SPEECH PROCESSING APPROACHES TOWARDS CHARACTERIZATION AND IDENTIFICATION OF DIALECTS

Thesis

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

by

NAGARATNA B. CHITTARAGI

(155112 CS15F09)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA,

SURATHKAL, MANGALORE - 575025

August 2020

This thesis is dedicated to my beloved parents and husband

DECLARATION

by the Ph.D. Research Scholar

I hereby declare that the research thesis entitled **SPEECH PROCESSING APPROACHES TOWARDS CHARACTERIZATION AND IDENTIFICATION OF DIALECTS** which is being submitted to the **National Institute of Technology Karnataka, Surathkal** in partial fulfilment of the requirements for the award of the Degree of **Doctor of Philosophy** in **Computer Science and Engineering** is a *bonafide report of the research work carried out by me.* The material contained in this research thesis has not been submitted to any university or institution for the award of any degree.

(155112 CS15F09, Nagaratna B. Chittaragi)
Department of Computer Science & Engineering

Place: NITK, Surathkal.
Date:

CERTIFICATE

This is to *certify* that the research thesis entitled **SPEECH PROCESSING APPROACHES TOWARDS CHARACTERIZATION AND IDENTIFICATION OF DIALECTS** submitted by **Nagaratna B. Chittaragi**, (Register Number: CS15F09) as the record of the research work carried out by her, is *accepted as the research thesis submission* in partial fulfilment of the requirements for the award of degree of **Doctor of Philosophy**.

Dr. Shashidhar G Koolagudi
Research Guide

Dr. Alwyn Roshan Pais
Chairman - DRPC

Acknowledgment

Writing the note of gratitude is the finishing touch on my dissertation. Research has been a period of intense learning for me, not only in the technical domain, but also on a personal level. I would like to reflect on the people who have supported and helped me so much throughout this period.

I would first like to thank and express my sincere gratitude to my supervisor Dr. Shashidhar G. Koolagudi for the constant guidance, support and encouragement at all stages of my research work. I am obliged for being introduced me to the area of speech, language processing and machine learning. I am especially thankful to my supervisor for giving me an opportunity to pursue this research under his guidance and also for patiently checking all my manuscript and thesis. I thank him for the excellent research environment he has created for all of us to learn.

I would like to express my heartfelt thanks to my Research Progress Assessment Committee (RPAC) members Dr. A. V. Narasimhadhan and Dr. Manu Basavaraju for their valuable suggestions and constant encouragement that consistently helped me in improving the research work. I would also like to thank Dr. Alwyn Roshan Pais, Head of the department for his help and support in carrying out research work. My special thanks to Prof. Annappa B. for his support during my work and presenting our paper in an international conference. I would like to mention the concern and encouragement shown to me by Prof. Santhi Thilagam and Dr. Soumya Hegde. I sincerely thank all teaching, technical and administrative staff of the Department of Computer Science and Engineering, NITK, for their help during my research work.

I am highly grateful to Siddaganga Institute of Technology (SIT), Tumkur, Karnataka, for deputing me to pursue Ph.D. at NITK and providing me the financial assistance. I am thankful to Prof. Jayanna H. S. and Prof. Aparna for their motivation, help, and moral support given to me to pursue research at NITK.

I will forever remember the wonderful time I have had with my friends and labmates Y. V. Srinivasa Murthy, Manjunath Mulimani, Pravin V. Ramteke, Fathima, Keerthi S. Shetty, Bheemappa H., Girish, Pradyoth Hegde and Siva Krishna P. M. I thank Manjunath for all his instant support and immense help on technical issues and machine learning concept doubts throughout my research work. I thank my lab-mates Keerthi, Rashmi, Apoorva, and Umapriaya for the wonderful moments we shared together. I thank all the past and present lab-mates for the friendly and conductive atmosphere in lab.

Last but not the least, a special thanks to my family. Words cannot express my sincere gratitude towards my father Basavaraj M. Chittaragi, mother Rachamma and husband Dr. Shankarappa H. Yadahalli. I am highly obligated to my parents and husband for all sacrifices they made for my better future and giving me freedom to take my own decisions. My engineering education would have not been completed without their love, motivation and support. The blessings and steadfast support of parents and parents-in-law always strengthened me. I never forget the timely help and extended support of my husband for all my works. I would also like to thank my sisters Dr. Vidya and Dr. Kavitha for supporting and encouraging me throughout this experience. Finally, my special thanks to my sons Nishanth and Nischith for their cooperation and incredible patience throughout the period of my study.

Finally I thank the almighty, for giving me the chance, the courage and the strength to endure and pursue till the end. You are the one, I always looked up to, for all my joys and trails. My trust in you will always stay firm. Thank you, God.

Place: Surathkal Nagaratna B. Chittaragi
Date:

Abstract

Dialects constitute the phonological, lexical, and grammatical variations in the usage of a language with very minor and subtle differences. These variations are mainly due to specific speaking patterns followed among the group of speakers. In the recent past, dialect identification from the speech is emerging as one of the prominent speech research areas. This is mainly due to the extensive increase in the use of interactive voice-based systems. Therefore, it is essential to address speech variabilities caused due to dialectal differences in order to achieve effective, realistic man-machine interaction. The existing research on characterization and identification of dialects has mainly focused on acoustic, phonetic and phonotactic approaches on several languages such as English, Chinese, Arabic, Hindi, Spanish, etc. However, these models are not proved to be language independent. Applying these models to other languages may not perform equally well as there are many fundamental differences between dialects of different languages. However, in the literature dialect processing models reported with respect to Indian regional languages are considerably less. In this thesis, an attempt is made to develop few useful language independent and dependent Automatic Dialect Identification (ADI) systems for the Kannada language. In the beginning, a new text-independent Kannada Dialect Speech Corpus (KDSC) is collected from native speakers belonging to five prominent dialectal regions of Karnataka.

This thesis investigates the significances of the excitation source, spectral, and prosodic features of speech for dialect identification. Additionally, spectro-temporal variations across dialects are captured through 2D Gabor features which are known to be biologically inspired ones. Further, the existence of non- conventional dialect-specific rhythmic and melodic correlations among dialects are explored using chroma features. These are well-established features in music-related applications. Robustness of these proposed features has been investigated under noisy background conditions and with small sized (limited data) audio clips. In addition, word and sentence based ADI systems are proposed using intonation and intensity variations representing the dynamic and static prosodic behaviors.

Further, language dependent dialect identification systems are proposed for Kannada language using basic phonetic unit level dialect information. Additionally, Kannada language specific ’*case*’ (Vibhakthi Prathyayas) based dialect identification approaches are proposed. A single classifier based Support Vector Machines (SVM) and multiple classifiers based ensemble algorithms are used for classification of dialects. Experiments are carried out using individual and combