recognize the nine British Isles dialects. The experiments were carried out using an Intonational variation of speech utterances and different classification methods SVM, Decision tree, and SVM ensemble classification. The authors used spectral features like Cepstral coefficients, Shift Delta coefficients (SDCs), Spectral Flux, Entropy and Prosodic features like Energy, Pitch of speech utterances to discriminate the nine dialects of British Isles. Overall Ensemble classifier outperformed compared to Decision tree and

SVM methods.

Chitturi, R et al. [32] used hybrid model to identify the Spanish Dialects by combining different dialect dependent features like Formants, Line spectral pairs and Hybrid features MFCCs+ Energy+ Pitch. The model GMM-SVM produced 30% improvement in the performance by applied the combination of features compared to individual feature performances.

Tzudir et al. [33] recognized the dialects of Nagaland, India. In this, authors considered three dialects Changki ,Chungli, and Mongsen. The idiosyncratic tone assignment was used to discriminate the different dialects. The GMM model and the spectral feature like MFCC and tonal features like F_0 , ΔF_0 , and $\Delta\Delta F_0$ were considered for the experiments. The model produced the 85.1% with MFCC and 86.2% with combined features.

Zergat et al. [34] proposed, a new approach for automatic speaker recognition using PCA method for reducing the feature vectors generated by model GMM-SVM at the back end. The results obtained by GMM-PCA-SVM were impressive than GMM-SVM alone. The error rate was reduced by 16% by GMM-PCA-SVM compared to GMM-SVM model.

Zissman et al. [35] identified the dialects of Latin American Spanish by using PRLM model. For this, authors created the extemporaneous and conversational database which is spoken in Spanish. The MFCC and

Delta MFCC feature were used to recognize the two dialects of Spanish i.e., Cuban and Peruvian dialects. The likelihood score was used to discriminate the corresponding dialects of Spanish.

Rong Tong et al. [36] proposed the hybrid features to identify the spoken language. Phonotactic, Prosodic, Acoustic features and combined features were considered and experiments were carried out with each feature set. Authors used NIST 1916 and 2003 LRE database for experiments. The different features extracted from speech utterances are spectrum, n-gram phonotactic, bag of sounds, duration and pitch. From the experiments, authors identified the 12- languages and concluded that prosodic features provide good results with shorter utterances and phonotactic features produced good results for longer speech utterances.

Wang et al. [37] proposed a new spectral subtraction method to increase the performance of Speaker Recognition System used in Vehicle interior. The traditional spectral subtraction method was produced less results in noisy conditions. To overcome the problem of traditional method, authors proposed a complex plane spectral subtraction method by modifying the phase difference between noise and clean signal is zero. The GMM model was used for classification purpose. The experiments were carried out on TIMIT database. Overall the model performance was impressive with complex spectral subtraction compared to traditional method.

Jain et al.[38] proposed the Language identification system with reduced features. Authors applied SVD (Singular Value Decomposition) method to reduce the feature vector size alternative to the traditional methods. Firstly, the GMM –UBM model with MAP method produces the super vectors. These super vectors are given input to the proposed SVD method to find the unique and important features by applying the proxy projection technique. The proposed method increased the accuracy by 8.4% compared to traditional i-vector based LID for 30s speech utterance. These experiments was conducted on CallFriend dataset which contains 12 languages.

Faragallah et al. [39] proposed a robust method to identify the speaker from noise speech. Authors proposed two methods Multiple Kernel Weighted Mel Frequency Cepstral Coefficient (MKMFCC) and support vector machine (SVM) to identify the speakers. Firstly, extracted the cepstral coefficients, accelerated and differential coefficients from speech utterances then MKMFCC and SVM methods were used. The MFCC-SVM produced good results compare to MKMFCC-SVM method.

2.3 Different Approaches for Dialect Identification

Zissman et al.[40] identified the dialects of Spanish in the Miami corpus by using the Phone Recognition followed by Language Modeling (PRLM). The PRLM model produced better results in classification of dialects of Cuban and Peruvian dialects in Spanish language. The accu-

racy produced by model is 84%.

Ma et al. [41] designed a system to recognize the three Chinese dialects by using MFCC features and multi-dimensional pitch flux features. The GMM model used to classify the dialects with test sample duration was 15s. The error rate is reduced by 30% by adding pitch flux features to MFCC. The dialect identification model reported 90% performance for identify three dialects.

Alorfi et al. [42] identified the Arabic dialects (Gulf and Egyptian Arabic) by using the ergodic HMMs with MFCC features. In this the phonetic differences between two Arabic dialects was considered.

Chittaragi et al.[43] Modelled Kannada dialect identification model to recognize five Kannada dialects with prosodic features and different speech processing models. The author's used SVM and Neural networks for the experiments and achieve good results for the shortest duration of test samples.

L. R. Arla, et al. [44] proposed a multiclass language identification using Convolution neural network method. For this, authors used MFCC spectral features which are extracted form short duration (2 to 4ms) of speech samples to identify the four Indian languages (Telugu, Bengali, Tamil and Gujarati). The CNN model gave 88.82% accuracy, which is good accuracy when compared with machine learning models.

Kim, H Park, J S [45] proposed the language identification by using speech rhythm as a feature for multi-lingual ASR. The SVM and i-vectors were used for the LID and proved that Rhythm provided the very good information about language-discriminative information, compare to remaining acoustic features. The computational cost and efficiency of the proposed system was quite impressive by using rhythm as a feature vector.

Reddy et al.[46] GMM importance was proved in speech processing applications like LID using spectral features in order to create a compact reference model.

Chen et al. [47] a speech-based NAO robot was proposed. The speech utterances were considered as inputs in the prototype of the NAO robot. The Hidden Markov model and GMM were used to implement this system. The results were impressive with HMM.

Sadanandam M and Prasad V [48] proposed spoken language identification system using robust features. For this, 12 MFCC features and five formant features were extracted from short duration of speech samples. The 17-dimnesional feature vectors were reduced to 8-dimentional feature vectors. This dimensionality reduction improved the performance in LID in Indian Languages.

Ibrahim et al. [49] Arabic dialects were recognized with GMM model. To carry out experiments, spectral and prosodic feature vectors were ex-

tracted from seven Quranic accents of Malay speakers. In this work, the authors used GMM and reported the performance of the system was more with combined features compared to individual feature vectors.

Tong et al. [50] explored experiments with different features to establish their role in speech processing. The authors also established the role of low-level features like MFCC and Pitch.

Chittaragi and Koolagudi [51] The dialects of Telugu were recognized using prosodic features with the Nearest neighbor algorithm. The authors considered the small database for the experiments. The reported performance was 78% with the prosodic feature.

Chittaragi et al.[52] Dialect identification was proposed to recognize nine dialects of British English using spectral features and different statistical methods. SVM, Gradient boosting methods were used and achieved the results of 78.8% and 80.5% respectively.

Ferrer et al.[53] proposed lower-level feature vectors-based system to identify the dialects in L1-English and L1-Japanese. The role of the MFCC feature vector and prosodic feature (pitch, duration, and intensity) was established in identifying the dialects in said two languages. The experiments were carried out using GMM with different individual and combined feature.

S. Shabani and Y. Norouzi [54] proposed a recognition system to identify

the discrete spoken words of Persian language by using the Neural networks with Principle Component analysis (PCA). For this, authors collected the data words by recording the predefined words from 20 speakers, each one uttered 10 words. 150-MFCC features are extracted from uttered speech words. Dimensionality reduction technique with PCA was used to report small training time and achieved good performance with 86%.

Manjushree B et al. [55] proposed a system, to enhancement the speech signal which is used in speech and emotion recognition systems. When the speech is recorded, the background noise may be added to the speech signal then it is reduce the performance of system. In this, authors extract the MFCC features from noisy signals and given to PCA model for reduced feature. By calculating Eigen values, eliminate the noise features from the signal and applied to HMM model to enhance the signal and identify the emotion from speech signal.

Shen et al. [56] analyzed PRLM framework and implemented dialect recognition system to identify the dialects in Mandarin and English languages. This PRLM gave a good performance with syllable features.

Torres-Carrasquillo et al. [57] developed a system to recognize dialects of Mandarin and Spanish languages using shift delta cepstral(SDC) features with the GMM model. They reported 70% of accuracy in the identification of dialects.

Torres-Carrasquillo et al.[58] proposed dialect identification system with SDC features using GMM modelling technique to recognize dialects of Chinese and Arabic. The hybrid features were derived using the SDC features and vocal tract length normalization.

Chen et al. [59] proposed a system which automatically identifies American vs. Indian English accents by using the set of biphones by recognizing phones of speech utterances using log-likelihood ratios.

Santosh Gaikwad et al.[60] proposed accent identification system by comparing the different acoustic features of speech signal. The acoustic features used in the system was formant frequency, energy and pitch to identify accents for English language.

Qin Yan et al.[61] published the comparative study for identify the accents of English (American and British) using acoustic features and prosodic features with different speech processing models.

Gang Liu et al.[62] proposed an automatic dialect identification system using hierarchical UBM model instead of normal UBM model and different features like SDC and perceptual minimum variance distortion less response (PMVDR) to identify Spanish dialect.

FadiBiadisy et al.[63] establish the importance of the prosodic feature by implementing the Automatic Arabic dialect recognition system and reported good accuracy for 30s test utterance.

Sadanandam et al.[64] proved the HMM performed well in speech processing applications by implementing LID using the spectral and prosodic feature vectors which are derived from the windowed speech signal. The HMM models represents huge feature vectors and their relationship in compact effectively.

H. C. Soumia et al.[65] studied the performance of SVM and DNN model by designing dialect identification of Algeria Arabic dialects using acoustic features. It was proved that the DNN performed well with the shortest duration of test samples.

Taylor J.H et al.[66] examined the performance of the Hidden Markov Model in speech processing applications: Speech recognition system (SRS) for identifying India Gujarati language. With HMM, the good performance was reported.

Ignacio Lopez et al.[67] studied the purpose and working of Deep Neural Networks (DNN) of sequential model for identifying the language in speech processing applications. The sequential DNN model performed well.

2.4 Difficulties in Dialect Identification

Identification of dialects in a particular language is a more complex task compare to recognize the language form from raw speech because there exists more similarity in different dialects in a particular language.

An initial problem in the fast years of Dialect Identification was the lack of availability of database.

In Dialect Identification, the feature extraction is the first phase. This step is part of training phase and testing phase of the system and plays a vital role in the accuracy and performance of the system. The representation of huge features to identify the dialect is very difficult and also, the selection of feature category and feature vector in particular category influences the performance of system so that feature selection and extraction is important task for dialect identification.

The speech produced by humans is estimated by the shape of their vocal tract (including tongue, teeth, etc.). In order to identify the sound produced by human being correctly, it is required to represent the shape of vocal tract. It is very difficult to represents the envelope of power spectrum and extract the features from the spectrum. In a language, some of the words are pronounced same way in different dialects. It is very difficult to discriminate those words with spectral features.

The nasal sounds play vital role to discriminate similar words of a language. The extraction of Prosodic features from short duration speech samples is complex task. The accuracy of system depends on the dimensionality of feature vectors and reducing of features and eliminate redundant data.

For any speech processing applications including dialect identifica-

tion, modelling techniques is also an important task in order to represent huge data samples in compact representation. So that, the selection and implement a modelling technique for dialect identification is an important task.

2.5 Need for Different Approaches of Research in Dialects of Telugu Language

From the review of the existing systems, it is observed that very few attempts were made to identify the dialects of Telugu language. It is observed that the dearth of database is primary reason to less research in identification of Telugu dialects. The Dialect Identification helps in improve some services like telephonic-services, e-health, e-education etc. for rural people also. The dialect identification is used in automatic speech based system like e-health, e-market system people calls directly are connected to the identification system without participation of tele call receiver etc., as it is very useful in different applications which are easy to rural people. As there is no standard database, Telugu language database has been created by us from collecting speech samples of different people of age in different places.

It is also observed that, most of the dialect identification system of different languages extracted either spectral features or combination of spectral features to increase the accuracy. The Dialect identification with respect to prosodic features is very less even thorough, these features provide very important cues with respect to nasal sounds to clearly

discriminate the dialects of a language. So, in this an attempt had been made to identify the dialect with Spectral features and prosodic features and also combined the spectral and prosodic features to clearly discriminate the dialects and also to increase the accuracy of models.

When combine the different features the resultant dimensional size of the feature vectors is huge and it takes long time to train the model. So, it is required to apply the different dimensionality reduction techniques to reduce, the time taken to train the model and also increase the accuracy of models. The models also very important to identify the dialects. The different Machine learning and Artificial Intelligence are techniques are used in identification of dialects. We made an attempt to apply the Deep learning techniques to identify the Telugu Dialects from short duration of speech samples.

2.6 Conclusion

In this chapter, a review of different methods to extract feature vectors used in speech processing including dialect identification has been discussed. The importance of different modelling techniques in the state of art system was presented with different kind of feature vectors of speech utterances. The difficulties need of new approach in designing dialect identification is also presented.

Chapter 3

SPECTRAL FEATURES BASED TELUGU DIALECTS IDENTIFICATION FROM SPEECH

3.1 Introduction

Dialect Identification system is to identify the dialects of speech utterance of human being in Telugu Language using shortest duration.

As specified in chapter 1, Telugu language consists, three dialects namely Telangana, Costa Andhra, and Rayalaseema. The challenge of research is to improve the accuracy of identification with shortest duration of utterance. There are different types of features are considered like acoustic, stress, intonation etc., to discriminate in a particular language [31]. The shape of the vocal tract determines the speech which is generated/uttered by human beings. The speech signal representation depends on the shape of the vocal tract such that if it is determined correctly, then the sound of speech is represented correctly [33]. The shape of the vocal tract is represented by the envelope of the power spectrum in time. The spectral feature of signals represents the envelope of the power spectrum [35]. The acoustic features which are derived from the raw speech signal are important features including spectral features and prosodic features.

The Mel Frequency cepstral coefficients (MFCC) features describe the spectral properties of speech signal which are useful to discriminate

the three dialects of Telugu Language. The general model of dialect identification is shown in the Fig.3.1.

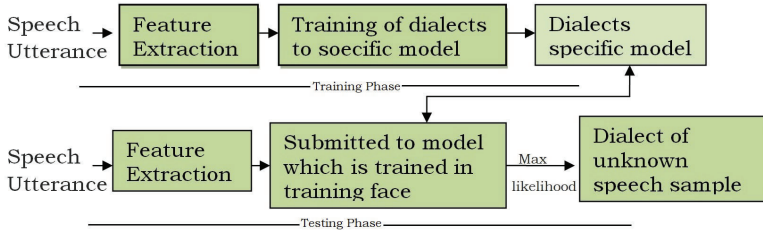


Figure 3.1: Basic working of general model

Like any speech processing model, the general model of dialect identification system consists of two phases: training phase and testing phase as specified in Fig.3.1.

As in the state-of-the-art systems, there are popular methods to extract the feature vectors from the raw speech signal by using framing and windowed speech like MFCC, Prosodic, TEO, SDC features, etc.[31], [33], [41]. These feature extract methods must be similar in training and testing phases.

In the first phase of general model, the suitable feature vectors are extracted from speech utterances of human beings. These extracted features are given as input to training model to create the reference model. In testing phase, the feature vectors of test speech utterances are given to the input for trained models of dialects and evaluated to get likelihood score. The model with maximum likelihood score gives the dialect of unknown utterance of speech.

In the literature, there are several speech processing modelling techniques like HMM, GMM, SVM, PRLM and DNN etc., which are derived in the training phase of the dialect identification system to create a reference model for compact representation of huge feature vectors.

In this Chapter, dialect identification system has been designed using MFCC feature vectors and its variants. By using HMM, GMM and DNN models, dialect identification systems have been designed. Initially, the database of Telugu language for dialect identification has been created by collecting speech samples from different age people and different places. Then HMM, GMM and DNN based dialect identification systems have been designed.

This chapter is organised as follows:

Section 3.2 explains the feature extraction process and phases of MFCC and delta MFCC and delta delta MFCC. Section 3.3 describes different statistical models like HMM, GMM and DNN. Section 3.4 represents the methodology used to design the Telugu dialect identification system. The results of different dialect identification systems have been presented in section 3.5 and conclusions are described in section 3.6.

3.2 Feature Extraction

The feature extraction from input data is the first phase of any pattern recognition system. This step is part of training and testing phases of the system and these are similar. The feature vectors play a vital role in the accuracy and performance of the system [75]. The selection of feature category and feature vectors in particular category influence the performance of system. In this proposed system, Mel Frequency Cepstral Coefficients (MFCC) and its derivative features are used to implement dialect identification system to identify Telugu dialects.

3.2.1 Mel Frequency Cepstral Coefficients (MFCC)

MFCCs are the most popular features which describe speech signal that is taken as an input from the speakers in ASR systems. It is useful for getting the spectral features from human voice. These features are derived based upon the frequency domain by using the Mel scale which is dependent upon human ear scale. Thus it will mimic the human ear. The main phases involved in MFCC feature extraction is shown in the Fig.3.2.

The extraction of MFCC feature vectors from short duration of speech involves Pre-emphasis of speech signal, Framing and windowing of speech signal, Fast Fourier Transformation (FFT), Mel spectrum calculation and applying Discrete Cosine Transformation (DCT).

Pre-Emphasis: It is used to maintain the higher frequencies in speech signal. This phase balances the spectrum of speech signal at higher

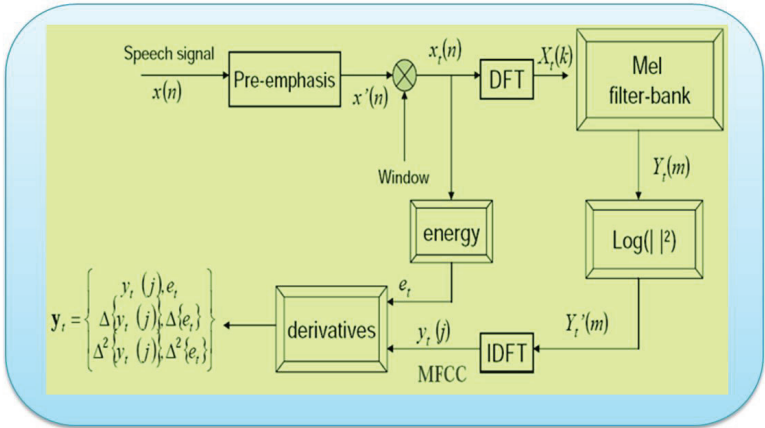


Figure 3.2: MFCC feature extraction

frequency regions because the glottal source for voiced sounds has an -12 dB/octave (approximately) slope [28]. However when we pronounce any speech signal, the acoustic energy generated from lips +6 dB/active (approximately) to sustain the spectrum. When the speech signal is recorded by micro phone, it deviates with original spectrum of vocal tract by -6 dB/active octave. To reduce the glottal effects of vocal tract, pre-emphasis is used. The equation 3.1 represents general transformation function used for pre-emphasis filter

$$H(z) = 1 - sz^{-1} \quad (3.1)$$

Where, z is the input vector. The slope of the filter is controlled by the value of s . In this the value of s used is 0.9.

Framing and Windowing: As speech signal is quasi stationary signal in nature, short period of speech signal is considered to extract the features of important characteristics of human speech signal. Therefore, the analysis of speech signal always is performed on short segments in which assumed that the characteristics of speech signal are stationary. In this, the size of the window is 20ms and advanced every 10ms. The 20ms analysis window was used to get good spectral features and 10ms was used to track the temporal characteristics of speech utterance.

While taking the DFT on the signal, a Hamming window is added to each frame for increasing harmonics, smoothening the edges, and decreasing the effects of edge by taper the signal to frame borders.

Fast Fourier Transform (FFT):

The speech signals initially in time domain format. To convert the signal from time domain to the frequency domain, applied the FFT. In FFT, while converting signal into the frequency domain, every frame should contain the same number of samples (N_m samples). The Fast Fourier transformation, on the given set of N_m samples is shown in equation 3.2.

$$T_k = \sum_{m=0}^{N_m-1} T_m e^{\frac{-j2\pi km}{N_m}} \tag{3.2}$$

where $k = 0, 1, 2, 3, \dots, N_m-1$

For converting the time domain signal to frequency domain, basically

FFT and DFT methods are used. The basic difference between these two methods is, in DFT, N-M samples are transformed to frequency domain whereas in FFT, the frame was divided into smaller DFT's here N is number of samples and M is overlapped samples.

Mel Spectrum:

After applying the FFT, the resultant signal was passed as input to the band-pass filters for compute the MEL Spectrum. The band-pass filters also known as Mel-filter bank. A *Mel* is nothing but the measure of perceived frequency by the human ear. It does not correlate linearly to the physical frequency of the speech because of the auditory system of human being where the pitch is not linearly perceived. The *Mel* scale is logarithmic spacing above $1kHz$ but below $1kHz$ it is linear in frequency spacing [4]. The conversation from physical frequency to *Mel* frequency is done by equation 3.3.

$$Mel(f) = 1125 \times \ln(1 + f/700) \quad (3.3)$$

where f denotes the physical frequency in Hz , and *Mel* denotes the perceived frequency [2].

The filter's centre frequencies are usually equally spaced on the frequency axis. The triangular band pass filters are used to get required cepstral coefficients and also smoothen the harmonics. The Fig.3.3 shows the triangle filter banks with Mel frequency warping.

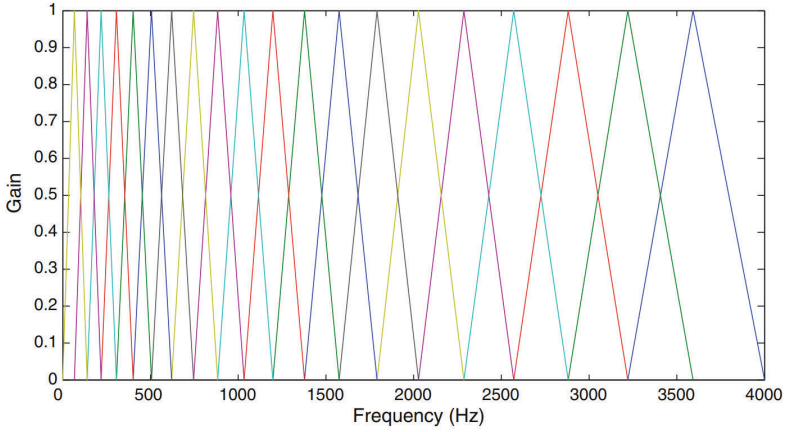


Figure 3.3: *Mel* filter bank

The *Mel* spectrum of the magnitude spectrum $X(k)$ is obtained by the multiplication of magnitude spectrum $X(k)$ and triangular *Mel* weighting filters given in equation 3.4.

$$s(m) \equiv \sum_{k=0}^{n-1} [X(k)]^2 H_m(k) \quad (3.4)$$

Where $0 \leq m \leq M - 1$ where M is triangular *Mel* weighting filters [5, 6]. The calculation of weighted filter given in equation (3.5).

$$H_m(k) = \begin{cases} 0, & = k < f(m-1) \\ \frac{2k-f(m-1)}{f(m)-f(m01)}, & = f(m-1) \leq k \leq f(m) \\ \frac{2(f(m+1)-k)}{f(m+1)-f(m)}, & = f(m) \leq k \leq f(m+1) \\ 0, & = k < f(m+1) \end{cases} \quad (3.5)$$

The weight of k_{th} energy spectrum bin is given by $H_m(k)$ by considering

the m th output band

Discrete Cosine Transform (DCT):

The energy levels in subsequent bands are correlated because of the smoothness of the vocal tract. The DCT is applied to convert the Mel frequency coefficients to cepstral coefficients. Because of the smoothness of vocal track, the adjacent bands in energy levels are correlated. The DCT is applied to get the uncorrelated coefficients by transform the *Mel* frequency coefficients. Before calculating the DCT, The *Mel* spectrum represents linearly in a log scale. As a result, the cepstral domain signal has a quefrequency peak matching to the signal's pitch and a more number of formats representing low quefrequency peaks. So, that most of the speech information is presented in first few MFCC coefficients and ignoring the components of higher orders DCT [10]. In last step, MFCC is computed as [10]

$$c(n) = \sum_{m=0}^{M-1} \log_{10} (s(m)) \cos \left(\frac{\pi n(m - 0.5)}{M} \right) \quad (3.6)$$

where $n = 0, 1, 2, \dots, C - 1$

where $c(n)$ represents cepstral coefficients, and C represents number of MFCCs

Only 8–13 cepstral coefficients are used in traditional MFCC systems.

The zeroth coefficient is usually ignored since it reflects the input signal's average log-energy, which contains very little speaker-specific information.

MFCC feature vectors capture the information from spectral envelope of each frames of speech signal. But it ignores dynamic information. As dynamic information is also important for some speech processing applications including dialect identification. The Delta and Delta-Delta MFCC features are considered. These features are calculated by derivation of original MFCC features.

3.2.1.1 Delta MFCC

Derivation of MFCC feature vectors gives Delta MFCC. These Delta MFCC features help to represent the related Delta features to the change in cepstral features with respect to time. They also represent the change between the frames. They give the temporal information in the speech signal for each frame. As MFCC feature vector size is 13, Delta MFCC size also 13.

3.2.1.2 Delta Delta MFCC

Derivation of Delta MFCC feature vectors gives Delta Delta MFCC feature vector ($\Delta\Delta$ MFCC). They represent the change in the delta features between the frames. They introduce even longer temporal context. They let us know if there is a peak or valley on the look over part of trajectory. As Delta MFCC feature vector size is 13, Delta Delta MFCC size also 13.

The delta MFCC is defined as

$$\Delta_k = f_k - f_{k-1} \quad (3.7)$$

Where f_k represents a feature and k represents time instance. And the delta-delta MFCC features calculated as

$$\Delta\Delta_k = \Delta_k - \Delta_{k-1} \quad (3.8)$$

These features cause increasing the performance and efficiency in extraction of data from speech utterances.

Steps involved in MFCC feature extraction:

1. Initially, speech signal is pre-processed and divided into frames with size of 20ms and overlapping window size is 10ms
2. Magnitude spectrum of each windowed frame of speech is computed using FFT.
3. Mel filter banks is applied on magnitude spectrum of speech signal to find Mel Spectrum.
4. Apply logarithm to Mel frequency in Mel spectrum of each frame.
5. DCT is applied to logarithmic Mel frames in Mel spectrum for get Mel frequency cepstral coefficients.

These DCT coefficients are considered as MFCC features. For each frame, 13 coefficients are extracted and the remaining features are discarded. If derivatives of MFCC are required, extract Δ MFCC, $\Delta\Delta$ MFCC using first and second derivation of MFCC.

Advantages:

1. It estimates the human system response very better than other methods.

2. MFCC features discriminate the spectral behavior of speech signal.
3. It provides good discrimination among the different dialects in a particular language.
4. The coefficients obtained have less correlation in cepstral coefficient.
5. It is helpful to get the important phonetic characters of each frame within the speech signal.

3.3 Different Statistical Models

The performance of any pattern recognition system including dialect identification system depends on the selection of training model which discriminate and study the features of input data. The training phase of any statistical model is used in order to get the behaviour and abstract representation of inputs. There are several statistical methods like Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), and Deep Neural Network (DNN) methods[7].

3.3.1 Hidden Markov Model (HMM)

The HMM is a popularly and mostly used in the design of speech processing applications as it is rich in effective mathematical structures and effectively describes and captures the sequence feature vectors of speech frames [26]. The Hidden Markov Model (HMM) is a probabilistic model that is used to find the hidden states by using a set of observed variables. It depends on Markov chain property in which the future possible event completely depends on the current possible event, but

not on the previous event. It is represented by $P(W_t|W_t - 1)$ means that event W_t is depends upon $W_t - 1$, not on $W_t - 2$. The basic model of HMM structure is shown Fig. 3.4.

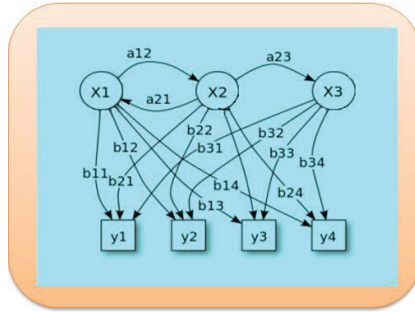


Figure 3.4: Basic structure of HMM model

The HMM model is described by different parameters namely states hidden and observable states, probabilities of Initial, Transition and Emission. The HMM, can extract the compact information from the input sequence and it describes, the behaviour of speech features well. From the Fig.3.4, the parameters are

$X = X_1, X_2, \dots, X_N$ are number of hidden states

$Y_i = Y_1, Y_2, \dots, Y_T$ are sequence of observable

$S = a_{11}, a_{12}, \dots, a_{mn}$ is transition probability matrix and $\sum_{i=1}^n a_{ji} = 1$

$E = b_{11}, b_{12}, \dots, b_{mn}$ are the emission probabilities.

$\Pi = \pi_1, \pi_2, \dots, \pi_n$ represents the initial probabilities.

The joint probability is used to, find the state of the event, The equation 3.9 represents the joint probability of two states.

$$P(Y, X) = P(Y|X)P(X) = P(x_i|z_i) \times P(z_i|z_{i-1}) \quad (3.9)$$

Where N is, number of observations.

Hidden Markov models are associated with three problems

1. Evaluation Problem of HMM which is used in testing phase of any pattern recognition application. It estimates the probability of observation sequence (O) against given Hidden Markov Model.
2. Decoding Problem of HMM finds the sequence of states produced for given observation sequence (O=O1,O2,... OT) and Hidden markov model (λ).
3. Learning Problem of HMM is used in training phase of Pattern recognition system including Speech Processing model and dialect identification system. It creates a reference model by using given a HMM model λ and sequence of inputs.

In this phase, compact representation of feature vectors is achieved by adjusting parameters of HMM (A, B, Π) in order to achieve maximum $p(O|\lambda)$. To adjust the model parameter, EM (Expectation-Maximization) algorithm is used [74].

In our work, MFCC feature vectors are extracted from Telugu speech utterances of different dialects and HMM is design using problem (3) as

specified above.

In testing phase of the system, MFCC feature vectors are extracted from unknown utterances of Telugu speech and evaluated against HMM of three dialects of observation sequence is calculated using evaluation problem of HMM.

Advantages:

1. HMM is an efficient learning algorithm means learning happen directly from raw data itself.
2. It is very flexible means it can handle input sequence of any length.
3. It has wide variety of applications including pattern discovery.
4. It has a good mathematical framework that provides a straight forward solution to related problems and structured frame work.

Disadvantages:

1. In general the successive observations are very often independent to each other, but HMM assumes that they are independent.
2. The amount of data that is necessary to train HMM is very high.
3. The no. of parameters that are required to establish HMM are also huge.
4. The trial and error methods are used for selecting the model topology.

3.3.1.1 Gaussian Mixture Model:

Gaussian Mixture Model is more famous to design the speech processing model. The GMM model contains certain Gaussian distributions where data points belong to any one of distribution [11]. The GMM model is adjusting to elliptic shape and it is a soft clustering technique. An Expectation-Maximization (EM) technique [75] is used in the GMM model to set the parameters. It is mainly used when there are latent variables in data. The probability density function(s) for a single variable is given in equation 3.10

$$M(d, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(d-\mu)^2}{2\sigma^2}} \quad (3.10)$$

When there are several observations then, the joint probability is given in equation 3.11.

$$P(d|\mu, \sigma^2) = \prod_{j=1}^n \sum_{k=1}^s \pi_s M(d_j, \mu_s, \sigma_s^2) \quad (3.11)$$

Where μ is mean, σ^2 is covariance, s is number of distributions and d is the observation.

EM algorithm is a two step process which is used to train the GMM model

1. **E-step:** Used to find the probability that each point d_i belongs, distributions C_1, C_2, \dots, C_s . It is represented given equation 3.12

$$r_{jc} = \frac{\pi_c M(d_j|\mu_c, \sigma_c)}{\sum_{i \in (0,s)} \pi_i M(d_i|\mu_i, \sigma_i)} \quad (3.12)$$

Where r is the calculation of responsibility.

2. **M-step** By considering the results of E-step update the μ, σ, π values. This can be calculated as like shown in equation 3.13

$$\mu_c^{New} = \frac{1}{T_c} \sum_j r_{jc} d_i \quad (3.13)$$

Where T_c is the number of points in cluster c ,

$$New \left(\sigma_c^2 \right) = \frac{1}{T_c} \sum_j r_{jc} \left(d_j - \mu_c^{New} \right)^t \left(d_j - \mu_c^{New} \right) \quad (3.14)$$

and the density function π is modified every time is given in equation 3.15

$$\pi_c = S_c/S \quad (3.15)$$

Where S_c is number of observations in cluster c and S is the total number of observations.

The number of Gaussian mixtures considered for GMM impacts the performance system. If the no. of Gaussian mixtures increases the accuracy will also get improved. GMM produces cluster which are non-convex and these are controlled with the variance of Gaussian distribution. In our work, MFCC feature vectors are extracted from Telugu speech utterances of different dialects and GMM model is trained. In testing phase of the system, MFCC feature vectors are extracted from unknown utterances of Telugu speech and evaluated against GMM of three dialects. Using the EM algorithm maximum probability is calculated.

Advantages

1. The training through this model is very fast.
2. We can easily update it with new data.

Disadvantages

1. The performance of GMM will vary if the signal is full of noise.
2. It requires huge amounts of data to estimate the parameters.

3.3.1.2 Deep Neural Networks (DNN)

Deep Neural Networks (DNN) are used to achieve the state of art performance in several fields such as speech recognition, speaker identification etc. The DNN can be considered as a branch of Machine learning where high level functions are retrieved from the input information and that can be incorporating several layers of nodes in the system. Through this mode creative and abstract component can be extracted from the input data.

The DNN is more popular because it is capable of extracting even complex features from large amount of data even from hours of speech data both Linear as well as non-linear because of their deeper architecture as shown in Fig.3.5.

In order accomplish the specific task, layers of data that exist in between the input and output must be processed by the system. A network is said to be deeper if there are more number of layers to be processed in order to obtain the result. In order to assess the number of layers that are required for the completion of task by the system, "Credit Assignment Path (CAP)" is popularly used and it is effective.

If the CAP index is greater than two, then we define the neural net-

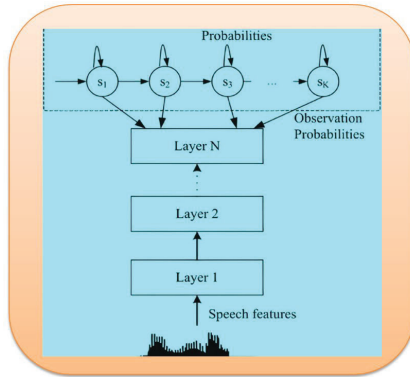


Figure 3.5: Basic structure of DNN model

work as deep. Autonomous work with high efficiency can be executed by a deep neural network without any human intervention. DNN is used in various applications like Artificial Intelligence, robot, image identification, AI cars, Speech processing etc. In order to define a network it mainly consists of building blocks such as neurone, layers, weights, input, output an activation function and learning mechanism (Optimizer), which plays a vital role in updating the weights. The Fig 3.6 represents the basic structure of DNN.

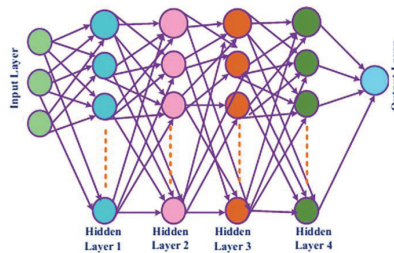


Figure 3.6: Basic DNN Structure contain N-hidden layer.

The neurons present in the first layer (not hidden) will take the input

data and its output will become as input to the neurons present in successive layer and so on thus it gives the final output as represented in Fig.3.6. The produced final output will be like yes or no which in turn represented as probability. Every layer may contain one or more neurons and uses a function on these neurons called as “activation function” This function normalize the output produced by neuron and sends the signal to next layer as input to the further connected neurons as shown in Fig 3.7.

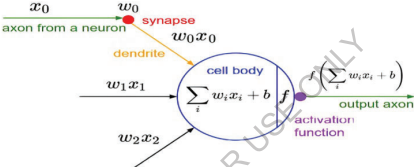


Figure 3.7: Neuron structure

The output of cell could be a weighted total of its inputs and bias terms, which is given to activation function F . The output of node in the n^{th} layer can be represented in equation 3.16.

$$y_{n,j} = F\left(\sum_i \left(y_{n-1,i} \times w_{n,i,j}\right) + b_{n,j}\right) \tag{3.16}$$

The output is further passed only if the results value of incoming neurons is more than the threshold. A weight will be assigned to the connection that exists between the neurons of successive layers. This weight defines the importance of the input on the output for the next neurone as a result for the overall final output. The weights are initialized to random value and the weights will get modified by back propa-

gation, in each iteration in the training phase of DNN. The input layers use the ReLU (Redressed straight unit) activation function. Scientifically, it is characterized as $y = \max(0, u)$. The working of ReLU activation function is shown in Fig.3.8.

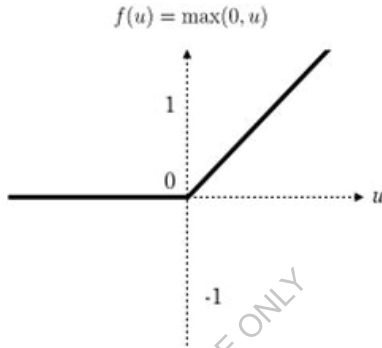


Figure 3.8: ReLU activation function

The activation function used in output layer is Softmax activation function and it is given in equation 3.17.

$$F(z_j) = \frac{e^{z_j}}{\sum_j (e^{z_j})} \quad (3.17)$$

Where z is input vector of j real numbers and e is constant.

Table 3.1: Parameters used in Training DNN

Parameters	Value
Total hidden layers	2
Number of neurons in each layer	Hidden1 = 30, Hidden2 = 12
Input Size	39 (MFCC + Δ MFCC + $\Delta\Delta$ MFCC)
Cross validation	True
Technology (to train)	Tensor flow

Advantages:

1. It can learn complex features as it is capable to learn with large input data.
2. It doesn't require any knowledge in advance.
3. It is well suitable for pattern recognition as a result it can generalise about the data.
4. It helps to obtain less word error rate.

Disadvantages:

1. It is not suitable when the data is small in size.
2. The computational cost is very high.
3. It is completely black box in nature.

3.4 Proposed Methodology to Identify the Dialects of Telugu Language

Dialect Identification system consists of Training phase and Testing phase including feature extraction.

3.4.1 Hidden Markov Model (HMM) based Dialect Identification System

3.4.1.1 Feature Extraction

In this work, 13-dimensional MFCC features, 13-dimensional Delta MFCC and 13-dimensional Delta Delta MFCC feature vectors are derived from windowed speech signals in Training and Testing phase and concatenated to form 39-dimensional feature vectors. These feature vectors are derived as specified in section 3.2.

3.4.1.2 Training of HMM for Dialect Identification

The Training phase of dialect identification system is a two-step procedure.

In first step, 39-dimensional MFCC features of speech samples of different regional dialects of Telugu language are derived as explained in section 3.2.1. These feature vectors are given as input sequence for HMM model and dialect specific HMM model is designed for each dialect. Training phase of HMM is done with the EM algorithm [8]. In this phase, dialect specific HMM is created that is one HMM for one dialect.

The model designed for each dialect as shown in the Fig.3.9. In this work, three HMM are created for three input sequence of Rayalaseema, Telangana, and Costa Andhra dialects of Telugu Language respectively.



Figure 3.9: Training phase of HMM based dialect identification

3.4.1.3 Testing phase of HMM for Dialect Identification

In testing phase, 39-dimensional feature extraction of unknown utterance of Telugu speech is derived as like in Training phase. This testing phase is also called as identification phase. These feature vectors are treated as observed sequence for HMM. These observation sequences of input speech are evaluated against three dialects specific HMMs in order to get likelihood score against each HMM. A dialect specific HMM which gives maximum likelihood score is considered as dialect of unknown utterance. The testing phase of HMM as shown in Fig.3.10.

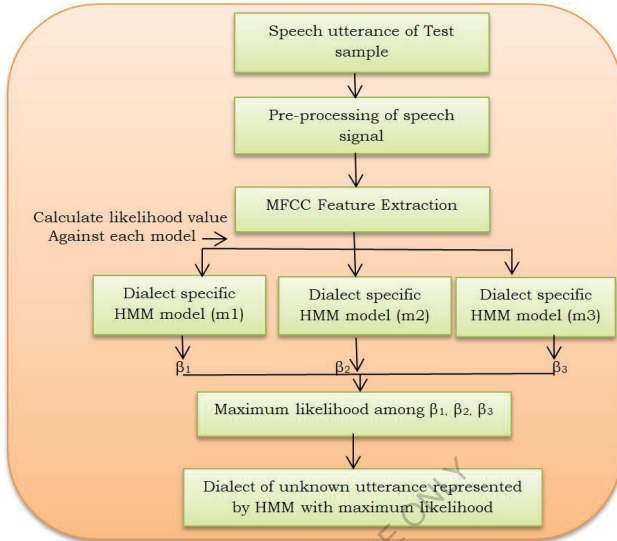


Figure 3.10: Testing phase of HMM based dialect identification

In this work, the feature vectors of unknown Telugu speech are extracted and evaluated against three dialect-specific HMMs.

3.4.2 GMM based Dialect Identification System

GMM based dialect identification system consists also two phases i.e., Training phase and Testing phase. In Training phase, dialect specific GMM is created. In the testing phase, the feature vectors of unknown speech are evaluated against dialect-specific GMM in order to identify the dialect of unknown speech. In Training and Testing phase, we considered 39-dimensional feature vectors of MFCC, Δ MFCC and $\Delta\Delta$ MFCC features.

3.4.2.1 Training phase of GMM:

Like HMM based dialect identification, In training phase of GMM, dialect specific GMM is designed i.e., one GMM for one specific dialect using Expectation Maximization (EM) algorithm[8] for Training speech of three dialects of Telugu language. The model designed for each dialect as shown in the Fig.3.11.



Figure 3.11: Training phase of GMM based dialect identification

3.4.2.2 Testing phase of GMM:

The Testing phase of GMM based Telugu dialect identification system gives the dialect of unknown utterance of Telugu speech. It involves 39-dimesional MFCC feature extraction from unknown utterance and evaluating against each GMM of three dialects of Telugu language. Like HMM based, GMM based Telugu dialect identification system gives the dialect of unknown utterance of speech signal based on the maximum likelihood. The general frame work of testing phase as shown Fig.3.12.

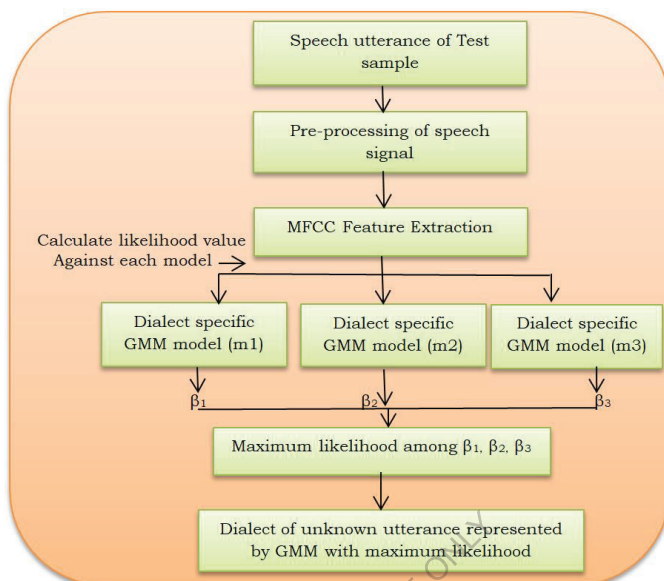


Figure 3.12: Testing phase of GMM based dialect identification

3.4.3 DNN based Dialect Identification System

Like GMM and HMM models, DNN based Telugu dialect identification system consists of two phases training phase and the testing phase. In the training phase, creates a reference model for input data and the testing phase reveals the dialect of input data of unknown speech.

3.4.3.1 Training phase of DNN

Like HMM based dialect identification, In training phase of DNN, dialect specific DNN is designed i.e., one DNN for one specific dialect using Stochastic gradient descent (SGD) and activation functions for Training speech of three dialects of Telugu language. The model designed for each dialect as shown in the Fig.3.13. In this work, MFCC feature vectors are extracted from Telugu speech utterances of different dialects of

huge speech training samples from frame and utterance level. These features are given as input to DNN model for learning. By applying the back propagation method, update the weights and learning rate in order to get expected value using error rate. The stochastic gradient descent (SGD) method is used to update the weights of the networks in order to reduce the error rate. It will repeat the process until the model is stabilized. In this, the activation function in output layer Softmax is used and it is used to normalize the output value given by model to 1.

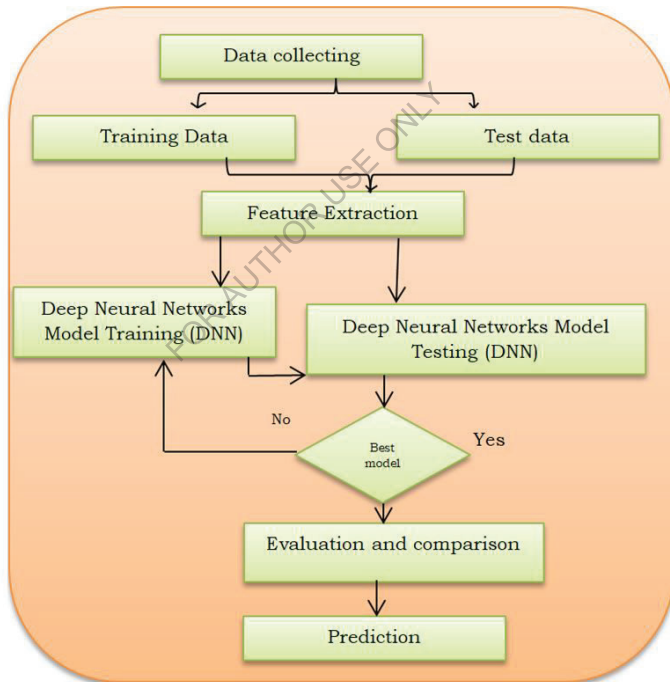


Figure 3.13: DNN based dialect identification

3.4.3.2 Testing phase of DNN

The testing phase of DNN based dialect identification system gives the dialect of unknown utterance of Telugu speech utterance of unknown speech. It involves 39-dimensional MFCC + Δ MFCC + $\Delta\Delta$ MFCC feature extraction from unknown utterance from frame level and utterance level and evaluation against each DNN of three dialects of Telugu language. Like HMM based, DNN based dialect identification system gives the dialect of unknown utterance of speech signal based on the maximum likelihood.

3.5 Results of HMM, GMM and DNN for Dialect Identification

In this work, HMM, GMM and DNN based dialect identification system are implemented using MFCC feature vectors. MFCC feature vectors discriminate the features of dialects in Telugu language effectively. As these dialects are various in the frequency of spectral properties of speech signal [25].

In this work, the experiments for identify dialects are carried out on data sets of three dialects of the Telugu language. The creation of the database is clearly explained in section 1.5 of chapter 1.

The Telangana dialect speech corpus duration is 2h 35 min, out of which 1h 55 min is Training speech samples and 40min is test speech samples which includes 361 test samples of 3-8sec speech utterances.

The Costa Andhra dialect speech corpus length is 2h 47min, out of which 2h 10 min is Training speech samples and 40min is test speech samples which includes 355 test samples of 3-8sec speech utterances.

The Rayalaseema dialect speech corpus length is 1h 43 min, out of which 1h 10 min is Training speech samples and 33 min is test speech samples which includes 287 test samples of 3-8sec speech utterances. The speech sample details are specified in Table.3.2.

Table 3.2: Telugu dialect database details

Dialect	Total time of speech data	Speakers for each dialect	Period of each test sample	Age of speakers	Sampling Frequency
Telangana	2h 35 min	80	3-8s	9-50	44,100Hz
Costa Andhra	2h 47 min	75	3-8s	9-50	44,100Hz
Rayalaseema	1h 43 min	75	3-8s	9-50	44,100Hz

In order to design a dialect identification system of Telugu language, initially, 13-dimensional MFCC feature vectors only are extracted from speech utterances and experiments are carried out using HMM, GMM and DNN. Then 39-dimensional feature vectors (MFCC + Δ MFCC + $\Delta\Delta$ MFCC) are extracted from the training speech of each dialect of the Telugu language and used these feature vectors as the input sequence. The HMM-based dialect identification system is designed with three states for the input sequence of 13-dimensional (MFCC) and 39-dimensional feature vectors (MFCC + Δ MFCC + $\Delta\Delta$ MFCC) and achieved the perfor-

mance is 78.6%, 82.6% respectively. GMM based dialect identification is designed with the same input sequence 13-dimensional (MFCC) as well as 39-dimensional feature vectors (MFCC + Δ MFCC + $\Delta\Delta$ MFCC) with 32 Gaussian mixtures and achieved the performance is 79.2%, 82.6% respectively. DNN based dialect identification is designed with 4 layers (2 hidden layers one Input and one output layer) and the performance achieved is 80.4%, 84.6% for 13-dimensional (MFCC) and 39-dimensional feature vectors (MFCC + Δ MFCC + $\Delta\Delta$ MFCC) respectively. The detailed performance of Telugu dialect identification system with each different models are shown in Table 3.3.

Table 3.3: Performance of Dialect Identification using MFCC + Δ MFCC + $\Delta\Delta$ MFCC

Feature Extraction	Model	Accuracy of Model		
		Telangana	Costa Andhra	Rayalaseema
MFCC	GMM	81.2	80.4	77.8
	HMM	80.4	79.5	76.3
	DNN	83.4	80.7	78.7
MFCC + Δ MFCC + $\Delta\Delta$ MFCC	GMM	85.3	82.6	79.9
	HMM	84.4	80.2	83.2
	DNN	85.1	84	84.6

The performance of Dialect Identification achieved with HMM, GMM and DNN using 13-dimensional feature vectors is specified in Fig.3.14.

The performance of Dialect Identification achieved with HMM, GMM and DNN using 39-dimensional feature vectors is specified in Fig.3.15.

It is observed from Fig.3.14 and Fig.3.15, 39-dimensional MFCC + Δ MFCC + $\Delta\Delta$ MFCC performed well in identification of Telugu Dialects.

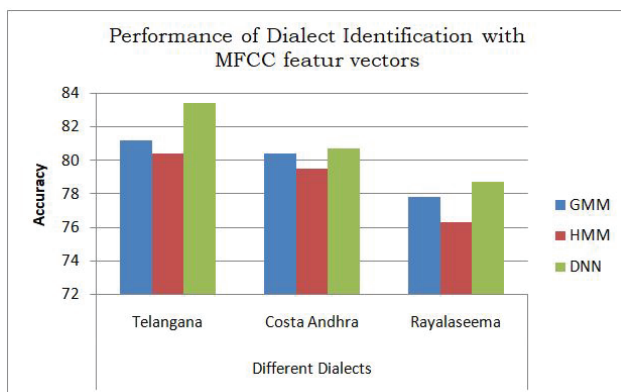


Figure 3.14: Performance of Dialect Identification with different models using MFCC

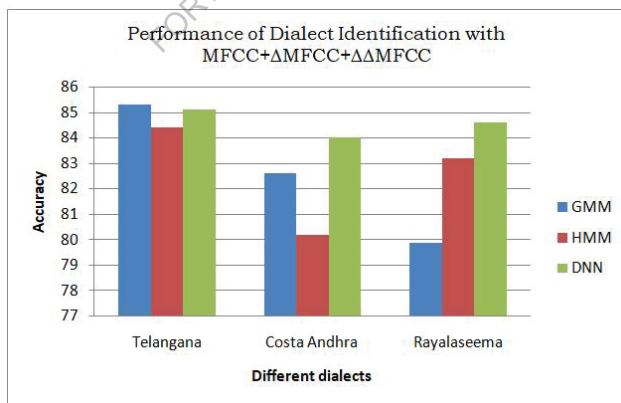


Figure 3.15: Performance of Dialect Identification with different models using MFCC + ΔMFCC + ΔΔMFCC

It is also observed that, DNN model gave the impressive performance with 84.6% accuracy, as this model learnt the discriminate features among the dialects of Telugu Language.

3.6 Conclusion

In this chapter, GMM, HMM and DNN based dialect identification system for Telugu language have been designed using spectral features (MFCC + Δ MFCC + $\Delta\Delta$ MFCC). A Telugu database with 7h 05min, out of which 2h 35min Telangana dialect, 2h 47 min Costa Andhra and 1h 43min Rayalaseema was created. In these models, 39-dimensional (13-MFCC, 13- Δ MFCC and 13- $\Delta\Delta$ MFCC) feature vectors of three dialects are derived and dialect specific models are designed in training phase. In second phase, MFCC features vectors of test speech samples are derived and give the result based on maximum likelihood. It is observed that the performance of Dialect Identification is 82.6% for HMM, 82.6% for GMM and 84.6% for DNN with MFCC + Δ MFCC + $\Delta\Delta$ MFCC features. MFCC feature vectors (MFCC + Δ MFCC + $\Delta\Delta$ MFCC) discriminate the spectral properties of dialect well but not in case of nasal sounds.

Chapter 4

PROSODIC FEATURE EXTRACTION TECHNIQUES TO IDENTIFY DIALECTS OF TELUGU LANGUAGE

4.1 Introduction

The Telugu language is an ancient language and it belongs to Dravidian family contains three dialects namely Rayalaseema, Telangana, Coastal Andhra as specified in chapter.1. These dialects are different from each other with respect to acoustic features and prosodic features of speech signal. They show significant variance with respect to prosodic features like pitch, intensity and energy etc. Prosodic feature extraction technique is one of the techniques which follows temporal feature extraction methodology [60]. In temporal feature extraction technique, the features are extracted directly from the input speech signal whereas Prosodic features of any audio changes from language to language and from region to region [70]. When the features are extracted using MFCC and its derivative features, drawback is that the particular word is spoken same in different region with different stress, intonation, rhythm, it is difficult to identify the dialect. This problem can be solved by using prosodic features because Prosodic features are extracted the features from supra segment level. The challenge of research is to improve the accuracy of identification with shortest duration of utterance. In this in order to increase the accuracy of model, new feature vectors are derived

by combining different prosodic features.

Prosodic features describe the features of speech signal with respect to vocal tract vibration, nasal sound, intonation and loudness etc., [70] [72]. These features discriminate the dialects with different pronunciation and accent. In the state of art systems, different prosodic features are derived from raw speech signals using supra segmental rather than phoneme and syllable level.

In this chapter, the importance of prosodic features has been established in the discrimination of similar utterances with different regional dialects of Telugu language. There are different prosodic acoustic features like pitch, loudness, Energy, Formants which are popular in speech processing applications like language identification, speaker identification, dialect identification [60].

In this Chapter, we explored the extraction of prosodic features: Pitch, Energy, Loudness, and Formants from framing and windowed speech and analyse the differences and effects of these features in three dialects of Telugu language. With these feature vectors, the performance of dialect identification system is evaluated using K-NN model by implementing Telugu dialect identification system for shortest duration of test samples.

In this chapter, section 4.2 describes the various prosodic features useful for dialect identification and its extraction from raw speech sig-

nal and also their variance in three dialects of Telugu language. The Statistical model K-NN model has been described in section 4.3 to implement dialect identification system.

Section 4.4 represents the proposed model for Telugu dialect identification system with prosodic features. The results are presented in section 4.5 and finally conclusions have been drawn in section 4.6.

4.2 Prosodic Features for Dialect Identification

The prosodic features provide the important cues to clearly discriminate the dialects of Language. The different prosodic features are extracted from frames of speech utterances to identify the Dialect Identification of Telugu Language. The prosodic features like Pitch, Intensity, and Energy etc. are considered and also to increase the performance of the system by using different combinations of prosodic features in Dialect Identification.

4.2.1 Pitch

Pitch is defined as a highness or lowness of a speech with respect to frequency of vibration [60]. Pitch is proportional to the energy of speech signal.

$$Pitch \propto Frequency$$

If the frequency of audio is high, then the audio has high pitch.

The frequency is inversely proportional to time period of speech utterance.

Table 4.1: Different pronunciations of same word corresponding to each dialect

Telangana	Costa Andhra	Rayalaseema
మట్టి	మట్టి	మట్టి

$$\text{Frequency} \propto 1/\text{Timeperiod}$$

The audios that take more time have the low pitch and audios that take less time has the high pitch. The relationship between frequency and time period depicted in the Fig.4.1.

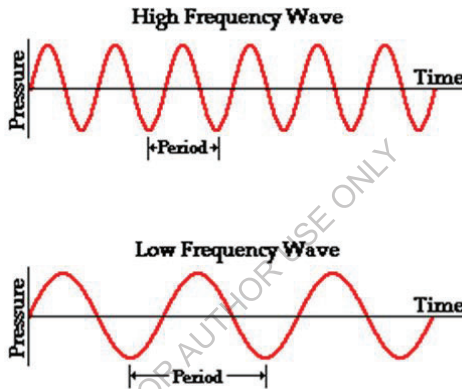


Figure 4.1: High and Low frequency

In this work, we analyze the pitch of the speech utterance of three different dialects: Telangana, Rayalaseema, Coastal Andhra. It was observed that Telangana has high pitch and Andhra has low pitch and Rayalaseema has medium pitch. For example, considered following example in Table.4.1 for pitch comparison for the word మట్టి

Pitch of speech signal varies between speech utterance of different persons belong to different regions. But there are some words which are similar to most of the dialects as shown in Table.4.1. These words are

similar and it is difficult to differentiate by using spectral feature extraction. To discriminate clearly prosodic features are used because if the words are similar but their vocal track vibration was different. In this context, Pitch plays important rule to differentiate the similar words according to the respective dialects. The Fig.4.2, 4.3 and 4.4 represent the prosodic features including Pitch of speech utterance of different persons from different regions Rayalaseema, Costa Andhra and Rayalaseema respectively.

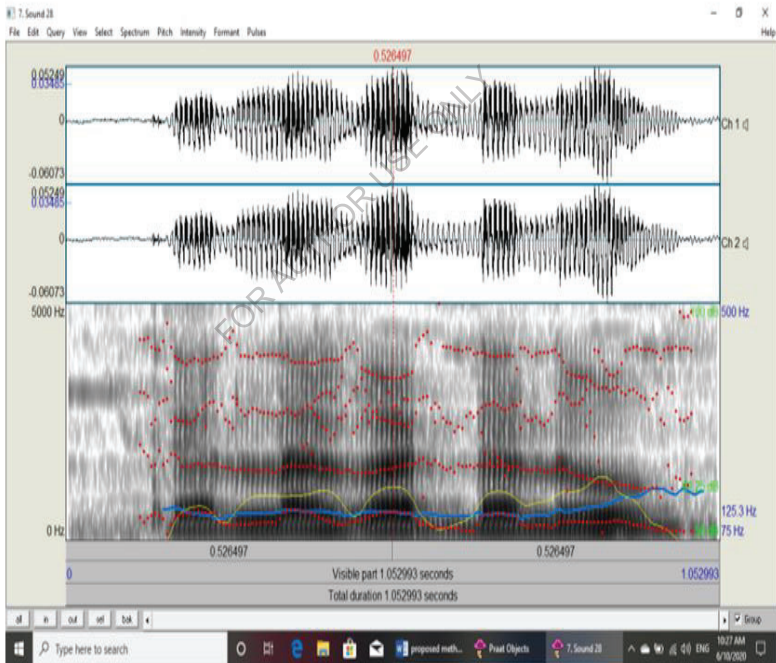


Figure 4.2: Different prosodic features of Rayalaseema dialect

In the above three figures the pitch of audios is represented in blue colour line.



Fig.4.3. Different prosodic features of Andhra dialect

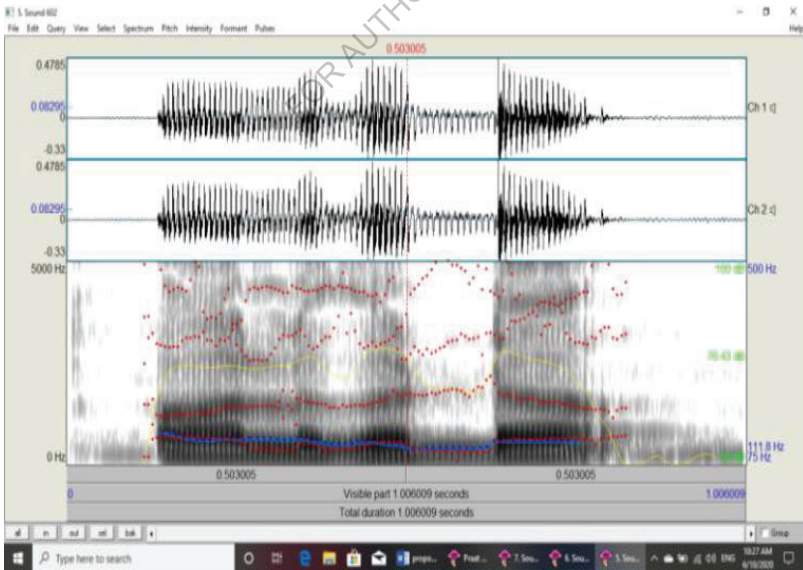


Fig.4.4. Different prosodic features of Telangana dialect

4.2.2 Intensity

Intensity of speech frame is defined as power carried by the audio waves per unit area which is perpendicular to the cross sectional area.

Intensity of speech signal defined in equation.4.1

$$I = 2 \Pi v^2 \delta^2 \rho c \quad (4.1)$$

where, I = Intensity of audio

v = frequency of audio

δ = amplitude of sound waves

ρ = density of sound waves

c = speed of sound waves

From the equation.4.1, Intensity is directly proportional to Frequency. The speech signal that has high pitch will high intensity also.

In this work, experiments are carried to analyze intensity of speech signal based on different dialects of Telugu Language and also draw the relationship between pitch and Intensity.

From the experiments, it is observed that Telangana dialects have more Intensity and An dhra has low Intensity. It is also observed that, if the words are similar in all dialects, then we used the intensity to differentiate the dialects. The Fig.4.2, 4.3 and 4.4 show the intensity of three dialects of Telugu Language in yellow colour lines.

4.2.3 Energy

Energy of the speech is produced when the vocal track of human being vibrates during speaking [60]. The sound vibrations cause wave of pressure that travel through certain medium. Sound energy is a form of mechanical energy as shown in equation.4.2 and equation.4.3.

$$E = \sum_m^{\infty} = -\infty s^2(m) \quad (4.2)$$

$$E = \sum_m^n = {}_{n-N+1} s^2(m) = s^2(n - N + 1) + \dots + s^2(n) \quad (4.3)$$

Where E is energy of speech signal $s(m)$.

The energy of a wave is directly proportional to Pitch of speech utterance. There are similar words also occurred in different dialects, in order to discriminate those words energy plays vital role as it completely depends upon vocal track. The experiments are done to analyze energy of speech signal and also draw the relationship between pitch and energy. From these experiments, it is observed that Telangana has high energy and Andhra has low energy.

4.2.4 Formants

A formant is defined as an acoustic energy that is present around a particular frequency of speech signal [33]. There are different formants present for a speech signal f_1 to f_4 . Each of the formants represents at different frequency levels. Formant corresponds to resonance of the vocal track.

In this work, speech utterance of three dialects of Telugu Language Telangana, Rayalaseema, and Coastal Andhra are analyzed based on the formants. The formant diagrams for the speech signal from different regions are as shown in Fig.4.2, 4.3 and 4.4 in red colour lines. Formants in speech signals are represented where the frequency of audio is high and which have high degree of energy. To increase the accuracy, Instead of comparing only single feature of speech signals sometimes required adding two or more features in order to clearly differentiate the dialects and get best results because some of the words are similar in more than one dialect shown in Table.4.2.

Table 4.2: Sample of words which have similar way of pronunciation

TELANGANA	COSTAL ANDHRA	RAYALASEEMA
దీపం	దీపం	దీపం
సంతోషం	సంతోషం	సంతోషం
పని	పని	పని
తెలివి	తెలివి	తెలివి
వారం	వారం	వారం

4.2.5 Loudness

To identify the dialect of speech from raw signal, Loudness is also play important role in order to discriminate different dialects of Telugu language. Loudness of speech signal is proportional to amplitude of that speech signal in which speech signal with more amplitude has more loudness. The amplitude of speech signal is depicted in Fig. 4.5.

The relation between loudness and amplitude is represented in equa-

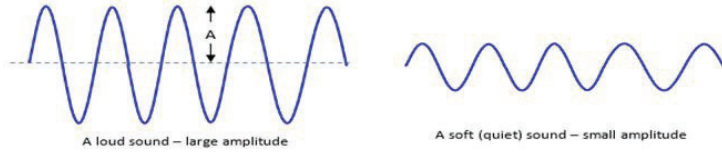


Figure 4.5: Amplitude of speech signal

tion 4.4

$$\text{Loudness} = k (\text{Amplitude})^2 \quad (4.4)$$

where k is constant value. Similarly, the loudness of speech signal is proportional to speech intensity and energy in the signal. That means a signal with more intensity has more loudness. Relation between loudness and intensity is given in equation 4.4.

$$\text{Loudness} = \log (\text{intensity}) \quad (4.5)$$

The three dialogues of Telugu language has different loudness in their speech utterances. So that it is an important cue to discriminate based on the accent.

In order to discriminate the words belongs to which region, it is proposed to combine the different prosodic features. This combining of two or more features of speech signals increase the performance of models and accurately differentiate the dialects of language [41]. The hybrid features considered in our work are Pitch + Intensity, Pitch + Intensity + Energy and Pitch + Intensity + Formant (f1). In this work, we have extracted different prosodic features like Pitch, Intensity, Loud-

ness, Energy, Formants, average of different combination of prosodic features are extracted from frames and windowed speech of three dialects (Telangana, Costa Andhra and Rayalaseema) of Telugu Language and presented in Tables 4.3, 4.4 and 4.5 respectively.

Table 4.3: Different prosodic values of Costa Andhra Speech samples

Pitch (1)	Intensity (2)	Energy (3)	Loudness (4)	(1)+(2)	(1)+(2)+(3)	(1)+(4)	(1)+(2)+F1
119.0	69.55	0.0035	69.72	94.28	47.145	92.178	228.6
123.3	68.02	0.0030	59.16	95.70	47.855	150.35	202.6
124.8	70.36	0.0017	56.74	97.61	48.807	89.152	227.3
124.8	70.36	0.0056	61.42	97.61	48.809	140.12	216.9
124.7	68.70	0.0035	52.99	96.71	48.356	151.35	212.3
121.2	68.04	0.0030	62.61	94.62	47.315	98.349	235.1
116.6	68.32	0.0030	55.18	92.48	46.243	101.46	192.2
117.9	66.03	0.0018	70.24	91.97	45.98	109.46	312.5
122.9	67.34	0.0025	53.96	95.14	47.57	119.46	241.6

The prosodic features of some speech utterances are shown in Table 4.3, 4.4 and 4.5 for Telangana, Costa Andhra and Rayalaseema. The hybrid features by combining the different prosodic features also considered for the dialect identification of Telugu Language. It is observed that Pitch, Intensity, Energy values of the speech signal of different regions are higher for the Telangana region and lower for the Andhra region, and Rayalaseema, it falls in between Telangana and Costa Andhra regions.

Table 4.4: Different prosodic values of Telangana Speech samples

Pitch (1)	Inten- sity (2)	Energy (3)	Loud- ness (4)	(1)+(2)	(1)+(2)+ (3)	(1)+(4)	(1)+(2)+ F1
248.3	70.37	69.12	159.35	0.021	151.4	79.68	325.7
334.6	70.14	61.35	202.37	0.013	184.5	101.1	346.1
244.2	61.04	51.29	152.64	0.002	141.4	76.32	398.4
310.4	66.57	58.20	188.50	0.006	190.5	94.25	340.4
116.0	71.46	72.17	93.74	0.14	98.57	46.94	253.1
259.2	70.38	68.35	164.80	0.016	142.3	82.41	308.5
208.4	78.94	76.18	143.69	0.008	155.3	71.85	281.3
212.8	72.33	69.34	142.61	0.123	132.3	71.36	321.1
174.0	80.55	55.12	127.28	0.015	121.4	63.65	294.8

Table 4.5: Different prosodic values of Rayalaseema Speech samples

Pitch (1)	Inten- sity (2)	Energy (3)	Loud- ness (4)	(1)+(2)	(1)+(2)+ (3)	(1)+(4)	(1)+(2)+ F1
149.2	72.63	52.18	110.96	0.006	93.38	55.48	265.6
286.9	64.92	49.36	175.94	0.003	91.23	87.97	321.3
169.5	65.45	32.46	117.52	0.001	89.94	58.76	296.0
191.4	68.66	56.36	130.045	0.009	93.23	65.02	288.9
277.0	62.56	49.81	169.81	0.008	90.12	84.90	347.6
141.9	67.26	67.47	104.62	0.007	90.47	52.31	249.0
125.5	79.59	70.15	102.57	0.123	88.34	51.35	266.8
140.4	67.81	50.12	104.12	0.001	91.36	52.06	272.9
126.0	69.19	56.18	97.60	0.001	80.23	48.80	288.4

4.3 Statistical Models used for Dialect Identification

In this work, a statistical model K-Nearest neighbour algorithm used to identify the dialect of Telugu Language using different features.

4.3.1 K-Nearest Neighbor (K-NN) Algorithm

The $K - NN$ model is a supervised statistical method that achieves high performance without any prior assumptions about the trained data. It is also called as lazy learner algorithm.

The $K - NN$ algorithm classifies the test samples into pre-defined class labels of trained data. Here K —represents the number of nearest neighbours to be considered for classifies test samples.

If $K = 1$ i.e., number of nearest neighbours is one class. In this case the class label of test sample is decided by the distance between test sample and classes of data. Whatever the class gives the minimum distance that class label is given to test sample.

The K-NN algorithm classifies the dialects of Telugu language into Andhra, Telangana, and Rayalaseema. In order to classify the test speech signal, extract the prosodic feature and calculate the statistical value corresponding to features. The distance between mean of the test sample prosodic feature and mean of prosodic feature of each dialect found in trained data is calculated using Manhattan distance. The Manhattan distance for one-dimensional data is calculated is given in

equation.4.6.

$$D = \sum_{i=1}^n |X_i - Y_i| \quad (4.6)$$

The basic steps of K-NN algorithm:

1. Find the distance of test label and classes of data
2. Finding closest neighbour
3. Assign class label to test sample.

The basic working of K-NN model is shown in Fig.4.6 and 4.7

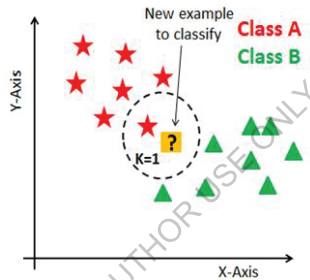


Figure 4.6: $K - NN$ model when $K = 1$ nearest neighbour

When the K value is more than 1, working of $K - NN$ model as shown Fig.4.7

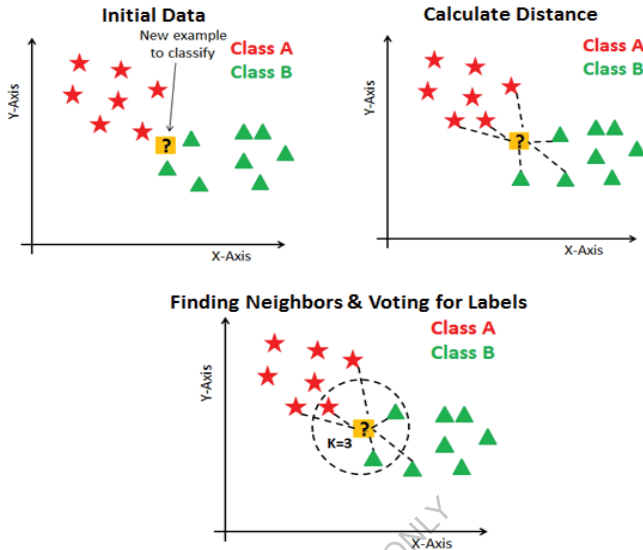


Fig.4.7 Working of K-NN model with K=3 nearest neighbours

4.3.1.1 Working of K-NN algorithm

Step-1: Choose the number of neighbours for test sample i.e. K

Step-2: Input data of Training samples is given along with the class labels.

Step-3: Give the input of test sample to K-NN model and calculate distance between test sample and training sample using any distance calculation method like Manhattan, Euclidian or Supreme distances

Step-4: Based on K- value, find the nearest neighbour of test samples.

Step-5: Classify the test samples into a class of train sample based on majority of nearest neighbours of same class.

4.4 Dialect Identification Using Prosodic Feature

As explained in previous chapter, Dialect Identification consists of two phases Training phase and Testing phase in which common part is feature vector extraction.

4.4.1 Training Phase

In the first phase of Dialect Identification, Prosodic feature vectors are extracted from speech utterances of Telugu language as explained previous sections using PRAAT tool. For these extracted dialect specific prosodic features, find the mean of all prosodic features of specific dialects. Such that one mean value of all the feature vector for one dialect. Using these three mean values K-NN model is designed. The detailed description of Training phase is shown in Fig. 4.8.

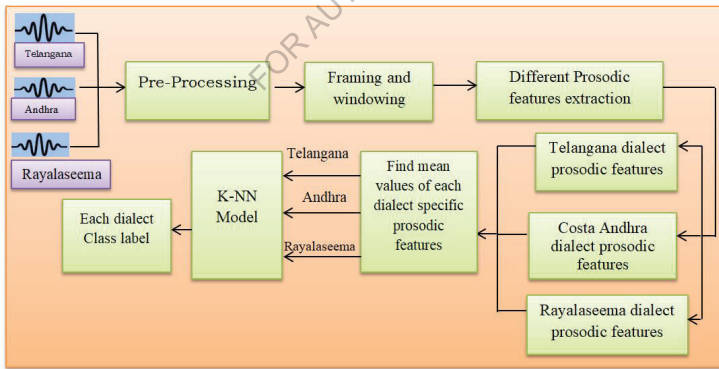


Figure 4.8: Methodology used in Training Phase

4.4.1.1 Testing Phase

The testing phase of dialect identification system identifies the dialect of test samples of input speech utterance. Using PRAAT tool, prosodic

features like Pitch, Energy, Loudness, and Formants are extracted from test speech utterance of short duration as like in Training phase. Find the mean value of prosodic features of test speech utterance of unknown speech and this mean value is treated as test data. This test data is evaluated against $K - NN$ model which is design in previous section. In this, K -value plays important role in deciding class of test data. Here K is chosen as 1. The distance using Manhattan is given by $K - NN$ model against three dialect classes are represented by β_1 , β_2 and β_3 . The dialects of testing sample is the dialect of train data which gives minimum of β_1 , β_2 and β_3 i.e. the distance between the mean of feature vector of test samples and dialect specific feature vector. The procedure applied in testing is shown in Fig.4.9.

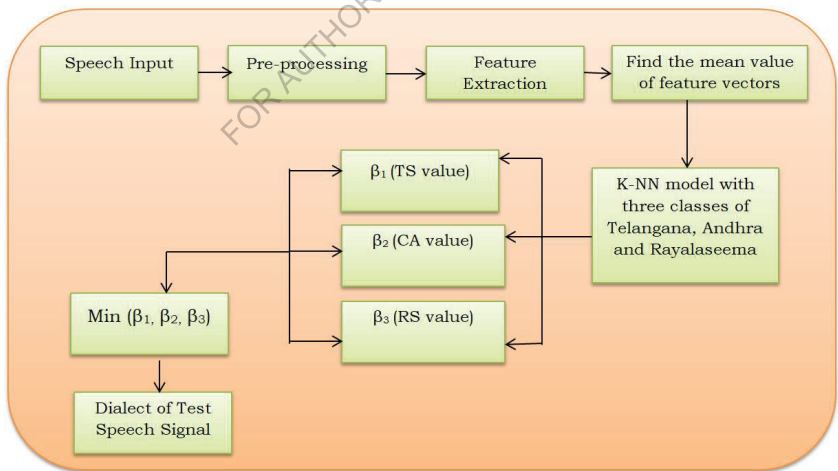


Figure 4.9: Methodology used in Testing Phase

4.5 Results

In this work, importance and behaviour of prosodic features in dialects of Telugu Language has been established. For this, we have extracted different prosodic features from windowed speech to get the average prosodic features of each dialect using PRAAT tool. The experiments are carried out on designed Telugu database.

The mean values of prosodic features which are calculated in the experiments for identification of different dialects of Telugu language as shown in Table.4.6.

Table 4.6: Mean values of different prosodic features related to each dialect

Feature Type	Telangana	Costa Andhra	Rayalaseema
Pitch	121.275	234.2495	176.97
Intensity	68.3340	71.31432	69.49
Energy	0.00344	0.046046	0.016
Loudness	59.4244	63.6638	54.12
Formant(f1)	360.163	484.9954	460.42
Formant(f2)	1621.69	1548.425	1418.3
Formant(f3)	2623.90	2675.073	2559.20
Formant(f4)	4102.91	3788.203	3781.35
Formant(avg)	2177.16	2124.174	2054.84
Pitch+ Loudness	116.8754	146.4618	89.82
Pitch+ Intensity	94.8048	152.78190	123.23
Pitch+Intensity+f1	227.484	318.8886	291.82
Pitch+Intensity+ Energy	47.4041	76.41033	61.62

From the above table, it is identified that Pitch, Intensity, Energy

values of the speech signal of different regions are higher for the Telangana region and lower for the Andhra region, and Rayalaseema, it falls in between Telangana and Andhra regions. As stated in section 4.3.2, the relation between pitch, intensity, and energy of speech signals are directly proportional to each other. Therefore, any combination of these three features results in the same order in terms of high and low. It is also observed that Pitch + Intensity and Pitch+Intensity+Energy are higher for Telangana low for Andhra. For formants, average considers it is more for Andhra and low for Rayalaseema.

As per the observation from prosodic features for three dialects of Telugu language, prosodic features are important cues to identify the dialects of Telugu Language from shortest duration.

The experiments are carried out the database of dialects of Telugu language which is designed as specified in section 1.5. In this work, different prosodic features are extracted from the huge speech utterances of three dialects i.e., Telangana, Costa Andhra and Rayalaseema and combined different prosodic features. Using these feature vectors, KNN-model is trained to design a reference model with three class labels corresponding to three dialects.

In identification phase, same prosodic features are extracted from shortest utterance of duration of test data (3s-8s) and evaluated in testing phase.

The results of Dialect Identification produced with prosodic features are shown in Table.4.7 and corresponding graph is shown in Fig.4.10.

Table 4.7: Overall accuracies produced by different prosodic features with the K-NN model

Feature Type	Accuracy of Model			
	Telangana	Costa Andhra	Rayalaseema	Average Accuracy
Pitch	100	80	45	75
intensity	60	70	50	60
Loudness	79	74	71.5	74.8
Energy	80	40	45	61.6
Formant(f1)	90	55.5	50	65
Formant(f2)	40	30	70	46.6
Formant(f3)	40	55.5	70	48.5
Formant(f4)	100	30	40	56.6
Formant(avg.)	50	30	50	43.3
Pitch+Intensity	80	70	80	76.6
Pitch+ Loudness	82	75.5	76	77.83

From the above table it is observed that, K-NN model provides good accuracy with Pitch+ Loudness is 77.83%. It has been observed that Pitch + Loudness are suitable features to identify the dialects of Telugu language.

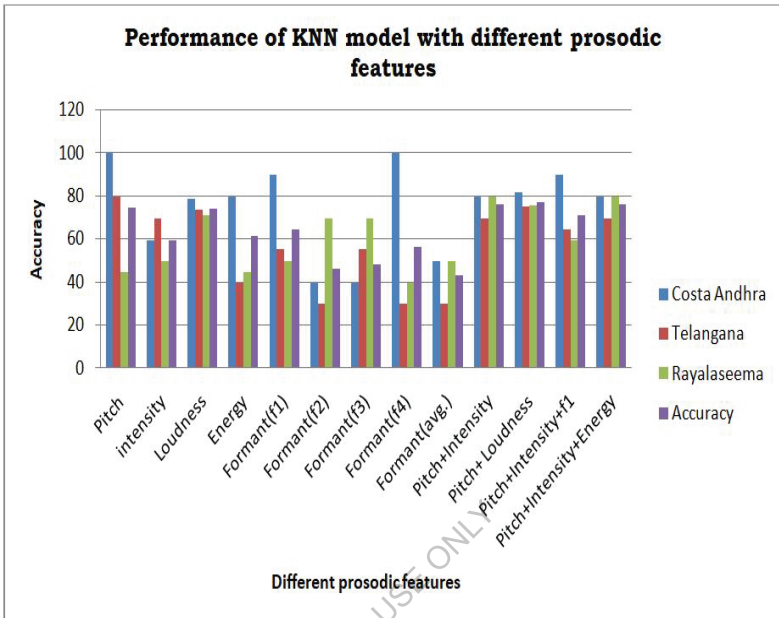


Figure 4.10: Performance of KNN model with different prosodic features

4.6 Conclusion

In this chapter, it has been established that prosodic feature vectors are important cues to discriminate three different dialects of Telugu Language. The dialect identification system for Telugu language was designed using prosodic features (Pitch, Intensity, Energy etc.). In first phase, the prosodic feature vectors of three dialects are derived and a reference model is designed using K-NN algorithm.

In second phase, prosodic features vectors of test speech samples are derived and evaluated to dialect specific reference model, give the results, based on minimum distance between test feature and mean

values which are developed in Training phase. The model gave the better performance with 77.83% for Pitch+ Loudness features.

Overall it is observed that Prosodic features are used to discriminate the dialects of Telugu language with respect to nasal level features. It is required to combine the spectral level and prosodic level features to clear discriminate the dialects of Telugu language.

FOR AUTHOR USE ONLY

Chapter 5

NEW FEATURES FOR TELUGU DIALECT IDENTIFICATION USING STATISTICAL APPROACHES

5.1 Introduction

The acoustic features of speech signals like spectral features i.e. MFCC and prosodic features are efficient features vectors in discriminating the different dialects in the Telugu language. These dialects are different from each other with respect to acoustic feature and prosodic features of speech signal. They show significant variance with respect to prosodic features like pitch, intensity, energy formants and Spectral features like MFCC etc. These features are important cues for recognizing the words with the same pronunciation or different one in different regional dialects.

The role and importance of these feature vectors was shown in identify the dialects in the Telugu language in the 3rd and 4th chapter. The spectral features can discriminate the different words of dialect in Telugu speech whereas prosodic features can discriminate the accent of different dialects in the Telugu language. As there are two kinds of features produced good results separately, the concatenation of these feature vectors are attracted for deriving new feature vectors. In this work, it is proposed to combine spectral features and prosodic features in order to discriminate the tonal and non-tonal dialects.

New features are derived by concatenating the prosodic features to spectral features. The spectral features (MFCC, Δ MFCC and $\Delta\Delta$ MFCC) and prosodic features (Pitch and Loudness) of speech signals are considered in identify the dialects of Telugu language and extracted using procedure explained in 3.2 section and 4.2 section.

As like other speech processing models, Dialect identification is also a two-step procedure. In the first step, the new features of training speech (huge data) are represented by compact notation. In the second step, new features of unknown utterance are evaluated with a reference model to find the dialect of unknown utterance.

In any pattern recognition problem including dialect identification, the performance depends on the size of the feature vector (number of Coefficients in a feature vector). It is very difficult for us to fix the size of the feature vector without loss of information [34].

Finding the optimal number of Coefficients in a feature vector is a difficult task from an original feature vector as if the number of coefficients chosen is few then loss of information might be occurred or if the number of the coefficients chosen is more the time for computation is increased. So that finding optimal feature vector is a challenging task in dialect identification.

There are several existing methods like PCA, ICA, wavelet transfor-

mation etc., to find the reduced feature vectors from original feature vectors [68, 69].

The PCA is a popular reduction technique which works by calculates the eigenvalues and eigenvector to eliminate redundant data and number of coefficient in a feature vector [76].

There are so many evidences of PCA that improves the performance of speech processing models by reducing the feature vectors for training and testing phases of the system [68, 69].

In order to improve the system performance and reduce the time taken for training and testing, PCA is considered for eliminating redundant information. In this chapter initially, new feature vectors have been derived from spectral features i.e. MFCC, Δ MFCC, and $\Delta\Delta$ MFCC, and prosodic features Pitch, Loudness. Using these new feature vectors, different Telugu dialect identification systems have been designed with HMM, GMM, and DNN. In order to find out optimized feature vectors, PCA has been applied on new feature vectors of MFCC, Δ MFCC, and $\Delta\Delta$ MFCC + Pitch, Loudness. Using these Optimizer feature vectors, HMM, GMM and DNN based dialect identification system have been designed. Then the performances of different systems have been compared.

This chapter is organized into seven sections. Section 1 is an introduction and the remaining sections are as follows:

Section 2 represents the extraction of new feature vectors from spectral and prosodic features. Section 3 explains the proposed model for dialect identification using new feature vectors. The Optimized feature vectors from new feature vectors are derived in Section 4 and dialect identification systems are design using optimized feature vectors in Section 5. Section 6 presents the results of the Telugu dialect identification system with new features and optimized feature vectors. Finally, conclusions are mentioned in section 7.

5.2 New Feature Vectors for Dialect Identification

Feature vectors are very important and prominent for any pattern recognition system in order to identify any pattern. Like this, feature vectors are essential for a dialect identification system. These describe the abstract behavior and content of speech signals in terms of data.

As specified in chapter 3, popular and familiar features in speech processing are MFCC. It has been established that MFCC has good results in the case of spectral behavior speech signals but not in nasal languages and tonal speech utterances. Whereas, the Prosodic features Pitch+ Loudness performed well in the case of tonal and nasal speech utterances.

To deal with spectral properties and tonal and nasal speech utterances, it is proposed to concatenate prosodic features Pitch+ Loudness

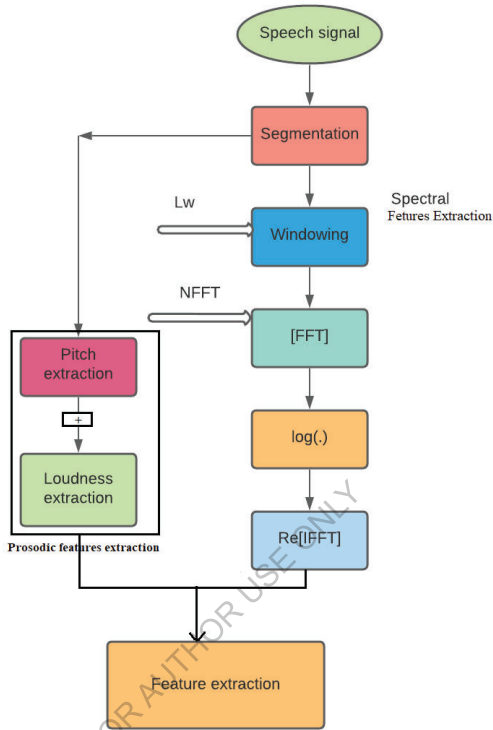


Figure 5.1: Basic diagram to extract new features

to 39-dimensional MFCC feature vector (13 MFCC+ 13 Δ MFCC + 13 $\Delta\Delta$ MFCC).

MFCC feature vectors are derived using the method as explained in the section 3.2 and prosodic feature vectors Pitch + Loudness are derived using the method explained in chapter 4. Fig.5.1 describes the extraction of new features vectors using acoustic and prosodic features.

5.3 Methodology Used for Dialect Identification

Dialect Identification system consists of two phases namely the Training phase and the Testing phase. The database which has been created for experiments is used for Training and Testing purposes. The speech utterances considered for training and testing are different. In both the training and testing phase, new feature vectors are derived for implementation.

5.3.1 Training Phase

The training phase of the proposed model has two steps i.e. feature extraction and designing the dialect specific reference model. In the feature extraction step, new features are derived by using the acoustic features and prosodic features as specified in section 5.2.

In this work, MFCC, Δ MFCC, $\Delta\Delta$ MFCC, Pitch and Loudness are extracted from windowed dialect specific speech corpus and concatenated to form new feature vectors for each dialect of Telugu Language. Different experiments are carried out by combining acoustic and prosodic features to increase the performance of the system.

In the second phase, a statistical method like HMM/GMM/DNN is trained using derived new feature vectors of three dialects of Telugu Language such that one HMM/GMM/DNN model for one kind of dialect of the Telugu language. The training phase gives the compact representation of huge feature vectors of each dialect of the Telugu language.

This compact representation is also known as the reference model which is used in the Testing phase.

5.3.2 Testing Phase

Feature extraction from shortest duration test samples of speech and finding the likelihood value of these feature vectors are the two steps in testing phase of Telugu dialect identification system. The feature vector of test samples are extracted as like training phase and these feature vectors are evaluated against each dialect reference model in order to get the likelihood of observation sequence of these feature vectors.

In this step, acoustic and prosodic features are extracted from the shortest duration of a test utterance and concatenated to form new feature vectors as like the training phase. These new feature vectors are considered as observation sequence and this observation sequence of Test speech data are evaluated against three dialect reference models of Rayalaseema, Telangana, and Costa Andhra which are created in Training phase. The model with maximum likelihood represents the dialect of unknown utterance. The block diagram of the proposed method for Telugu dialect identification as shown in Fig.5.2.

5.4 Optimized features for Dialect Identification

To reduce the time taken by model for training and to increase the accuracy of model to identify the dialects of Telugu language, number of coefficients in each feature vectors are reduced using different dimen-

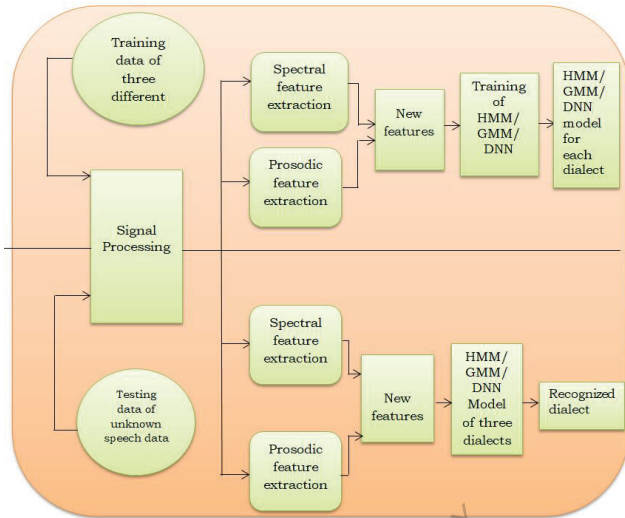


Figure 5.2: Block diagram of the proposed method for dialect identification

sionality reduction techniques. The new feature vectors are extracted by combining the MFCC, Pitch and Loudness as discussed in section 5.2, the resultant feature vector size is 41, but this much size feature vector, the model takes long time for training and testing. In order to reduce the size of the feature vector, there are several methods in literature survey like PCA, wavelet transformation, rough set theory etc. [64].

PCA also performed well in reducing the dimensionality. In this work, PCA is used to reduce the 41-dimensional new feature vector (MFCC+Pitch+ Loudness) to optimize feature vector of 30 size.

5.4.1 Principal Component Analysis (PCA)

In huge data, the feature vector size is very important and which impacts the training and testing time. It is also very difficult for a model

to analyse the feature vectors with more dimensionality for establishing that feature vectors are informative are not. The model may take more time for learning the data if the number of coefficients are more in feature vectors.

To improve the efficiency of modelling, it is essential to eliminate redundant or unnecessary data and reduce the number of coefficients in feature vector. In the literature survey, there are number of dimensional reduction techniques like PCA, ICA, rough set model, wavelet transformation [34, 64]. PCA is a popular and easy technique to reduce the dimensionality of huge dataset by decreasing the number of coefficients in feature vectors into smaller which contains almost same information as huge dataset.

The working principle of PCA is simple, It finds the matrix X with size d for the input matrix W with size e such that $d < e$ by capturing maximum variance of input data. The new data is linear function of input data and uncorrelated each other. The basic steps of PCS is:

1. Normalize the input data to the continue initial range value.
2. To identify correlation between input data, covariance matrix is calculated.
3. Principle components are found by calculating Eigen vectors from covariance matrix.

4. Derive features vectors based on the selection of Principle components.
5. Recast the data along with principle components exists.

In the first step, continuous input data is standardized using statistical measure on data i.e. mean and standard deviation. So that, all input continuous values are to be same scale. Standardization is done with the equation 5.1 on each value of every variable.

$$Z = \frac{d - \mu}{\sigma} \quad (5.1)$$

where d is input value, μ is mean and σ is standard deviation.

In step2, Correaltes among the feature vectors are computed using covariance matrix by understanding the variance of input and mean with repeat to each outlier. This step is useful to identify redundant information, if they are more correlated.

In step3, Eigen vectors and Eigen values are computed from covariance matrix of input data inorder to find the principle components of data. These principle componets are constituted as liner combination or mixture of actual intial variables inorder to get uncorrelated data. Such that, first component has maximum information of intial variable, then next maximum information in second component and so on as like shown in Fig. 5.3.

From the above steps, remove the principle components which have less information and repeat the step3 by recasting the data. Principle

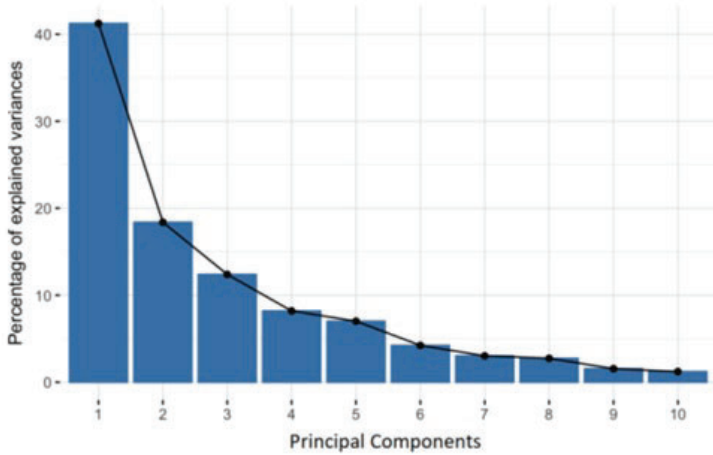


Figure 5.3: Principle components

components with more information are considered as new feature vectors which have maximum information of original feature vectors. The detailed work flow of PCA is presented in Fig.5.4.

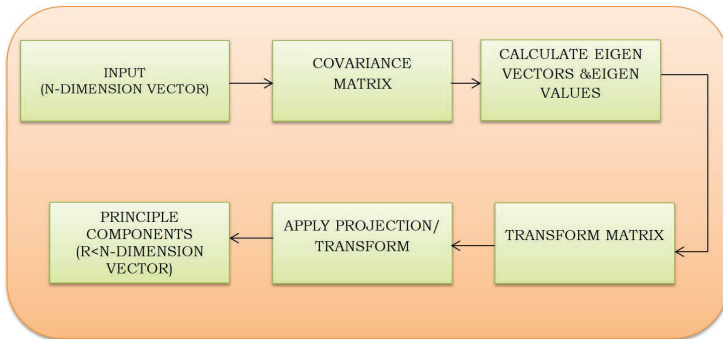


Figure 5.4: Methodology to calculate PCA.

5.5 Dialect Identification using Optimized Feature Vectors

To find the dialects of the Telugu language with the shortest duration of speech utterances, new feature vectors with 41-dimensionality are derived by using spectral and prosodic features. In order to reduce the dimensionality of feature vectors, PCA is applied to form 30 dimensional feature vectors using the algorithm explained in section 5.5.

The detailed procedure to extract optimize features from new feature vector is given in Fig.5.5.

The proposed dialect identification system with an optimized feature vector is also a two-step method i.e., Training phase and testing phase.

FOR AUTHOR USE ONLY

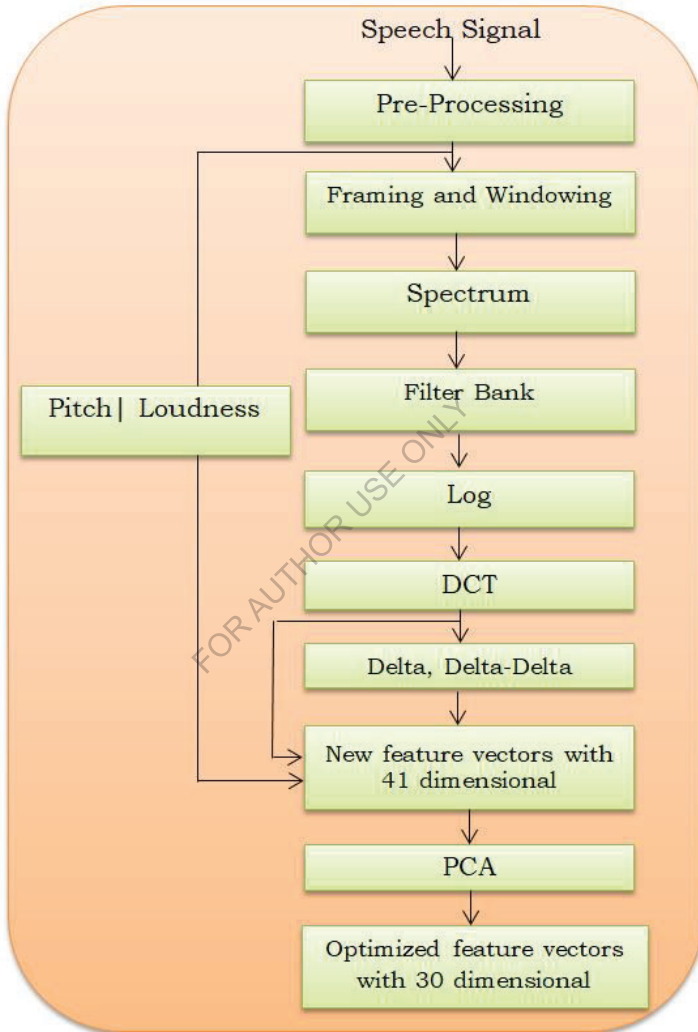


Figure 5.5: Optimized feature extraction

5.5.1 Training Phase

This Training phase creates a reference model for optimized feature vectors of Training speech samples of different dialects. Initially, optimized feature vectors are extracted from three dialects of the Telugu language using the method explained in section 5.4.

In the training phase, a reference model for optimize feature vectors of Huge training samples are created using GMM/HMM/DNN. Such that one model for one dialect is created. In the testing phase, optimize feature vectors are evaluated against the reference model.

Then, the optimized feature vectors are given input to HMM/GMM/DNN to design a HMM/GMM/DNN for each dialect for the compact representation of huge data. The schematic diagram of the training phase is given in Fig.5.6.

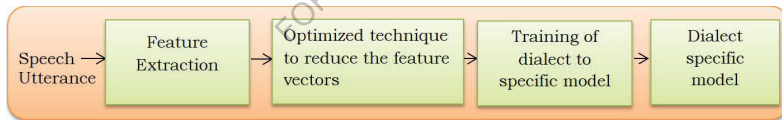


Figure 5.6: Training phase

5.5.2 Testing Phase

In the testing phase of the dialect identification system, 30-dimensional optimized feature vectors are derived from 41-dimensional feature vectors of the shortest utterance speech sample. These 30-dimensional optimized feature vectors of test utterance are considered as observation sequences. This observation sequence of test utterance is evaluated against reference models of three dialects. The dialect-specific reference

model which gives the maximum likelihood score is the dialect of test utterance. Fig.5.7 describes the schematic representation of the testing phase.

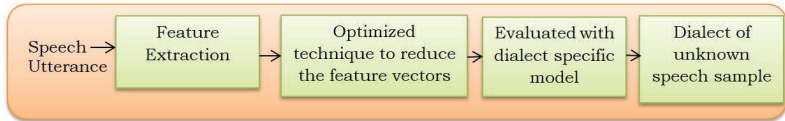


Figure 5.7: Testing Phase

5.6 Results of Dialect Identification with New Feature Vectors and Optimized Feature Vectors

To carry out experiments, we have used Telugu dialect speech database created by us as specified in chapter 1. The Telangana, Costa Andhra and Rayalaseema has 2h 35min duration, 2h 47min duration, and 1h 43min collected nearly 7h 05min speech has been used for training and testing.

In this work, MFCC feature vectors and prosodic feature Pitch and Loudness are extracted from the training speech and testing speech to derive new feature vectors.

MFCC features captured the difference among three dialects but in some cases it does not capture. So that it gives some less accuracy in case of same word with different accents of different dialects. For example:

It has proved Pitch and Loudness perform well in this case in chapter 4. So in this work, MFCC features and prosodic features are concate-

Telangana	Costa Andhra	Rayalaseema
సంతోషం	సంతోషం	సంతోషం

nated to form a new feature vectors for training and testing phases of dialect identification system.

In this, different experiments are carried out with new features of 41 dimensional which are derived by concatenating 39-dimensional MFCC features, prosodic features (Pitch and Loudness) using Hidden Markov Model (HMM), Gaussian Mixture Model (GMM) and Deep Neural Network (DNN).

In these experiments, 41-dimensional new feature vectors of train speech are used to create reference model using HMM/GMM/DNN. In testing phase, 41-dimensional new feature vectors are derived from testing sample of 3s-8s and evaluated against the reference models to identify the dialects of test samples.

The performance of GMM based dialect identification system using new features for three dialects of Telugu is depicted in Table 5.1 and corresponding graph is given in Fig.5.8.

Table 5.1: Performance of GMM with new feature vectors

Feature Extraction	Performance of GMM Model		
	Telangana	Costa Andhra	Rayalaseema
New Feature vectors (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	89.12	88.7	88.8

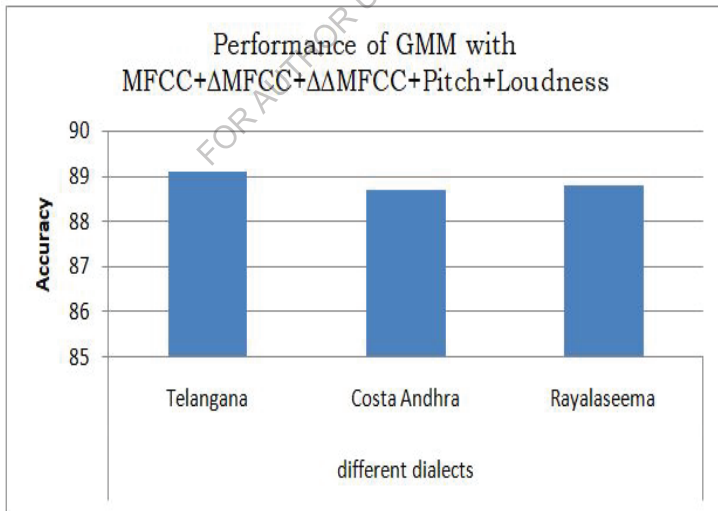


Figure 5.8: Performance of GMM with new feature vectors

From the results, It is observed that GMM model gave the better performance with new feature vector MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness with 89.12%, 88.7% and 88.8% with respect to Telangana, Costa Andhra and Rayalaseema.

The performance of HMM based Dialect Identification using new features is shown in Table.5.2 and corresponding graph is shown in Fig.5.9.

Table 5.2: Performance of HMM with new feature vectors

Feature Extraction	Performance of HMM Model		
	Telangana	Costa Andhra	Rayalaseema
New Feature vectors (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	89.12	88.7	88.8

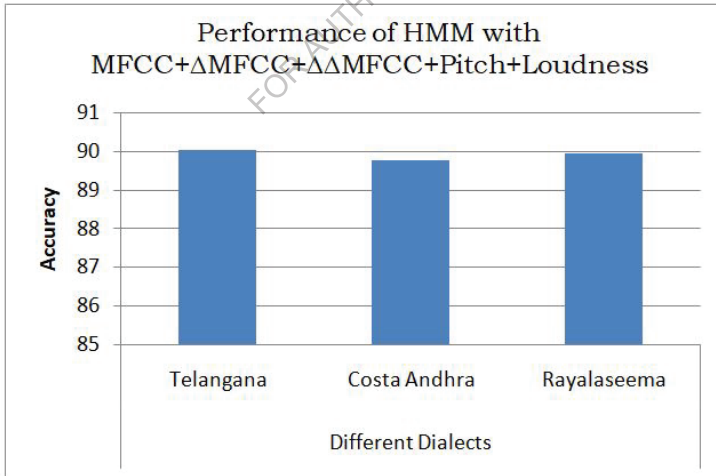


Figure 5.9: Performance of HMM model with new feature vectors

From the results, it is observed that like GMM , HMM also produced the good performance with new feature vector MFCC+ Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness is 88.41%, 87.63% and 86.16% with respect to Telangana, Costa Andhra and Rayalaseema.

The performance of DNN based Dialect Identification using new features is shown in Table.5.3 and corresponding graph is shown in Fig.5.10.

Table 5.3: Performance of system with DNN with new feature vectors

Feature Extraction	Performance of DNN Model		
	Telangana	Costa Andhra	Rayalaseema
New Feature vectors (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	90.03	89.75	89.95

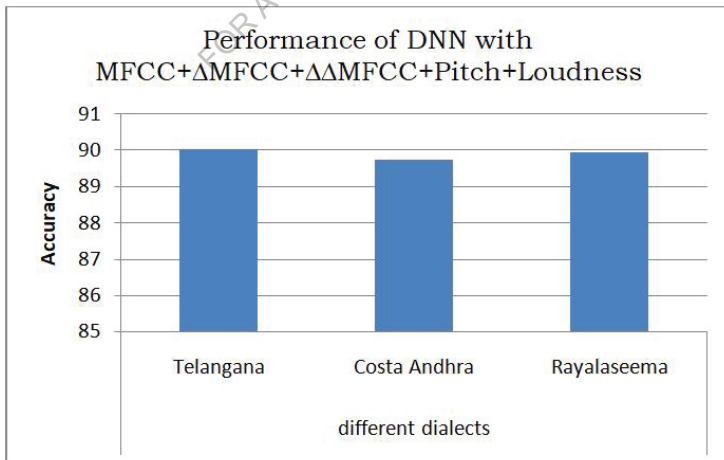


Figure 5.10: Performance of DNN model with new feature vectors

From the results, it is observed that DNN model produced the good performance with new feature vectors MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness as like GMM and HMM model. The accuracies of model is 90.03%, 89.75% and 89.95% with respect to Telangana, Costa Andhra and Rayalaseema.

From the results table 5.1, 5.2, 5.3 , it is observed that, overall DNN model gave the better performance with new feature vectors MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness with 90.03%, 89.75 and 89.95% for Telangana, Costa Andhra and Rayalaseema respectively compare to HMM and GMM models.

To reduce the dimensionality of feature vectors in order to improve the performance of dialect identification, the dimensionality of new feature vectors (41 dimensional) is reduced to optimize feature vectors of 30-dimensional using PCA.

Using these optimized feature vectors, dialect identification systems using HMM, GMM and DNN have been designed on the same Telugu dialect database.

The performance of HMM-based dialect identification using an optimized feature vector is shown in Table 5.4 and corresponding graph in Fig.5.11.

Table 5.4: The performance of HMM-based dialect identification using optimized feature vectors

Feature Extraction	Performance of HMM Model		
	Telangana	Costa Andhra	Rayalaseema
New Feature vectors (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	88.41	87.63	86.16
Optimized feature vectors - 30 dimensional (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	89.03	88.12	86.55

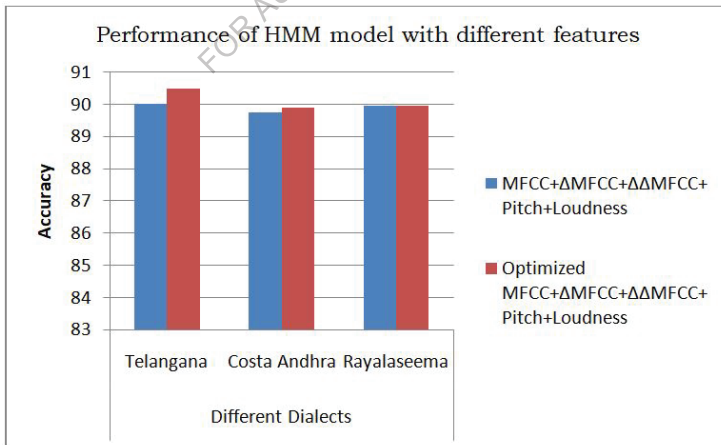


Figure 5.11: Performance of HMM model with optimized features.

The performance of GMM-based dialect identification using an optimized feature vector is shown in Table 5.5 and corresponding graph in Fig.5.12.

Table 5.5: The performance of GMM-based dialect identification using optimized feature vectors

Feature Extraction	Performance of GMM Model		
	Telangana	Costa Andhra	Rayalaseema
New Feature vectors (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	89.12	88.7	88.8
Optimized feature vectors-30 dimensional (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	90.1	89.09	89.01

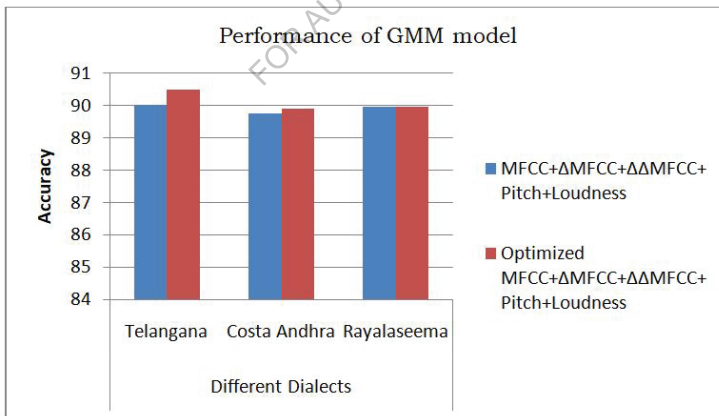


Figure 5.12: Performance of GMM model with optimized features

The performance of DNN-based dialect identification using an optimized feature vector is shown in Table 5.6 and corresponding graph in Fig.5.13.

Table 5.6: The performance of DNN-based dialect identification using optimized feature vectors

Feature Extraction	Performance of DNN Model		
	Telangana	Costa Andhra	Rayalaseema
New Feature vectors (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	90.03	89.75	89.95
Optimized feature vectirs - 30 dimensional (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	90.5	89.9	89.96

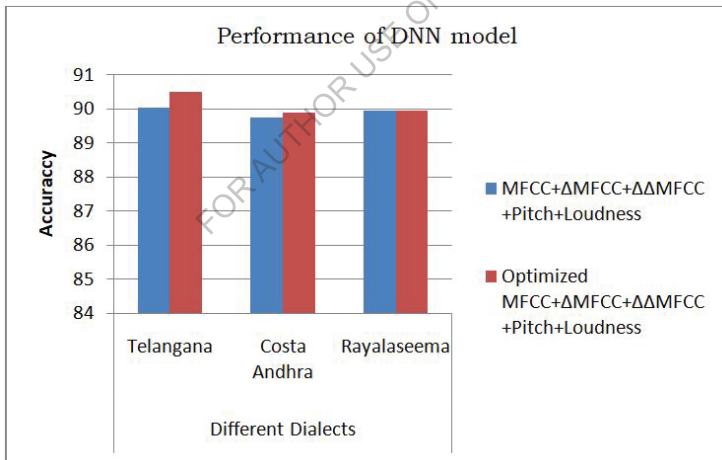


Figure 5.13: Performance of DNN model using Optimized features

The results of different dialect identification systems using optimized feature vectors are improved compared to new feature vector, it is observed that DNN model with 90.12% performed well compare to GMM and HMM in case of optimized feature vectors also.

5.7 Conclusion

In this chapter, the significance of new features which are derived from acoustic and Prosodic features has been established. The 39-dimensional MFCC feature vectors and Pitch and Loudness are considered and concatenated to form new feature vectors. Different dialect identification systems with GMM, HMM and DNN have been created using 41-dimensional new feature vectors (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness). The performance of dialect identification with DNN is impressive.

To reduce the dimensionality of feature vector and increase the accuracy, PCA has been used to reduce the feature vector size from 41 to 30. Using these optimized feature vectors, Dialect Identification system have been designed using HMM, GMM and DNN.

The performance of Dialect Identification is improved with optimized feature vectors compare to new feature vectors. Out of these three modelling techniques, DNN based Dialect Identification gave the good performance with 90.5%, 89.9% and 89.96% for Telangana, Costa Andhra and Rayalaseema respectively.

Chapter 6

PERFORMANCE EVALUATION OF DIALECT IDENTIFICATION SYSTEMS WITH DIFFERENT MODELLING TECHNIQUES

6.1 Introduction

In this task, a dialect identification system has been implemented with different feature vectors and different modeling techniques. The selection of feature vectors and model, to represent features are important for the performance of the system. The performance is measured in terms of the accuracy of the system that recognizes the test samples correctly.

To carry out experiments, the Telugu database with three dialects Telangana, Costa Andhra, Rayalaseema has been created by recording speech utterances in different environments.

The database consists of 2hours 35minutes for Telangana, 2hours 47 minutes speech for Costa Andhra, and 1hour 43 minutes speech for Rayalaseema utterances.

This chapter explains the new experimental setup and the performance of different dialect identification systems with different modeling techniques and features. Initially spectral feature based (MFCC) dialect identification system with HMM, GMM and DNN has been implemented. These systems performed well but gave with some less accuracy results

in case of nasal sounds. In order to avoid this, prosodic features importance has been established and new features were derived by concatenating MFCC and prosodic features like pitch and loudness. Finally, optimized features were derived and dialect identification system using HMM, GMM and DNN has been implemented with this feature vectors. A comparison study has been established with reputed published work to identify the dialects.

6.2 Experimental Setup

The database has been created using the PRAAT tool, Sony digital voice recorder, online streaming recorder and pre-processed with average filter.

The experiments are implemented using the Python programming language. Dialect identification systems have been implemented using HMM, GMM, and DNN with different types of feature vectors. These, dialect identification are evaluated to identify dialect of the shortest duration of unknown utterance (3-8 sec). In case of HMM experiments are explored with different number of states and mixtures. As it gave the good results with, 3-states and 32-mixtures. The HMM based Dialect Identification systems have been implemented with 3 states and 32 mixtures. In case of GMM, 32 mixtures are considered to implement Dialect Identification system. In case of DNN, 4 layers are used to implement Dialect Identification system.

6.3 Performance Evaluation of Dialect Identification using MFCC with Different Models

The performance evaluation of the dialect identification system using MFCCs has been done with the modeling techniques HMM, GMM, and DNN on the Telugu database for dialects. The test utterance duration is 3 to 8 Sec. The performance of dialect identification for the Telugu database is showed in Table 6.1.

Table 6.1: Performance of Dialect Identification system of Telugu Language using different models with MFCC + Δ MFCC + $\Delta\Delta$ MFCC

Feature Extraction	Model	Accuracy of Model		
		Telangana	Costa Andhra	Rayalaseema
MFCC+ Δ MFCC+ $\Delta\Delta$ MFCC	GMM	85.3	82.6	79.9
	HMM	84.4	80.2	83.2
	DNN	85.1	84	84.6

From the above table, it is observed that, the GMM model gave the performance 85.3%, 82.6% and 79.9% for Telangana, Costa Andhra and Rayalaseema respectively. The HMM model gave the performance is 84.4%, 80.2% and 83.2% for Telangana, Costa Andhra and Rayalaseema respectively. The DNN model gave the performance 85.1%, 84% and 84.6% for Telangana, Costa Andhra and Rayalaseema respectively.

Overall DNN model performed well in dialect identification with 39-dimensional MFCC features as shown in Fig.6.1.

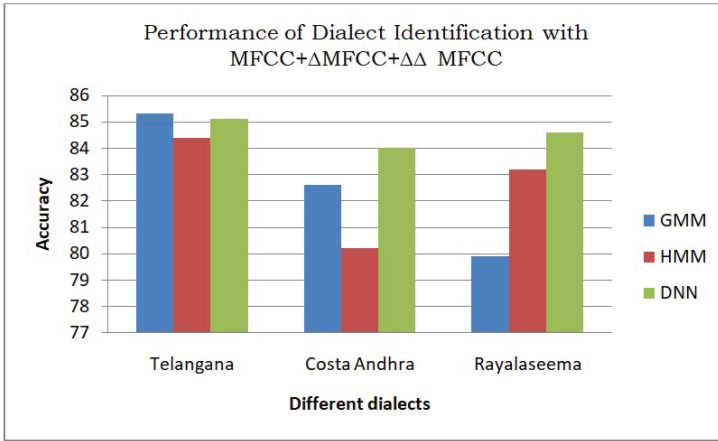


Figure 6.1: Performance of Dialect identification system with different models using MFCC + Δ MFCC + $\Delta\Delta$ MFCC

From the Fig.6.1, it is observed that DNN model has performed well with 85.1%, 84% and 84.6% for Telangana, Costa Andhra and Rayalaseema respectively.

6.4 The Performance Evaluation of Dialect Identification with MFCC and Prosodic Features using HMM, GMM and DNN

The dialect identification performance has been evaluated using different modeling techniques HMM, GMM and DNN with new features which are derived from MFCC and Prosodic features as specified in section 5.4. The performance of Dialect Identification with HMM, GMM and DNN has been depicted in Table 6.2, 6.3 and 6.4 respectively. The corresponding graphs are shown in Fig. 6.2, 6.3 and 6.4.

Table 6.2: Performance of Dialect Identification System for GMM with new feature vectors

Feature Extraction	Performance of GMM Model		
	Telangana	Costa Andhra	Rayalaseema
MFCC + Δ MFCC + $\Delta\Delta$ MFCC	85.3	82.6	79.9
New feature vectors (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	89.12	88.7	88.8

From the results, it is observed that GMM model gave the good performance with new feature vectors with 89.12%, 88.7% and 88.8% for Telangana, Costa Andhra and Rayalaseema respectively. The corresponding graph is shown in Fig.6.2.

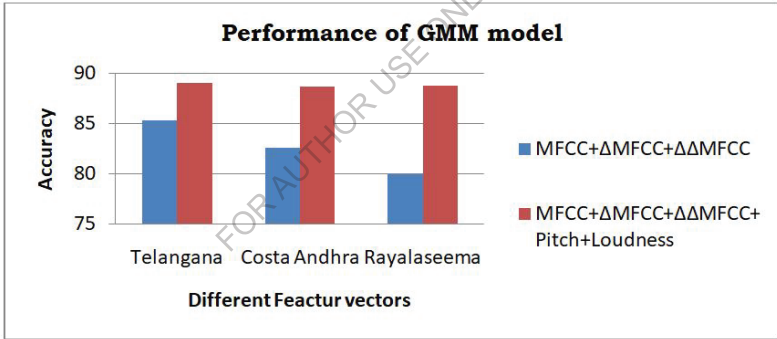


Figure 6.2: GMM based Dialect Identification with MFCC and New feature Vectors

Table 6.3: Performance of Dialect Identification System for HMM with new feature vectors

Feature Extraction	Performance of HMM Model		
	Telangana	Costa Andhra	Rayalaseema
MFCC + Δ MFCC + $\Delta\Delta$ MFCC	84.4	80.2	83.2
New feature vectors (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch+Loudness)	88.4	87.6	86.1

From the results, it is observed that HMM model gave the good performance with new feature vectors with 88.4%, 87.6% and 86.1% for Telangana, Costa Andhra and Rayalaseema respectively. The corresponding graph is shown in Fig.6.3.

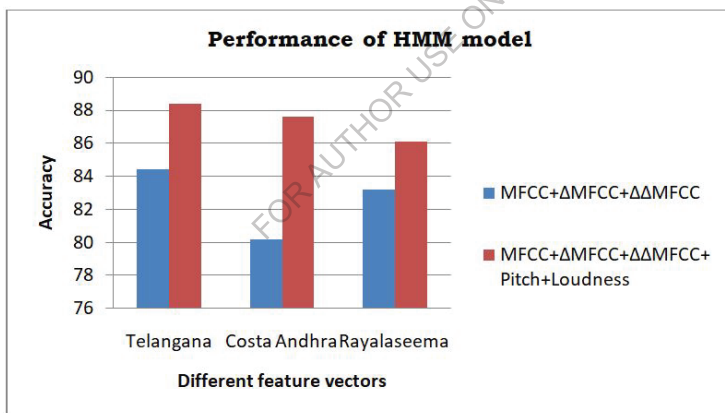


Figure 6.3: HMM based Dialect Identification with MFCC and New feature Vectors

Table 6.4: The performance of Dialect Identification System for DNN with new features

Feature Extraction	Performance of DNN Model		
	Telangana	Costa Andhra	Rayalaseema
MFCC + Δ MFCC+ $\Delta\Delta$ MFCC	85.1	84	84.6
New feature vectors (MFCC Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	90	89.7	89.9

From the results, it is observed that DNN model gave the good performance with new feature vectors with 90.4%, 89.7% and 89.9% for Telangana, Costa Andhra and Rayalaseema respectively. The corresponding graph is shown in Fig.6.4.

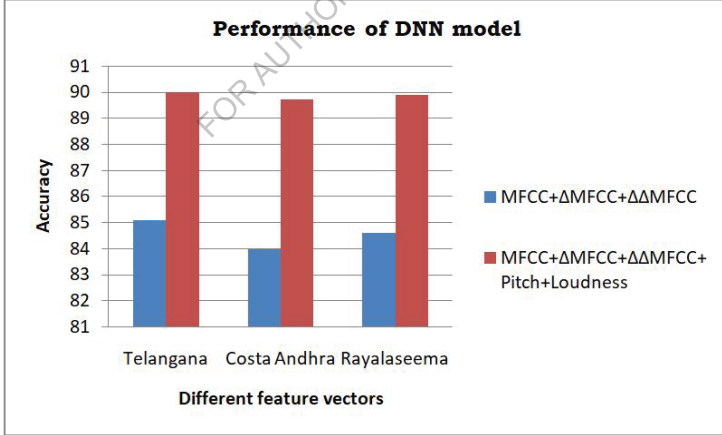


Figure 6.4: DNN based Dialect Identification with MFCC and New feature Vectors

It is observed that, the performance of Dialect Identification is improved with new features for HMM, GMM and DNN compared to spectral features i.e., MFCC.

It is observed that the performance of Dialect Identification System with MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness feature vector for HMM is 87.4%, GMM is 88.87% and DNN is 89.91%. It is also observed the performance of DNN based Dialect Identification System using MFCC+prosody features is improved compare to MFCC features in three dialects. The overall performance of system with new feature has improved for DNN modeling techniques.

6.5 Performance Evaluation of Optimized Feature Vectors

In this section, the performance of Dialect Identification system using optimized features vectors has been evaluated. The optimized feature vectors are derived using Principle component Analysis (PCA). 30-dimensional feature vectors (Optimized feature vector) are derived from 41-dimensional new feature vectors. The performance of Dialect Identification system with optimized features and new features for HMM, GMM and DNN models shown in Table 6.5, 6.6 and 6.7 respectively.

Table 6.5: The performance of GMM based Dialect Identification System using optimized feature vectors

Feature Extraction	Performance of GMM Model		
	Telangana	Costa Andhra	Rayalaseema
New Feature vectors - 41 dimensional (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	89.12	88.7	88.8
Optimized Feature vectors - 30 dimensional (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	90.1	89.09	89.01

From the results, it is observed that GMM model gave the good performance with optimized feature vectors with 90.1%, 89.09% and 89.01% for Telangana, Costa Andhra and Rayalaseema respectively. The corresponding graph is shown in Fig.6.5.

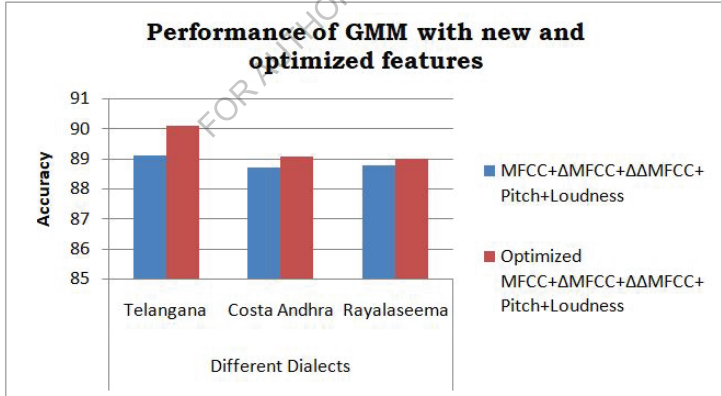


Figure 6.5: GMM based Dialect Identification with New and optimized feature Vectors

Table 6.6: The performance of HMM based Dialect Identification System using optimized features

Feature Extraction	Performance of HMM Model		
	Telangana	Costa Andhra	Rayalaseema
New Feature vectors - 41 dimensional (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	88.41	87.63	86.16
Optimized Feature vectors - 30 dimensional (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	89.03	88.12	86.55

From the results, it is observed that HMM model gave the good performance with optimized feature vectors with 89.03%, 88.12% and 86.55% for Telangana, Costa Andhra and Rayalaseema respectively. The corresponding graph is shown in Fig.6.6.

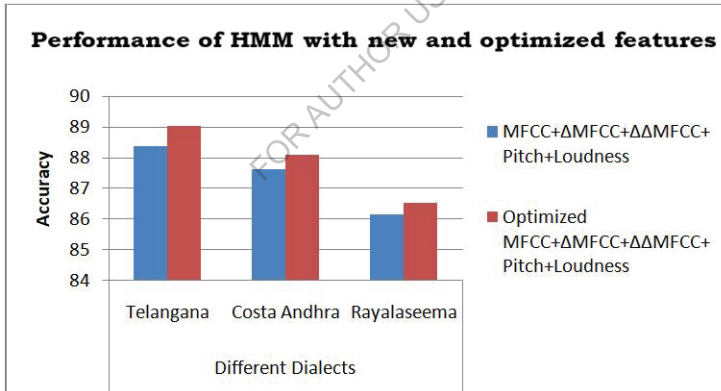


Figure 6.6: HMM based Dialect Identification with New and optimized feature Vectors

Table 6.7: The performance of DNN based Dialect Identification System using optimized features

Feature Extraction	Performance of DNN Model		
	Telangana	Costa Andhra	Royalaseema
New Feature vectors - 41 dimensional (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	90.03	89.75	89.95
Optimized Feature vectors - 30 dimensional (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	90.5	89.9	89.96

From the results, it is observed that DNN model gave the good performance with optimized feature vectors with 90.5%, 89.9% and 89.96% for Telangana, Costa Andhra and Royalaseema respectively. The corresponding graph is shown in Fig.6.7.

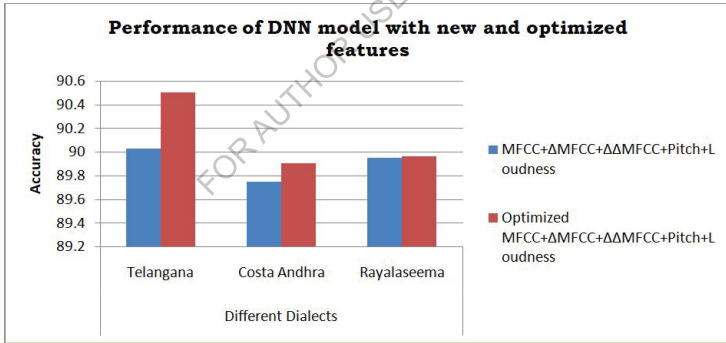


Figure 6.7: DNN based Dialect Identification with New and optimized feature Vectors

From the results of Dialect Identification System, it is observed that, the DNN model performed well with optimized feature vectors compare to HMM and GMM. Overall, the DNN model gave the good performance with optimized feature vectors with 90.5%, 89.9% and 89.96% for Telangana, Costa Andhra and Royalaseema respectively.

The time for computational task has been analyzed for the experiments carried out. The computational time has been reduced using optimized features (30-dimensional vector) compared to actual feature vector (41-dimensional vector). The time taken for identify test utterance of speech with different duration of speech 3sec, 5sec and 8 sec was analyzed and average time taken for identifying the dialects using 30-dimensional optimized features and 41-dimensional proposed new feature vectors is shown table 6.8.

Table 6.8: Analysis of time taken for identify the dialect of test utterance using DNN

Feature Vector	Average Time taken to identify the dialects of test sample in milli seconds		
	3 Sec	5 Sec	8 Sec
New Feature vectors - 41 dimensional (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	69.01	69.80	76.60
Optimized Feature vectors - 30 dimensional (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	67.05	68.43	75.49

A significant improved was found in identify the dialect of test samples with respect to computational time using optimized feature vectors. The time taken for identifies the test utterance for 3sec, 5sec and 8sec is shown in Fig.6.8.

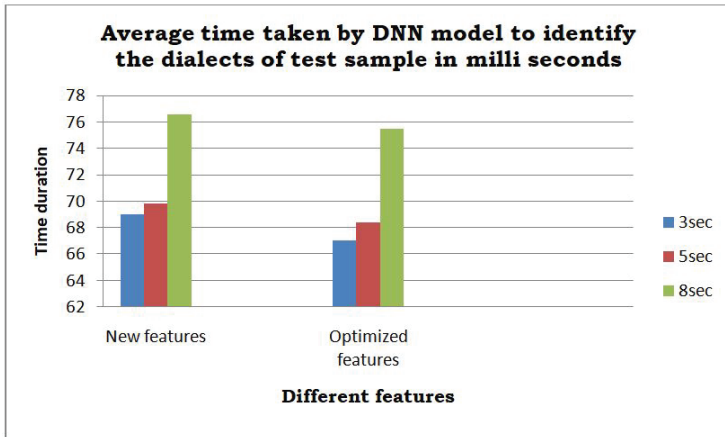


Figure 6.8: Analysis of DNN model identification time with different duration of test samples

It is observed that optimized feature vectors performed well in case 3s,5s and 8s test samples to identify the dialects using DNN.

6.6 Comparison of different Dialect Identification System

In this section, the average performance of the dialect identification systems has been compared. Performance of HMM, GMM and DNN based dialect identification system has been explored using different feature types in order to identify dialect of Telugu language from shortest utterance of speech.

In case of HMM, considered 3 states and 32 mixtures at each state. The GMM has been implemented with 32 mixtures and DNA has one input layer, two hidden layers and one output layer. In this task, MFCC, Prosodic features, new features by combining MFCC+ Pitch+ Loudness

and optimized feature vectors have been used for the experiments to identify the dialects. The average performance of different model techniques with different features has been mentioned in Table 6.9.

Table 6.9: Average Performance of Different models in dialect identification with different features

Model	Feature Vector	Accuracy
HMM	MFCC + Δ MFCC + $\Delta\Delta$ MFCC	82.6
	New Feature vectors - 41 dimensional (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	87.4
	Optimized Feature vectors - 30 dimensional (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	87.9
GMM	MFCC + Δ MFCC + $\Delta\Delta$ MFCC	82.6
	New Feature vectors - 41 dimensional (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	88.87
	Optimized Feature vectors - 30 dimensional (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	89.4
DNN	MFCC + Δ MFCC + $\Delta\Delta$ MFCC	84.6
	New Feature vectors-41 dimensional (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	89.91
	Optimized Feature vectors - 30 dimensional (MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness)	90.12

From, the results, the average performance of HMM, GMM and DNN model with optimized feature vectors is 87.9%, 89.4% and 90.12% respectively. Whereas, 87.4%, 88.87% and 89.91% with proposed new features. The corresponding graph is depicted in Fig.6.9 and 6.10.

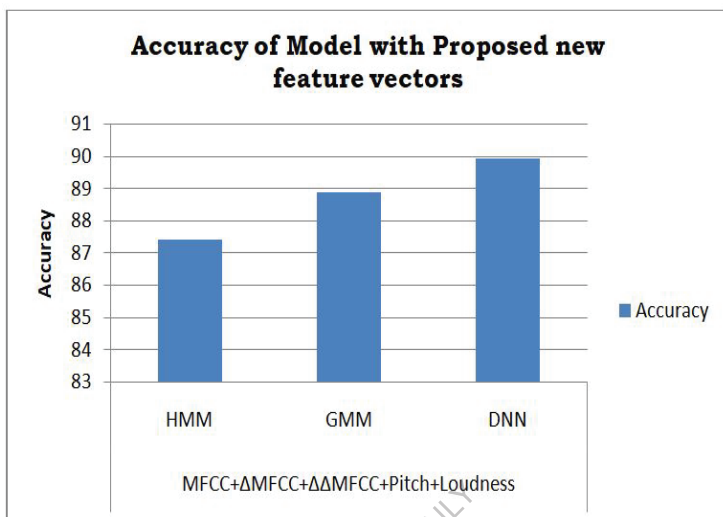


Figure 6.9: The performance of HMM, GMM and DNN using proposed new features

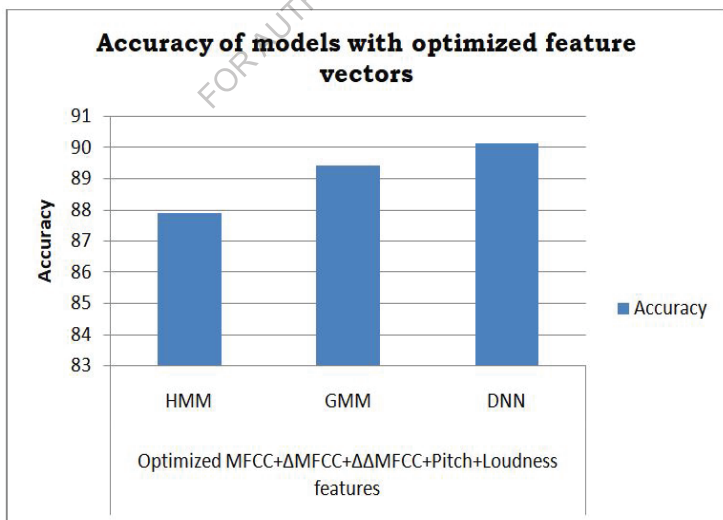


Figure 6.10: The performance of HMM, GMM and DNN using optimized features

From Fig. 6.8 and 6.9, it is observed that average performance of DNN is impressive with optimized features. Overall, DNN model with Optimized (30-dimensional) MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness features provided good performance in dialect identification of Telugu language with 90.12%.

6.7 Comparison study with reputed published work

In this thesis, Dialect Identification system with new features and optimized features has been implemented with HMM, GMM and DNN. DNN has well performed with optimized feature vector (derived from MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness feature vector). These results are compared with few recent published work where in identified dialects in different languages.

In this comparison study, type of features, modelling technique, training duration and testing duration are compared with published work. The comparison study done with DNN with optimized MFCC + Δ MFCC + $\Delta\Delta$ MFCC + Pitch + Loudness feature vectors are shown in Table.6.10.

Table 6.10: Comparison of proposed model with published work

Model	Language	Feature vector	Training duration / Samples	Test duration	Accuracy
SVM[19]	Malayalam Language	MFCC + TEO	3h 10 min	5 to 10s	78%
GMM[24]	Assamese Language	MFCC	13h 30 min	3 to 7s	85%
HMM[77]	Gujarati Language	MFCC	1hour 38min	3 to 5s	87.23%
GMM[20]	Pashto Language	MFCC	0.40hours	5 to 10s	88.43%
DNN (Proposed model)	Telugu language	Optimized feature vector (MFCC+ Pitch+ Loudness)	6h 25min	3 to 8s	90.12%

It is observed the proposed model with optimized feature vectors which are derived from MFCC+Pitch+Loudness is performed well compared to existing works. The corresponding comparison graph is shown in Fig.6.11.

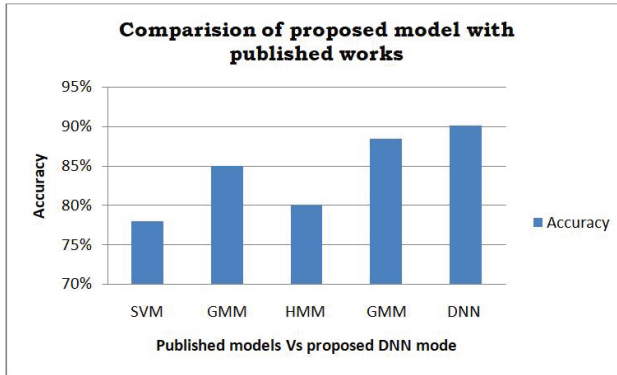


Figure 6.11: Comparison of proposed model with published work

Chapter 7

SUMMARY AND CONCLUSION

7.1 Summary

In this work, Dialect identification system has been implemented using the different feature vectors, extracted from raw speech signal and different modeling techniques. As there is no standard database for Telugu Language, database has been created by collecting and recording the speech samples from different peoples, places and grounds. Data base consists of three dialects of Telugu Language namely Telangana, Costa Andhra and Rayalaseema.

The importance and role of spectral features and prosodic features has been established by modelling HMM, GMM, and DNN. The significant results are achieved using these features to identify dialects from shortest speech utterances of Telugu Language.

The role of Pitch and Loudness has been proved in identifying the dialects efficiently from the shortest duration of the Telugu language.

New feature vectors have been derived by combining spectral features (MFCC) and the Prosodic features (Pitch+ Loudness) and experiments are carried with different modeling techniques. It was proved that, the new feature vector well discriminates the speech utterances among three dialects.

The Optimized feature vectors have been derived in order to remove redundant data and reduce the dimensionality of feature vector. The PCA approach has been used for dimensional reduction, which improves the system's performance in terms of identification rate and fast response in identification. A comparative study of the proposed dialect identification system using HMM, GMM, and DNN with the reputed published work has been carried out. The performance of the system is very impressive.

7.2 Scope for Future Work

1. Need to explore hybrid modeling techniques using SVM, Deep Learning techniques etc.
2. Data base size may be increased by collecting samples from the child age of 6 to 9 also.
3. It is also possible to extract the different feature reduct vectors to reduce the time complexity of the model.
4. The duration of Test utterance might be reduced.
5. The rough set and fuzzy methods may be used in case of feature selection and dimensionality reduction.

Bibliography

- [1] Subhash Chand Samota, Gaurav Kumar Sharma "Speech Signal Processing: A Technical Review" Journal of Engineering and Technology, Vol 5, issue 1, 2019.
- [2] M.A.Anusuya, S.K.Katti "Speech Recognition by Machine: A Review" International Journal of Computer Science and Information Security (IJCSIS) Vol. 6, No. 3, 2009.
- [3] Ashok Kumar, Vikas Mittal "Speech Recognition: A Complete Perspective" IJRTE, ISSN: 2277-3878, Volume-7 Issue-6C, April 2019.
- [4] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh and K. Shaalan, "Speech Recognition Using Deep Neural Networks: A Systematic Review," in IEEE Access, Vol. 7, pp. 19143-19165, 2019.
- [5] Etman, Asmaa & Beex, A. A. (Louis) "Language and Dialect Identification: A survey." 10.1109/IntelliSys.p.p.-7361147, 2015.
- [6] S. Weninger, G. Khan, M. Streck, J. C. E. Watson and Eds.Berlin, Walter de Gruyter, "Arabic dialects (general article)," in The Semitic Languages:An International Handbook (Handbooks of Linguistics and Communication Science (HSK)),2011, pp. 851-896
- [7] M. Venkatalakshamma, N. Munirathnamma "Importance of Andhra Pradesh Mother Tongue-a Study on Telugu Language" Indian Journal of Applied Research, Vol. 4 issue 12 2014.
- [8] B.A. PrabhakarBabu "A Phonetic and Phonological Study of some Characteristic Features of Telugu English Including reference to the Source And Target Languages" University Of London, M.Phil, 1976.
- [9] J VenkateswaraSastry "A Study of Telugu Regional and social dialectsA Prosodic analysis": Department of Phonetics and Linguistic, School of Oriental and African studies, University of London 1989.

- [10] Boersma, Paul & Van Heuven, Vincent "Speak and un Speak with PRAAT", Glot International, 2001.
- [11] Ambalika, Er. Sonia Saini "Speech Analysis in Praat Tool Using Hybrid Filter" IJARECE, Volume 5, Issue 10, October 2016.
- [12] N. D. Londhe and G. B. Kshirsagar "Speaker independent isolated words recognition system for Chhattisgarhi dialect," 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017, pp. 1-6, doi:10.1109/ICIIECS.2017.8276169.
- [13] A power soft "Streaming Audio Recorder for windows" <https://apowersoft-streaming-audio-recorder.en.softonic.com>, 2020
- [14] Rafael C.Gonzalez, Richard E.Woods "Mean Filters: Digital image Processing" p.231-235, 2019.
- [15] Q. Song, L. Ma, J. Cao and X. Han, "Image Denoising Based on Mean Filter and Wavelet Transform," 2015 4th International Conference on Advanced Information Technology and Sensor Application (AITS), 2015, pp. 39-42, doi: 10.1109/AITS.2015.17.
- [16] Sir George Grierson "Linguistic Survey of India" Journal of the Royal Asiatic Society of Great Britain and Ireland, no. 3 1928.711-18.
- [17] Laszlo Czap, Lu Zhao. "Phonetic aspects of Chinese Shaanxi Xi'an dialect", IEEE International Conference on Cognitive Info communications (CogInfoCom), 2017.
- [18] ImeneGuellil, FaicalAzouaou. "Arabic Dialect Identification with an Unsupervised Learning (Based on a Lexicon) Application Case: ALGERIAN Dialect", IEEE Intl Conference on Computational Science and Engineering. DCABES 2016.

- [19] V VSreeraj, Rajeev Rajan. "Automatic dialect recognition using feature fusion", International Conference on Trends in Electronics and Informatics (ICEI), 2017.
- [20] Saud Khan, Haider Ali, Khalil Ullah. "Pashto language dialect recognition using mel frequency cepstral coefficient and support vector machines", International Conference on Innovations in ICIEECT, 2017.
- [21] Suwon Shon, Ahmed Ali, James Glass. "MITQCRI Arabic dialect identification system for the 2017 multi-genre broadcast challenge", IEEE Automatic Speech Recognition 2017.
- [22] Mehrabani, Mahnoosh, and John H. L. Hansen. "Automatic analysis of dialect/language sets", International Journal of Speech Technology, 2015.
- [23] Jacqueline Ibrahim, DessiPuji Lestari. "Classification and clustering to identify spoken dialects in Indonesian", 2017 International Conference on Data and Software Engineering(ICoDSE), 2017
- [24] Mona Abdullah Al-Walaie, Muhammad Badruddin Khan. "Arabic dialects classification using text mining techniques", International Conference on Computer and Applications (ICCA), 2017
- [25] Tanvira Ismail, L. Joyprakash Singh. "Identification of Goalparia dialect and similar languages", International Conference on Multimedia, Signal Processing and communication Technologies (IMPACT), 2017.
- [26] Grabe, E., Kochanski, G. and Coleman, J. The intonation of native accent varieties in the British Isles - potential for miscommunication? English pronunciation models: a changing science. Linguistic Insights Series, Peter Lang, pp. 311-337.2005.

- [27] Abrham Debasu Mengistu and Dagnachew Melesew "Text Independent Amharic Language Dialect Recognition: A Hybrid Approach of VQ and GMM", International Journal of Signal Processing 2017.
- [28] Trang, H., Loc, T.H., Nam, H.B.H. Proposed combination of PCA and MFCC feature extraction in speech recognition system. In 2014 International Conference on Advanced Technologies for (ATC 2014), pp. 697-702.
- [29] Ghosal, A., Chakraborty, R., Chakraborty, R., Haty, S., Dhara, B.C., Saha, S.K. Speech/music classification using occurrence pattern of zcr and ste. In 2009 Third International Symposium on Intelligent Information Technology Application, 3: 435-438.2009.
- [30] B Chittaragi, Nagaratna & Limaye, Asavari & Chandana, N. Basava, Annappa & Koolagudi, Shashidhar "Automatic Text-Independent Kannada Dialect Identification System: Proceedings of Fifth International Conference INDIA 2018 Volume 2.
- [31] Chittaragi, N.B., Prakash, A. & Koolagudi, S.G. Dialect Identification using Spectral and Prosodic Features on Single and Ensemble Classifiers. Arab J SciEng 43, 4289-4302 (2018).
- [32] Chitturi, R. and J. Hansen. "Multi-stream dialect classification using SVM-GMM hybrid classifiers.", IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU), 2007: 431-436.
- [33] Tzudir, Moakala & Sarmah, Priyankoo & Prasanna, S. Dialect Identification Using Tonal and Spectral Features in Two Dialects of Ao, 2018. 10.21437/SLTU.2018-29.
- [34] Zergat, Kawthar & Amrouche, Abderrahmane. New scheme based on GMM-PCA-SVM modelling for automatic speaker recognition. International Journal of Speech Technology, 17.10.1007/s10772-014-9235-7.2014.

- [35] Zissman, Marc & Gleason, T.P. & Rekart, D.M. & Losiewicz, B.L. Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech. 777-780, Vol. 2, 1996. 10.1109/ICASSP.1996.543236.
- [36] Rong Tong, Bin Ma, Donglai Zhu, Haizhou Li and EngSiongChng, "Integrating Acoustic, Prosodic and Phonotactic Features for Spoken Language Identification," 2006, IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, 2006, pp. I-I, doi: 10.1109/ICASSP.2006.1659993.
- [37] Huang, Hai & Huang, Xiao & Li, Ren & Lim, Teik & Ding, Wei. Sound quality prediction of vehicle interior noise using deep belief networks. Applied Acoustics.2016. 113. 149-161. 10.1016/j.apacoust.2016.06.021.
- [38] M. Jain, M. S. Saranya and H. A. Murthy, "An SVD Based Approach for Spoken Language Identification," 2018 International Conference on Signal Processing and Communications (SPCOM), 2018, pp. 312-316, doi: 10.1109/SPCOM.2018.8724477.
- [39] Faragallah, Osama. Robust noise MKMFCC-SVM automatic speaker identification. International Journal of Speech Technology, 2021.10.1007/s10772-018-9494-9.2018.
- [40] [Zissman] M. A. Zissman, "Language identification using phoneme recognition and phonotactic language modeling," 1995 International Conference on Acoustics, Speech, and Signal Processing, 1995, pp. 3503-3506 vol.5, doi: 10.1109/ICASSP.1995.479741.
- [41] Bin MA, Donglai ZHU and Rong Tong, "Chinese Dialect Identification using Tone Features based on pitch flux" ICASSP, pp I1029-I1032,06
- [42] F. S. Alorfi, Automatic Identification of Arabic Dialects Using Hidden Markov Models, Ph.D. thesis, University of Pittsburgh, USA (2008)

- [43] Chittaragi N.B., Limaye A., Chandana N.T., Annappa B., Koolagudi S.G. "Automatic Text-Independent Kannada Dialect Identification System." *Information Systems Design and Intelligent Applications. Advances in Intelligent Systems and Computing*, vol 863. Springer, Singapore, 05 Jan 2019, pp 79-87. Doi: 0.1007/978-981-13-3338-3-8.
- [44] L. R. Arla, S. Bonthu and A. Dayal, "Multiclass Spoken Language Identification for Indian Languages using Deep Learning," 2020 IEEE Bombay Section Signature Conference (IBSSC), 2020, pp. 42-45, doi: 10.1109/IBSSC51096.2020.9332161.
- [45] Kim, Hwamin & Park, Jeong-Sik. Automatic Language Identification using Speech Rhythm Features for Multi-Lingual Speech Recognition Applied Sciences. 2020. 10. 2225. 10.3390/app10072225.
- [46] Reddy VR, Maity S, Rao KS. Identification of Indian languages using Multi-Level Spectral and Prosodic Features. *International Journal of Speech Technology*. 2013; 16(4):489-511.
- [47] Ming CHEN, Lujia WANG "A Novel Approach of System Design for Dialect Speech" Interaction with NAO Robot by, Cheng-zhong XU3, ICAR 2017
- [48] Sadanandam, M. & Prasad, V. Automatic text independent language Identification using reduct set of feature vectors. *IEEE International Conference on Fuzzy Systems*. 1-5.2013.10.1109/FUZZ-IEEE, 2013.
- [49] IBRAHIM, Noor Jamaliah et al. Robust Feature Extraction Based on Spectral and Prosodic Features For Classical Arabic Accents Recognition. *Malaysian Journal of Computer Science*, [S.l.], pp. 46-Dec. 2019 doi:10.22452/mjcs.sp, 2019.
- [50] Tong R, Ma B, Lee KA, You C, Zhu D, Kinnunen T, Sun H, Dong M, ChngES, and Li H "Fusion of acoustic and tokenization features

- for speaker recognition,” in Lecture Notes in Computer Science vol. 42-74 2016 pp. 566–577.
- [51] Chittaragi, N.B., Prakash, A. & Koolagudi, S.G. Dialect Identification Using Spectral and Prosodic Features on Single and Ensemble Classifiers. Arab J SciEng 43, 4289–4302, (2018).
- [52] N. B. Chittaragi and S. G. Koolagudi, "Acoustic features based word level dialect classification using SVM and ensemble methods," 2017 Tenth International Conference on Contemporary Computing (IC3), 2017, pp. 1-6, doi: 10.1109/IC3.2017.8284315.
- [53] Luciana Ferrer et.al. "Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems" Speech Communication, Vol 69. Issue C.M ay 2015. pp: 31-45. <https://doi.org/10.1016/j.specom.2015.02.002>.
- [54] S. Shabani and Y. Norouzi, "Speech recognition using Principal Components Analysis and Neural Networks," 2016 IEEE 8th International Conference on Intelligent Systems (IS), 2016, pp. 90-95 doi: 10.1109/IS.2016.7737405
- [55] Manjushree B. Aithal, Pooja R. Gaikwad, & Shashikant L. Sahare "Speech Enhancement Using PCA for Speech and Emotion Recognition" G.J. E.D.T., Vol.4(3):6-12,2015.
- [56] Shen, Wade & Chen, Nancy & Reynolds, Douglas. Dialect recognition using adapted phonetic models. 763-766,2008.
- [57] P. A. Torres-Carrasquillo and D. A. Reynolds, "Dialect identification using Gaussian mixture models," in proceedings of The Speaker and Language Recognition Workshop in Toledo, Spain, ISCA, pp. 297–300, May 2004.
- [58] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R.J. Greene, D. A.Reynolds, and J. Deller, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral fea-

- tures,” in proceedings of International Conference on Spoken Language Processing, pp. 89–92, September 2002.
- [59] N. F. Chen, et al., "Characterizing Phonetic Transformations and Acoustic Differences Across English Dialects". IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 22 No.1, 2014. pp. 110-124.
- [60] Santosh Gaikwad, Bharti Gawali and K V Kale, Accent Recognition for Indian English using Acoustic Feature Approach. International Journal of Computer Applications 63(7):25- 32, Foundation of Computer Science, New York, USA.2013.
- [61] Qin Yan, Saeed Vaseghi, "A Comparative Analysis of UK and US English accents in recognition And Synthesis" 0-7803-7402,2002.
- [62] Gang Liu and John L. Hansen, 2011. A systematic strategy for robust automatic dialect identification. In EUSIPCO, pp: 2138-2141
- [63] Fadi Biadsy, Julia Hirschberg, 2009. "Using Prosody and Phonotactics in Arabic Dialect Identification" interspeech09.
- [64] Sadanandam, M. HMM based language identification from speech utterances of popular Indic languages using spectral and prosodic features. Traitement du Signal, Vol. 38, No. 2, pp. 521-528. <https://doi.org/10.18280/ts.380232>
- [65] H. C. Soumia Bougrine and A. Abdelali, "Spoken Arabic Algerian dialect identification," 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP), 2018, pp. 1-6.
- [66] Tailor J.H., Shah D.B HMM-Based Lightweight Speech Recognition System for Gujarati Language. In: Mishra D., Nayak M., Joshi A. Information and Communication Technology for Sustainable Development. Lecture Notes in Networks and Systems, vol 10. Springer, Singapore. 2018.

- [67] Ignacio Lopez-Moreno, Javier Gonzalez-Dominguez, Oldrich Plchot, David Martinez, Joaquin Gonzalez-Rodriguez², Pedro Moreno¹ "Automatic Language Identification using Deep Neural Networks" 2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP) pp: 5374-5378
- [68] Kwon, O.W., Chan, K., Lee, T.W. Speech feature analysis using variational Bayesian PCA. *IEEE Signal Processing Letters*, 10(5):137-140.2013. <https://doi.org/10.1109/LSP.2003.810017>
- [69] Abolhassani, A.H., Selouani, S.A., O'Shaughnessy, D. Speech enhancement using PCA and variance of the reconstruction error in distributed speech recognition. In 2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU), 19-23.
- [70] K. Mannepalli, P. N. Sastry and V. Rajesh, "Accent detection of Telugu speech using prosodic and formant features," 2015 International Conference on Signal Processing and Communication Engineering Systems, 2015, pp. 318-322,
- [71] Chadawan Ittichaichareon, SiwatSuksri and Thaweesak Yingthaworn-suk, "Speech recognition using MFCC", International Conference on Computer Graphics Simulation and Modeling, 2012.
- [72] Shrawanka U, Thakare V. Techniques of feature extraction in speech recognition system: A comparative study. *International Journal of Computer Applications in Engineering Technology and Sciences*.2010; 2:412-8.
- [73] T. K. Moon, "The expectation-maximization algorithm," in *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47-60, Nov. 1996
- [74] Q. Legros, S. Meignen, S. McLaughlin and Y. Altmann, "Expectation Maximization Based Approach to 3D Reconstruction From Single-Waveform Multispectral Lidar Data," in *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1033-1043, 2020.

- [75] Magre, Smita & Janse, Pooja & Deshmukh, Ratnadeep A Review on Feature Extraction and Noise Reduction Technique. IJARCSSC, Volume 4, Issue 2, February 2014
- [76] Xuan Huang, Lei Wu and Yinsong Ye "A Review on Dimensionality Reduction Techniques" International Journal of Pattern Recognition and Artificial Intelligence Vol. 33, No. 10, 1950017, 2019.
- [77] Tailor, J. H., & Shah, D. B. (2017). HMM-Based Lightweight Speech Recognition System for Gujarati Language. Lecture Notes in

FOR AUTHOR USE ONLY

LIST OF PUBLICATIONS

Published Papers

1. **S.Shivaprasad**, Dr.M Sadanandam “Speech Based Query Searching Technique And Its Application In Library Management System”, International Journal of Recent Technology And Engineering, Vol.8 September 2019. (SCOPUS)
2. **S.Shivaprasad**, M Sadanandam “Identification Of Dialects: Survey”, International Journal Of Advanced Science And Technology, Vol29 issue 3, 2020. (Scopus)
3. **S.Shivaprasad**, Dr.M Sadanandam “Identification of regional dialects of Telugu language using text independent speech processing models” International journal of speech technology Vol 23, issue 1 pages251–258, 2020. (Springer(ESCI)-SCOPUS)
4. **Shivaprasad Satla**, M. Sadanandam “Dialect recognition from Telugu speech utterances using spectral and prosodic features” International Journal of Speech Technology.(Springer(ESCI)- SCOPUS) DOI:<https://doi.org/10.1007/s10772-021-09854-8>
5. **Satla Shivaprasad**, Manchala Sadanandam “Optimized Features extraction from Spectral and Temporal Features for Identify the Telugu Dialects by Using GMM and HMM” Ingénierie Des Systèmes D Information vol.23 issue 6, 2021.
6. **S.Shivaprasad**, M. Sadanandam “Comparison of Different Feature Extraction Techniques In Telugu Dialects Identification” Turkish Journal of Computer and Mathematics Education Vol.12 No. 9 (2021), 3196-3206.(Scopus)

Accepted Papers

1. Satla Shivaprasad, Manchala Sadanandam, Pranay “Random Search Technique to find the Dialects of Telugu Language” AIP conference proceedings. SRITW

Communicated Papers

1. **Satla Shivaprasad**, Manchala Sadanandam “ New model to identify the dialects o Telugu language: Deep Neural Network”, Treatment du signal.(SCIE)
2. **Satla Shivaprasad**, Manchala Sadanandam “Hybrid Prosodic Feature Extraction Techniques to Identify Dialects of Telugu Language” PATTERN RECOGNITION LETTERS-(SCI)

FOR AUTHOR USE ONLY

FOR AUTHOR USE ONLY

**More
Books!**



yes
I want morebooks!

Buy your books fast and straightforward online - at one of world's fastest growing online book stores! Environmentally sound due to Print-on-Demand technologies.

Buy your books online at
www.morebooks.shop

Kaufen Sie Ihre Bücher schnell und unkompliziert online – auf einer der am schnellsten wachsenden Buchhandelsplattformen weltweit! Dank Print-On-Demand umwelt- und ressourcenschonend produziert.

Bücher schneller online kaufen
www.morebooks.shop

KS OmniScriptum Publishing
Brivibas gatve 197
LV-1039 Riga, Latvia
Telefax: +371 686 20455

info@omniscryptum.com
www.omniscryptum.com

OMNIScriptum



FOR AUTHOR USE ONLY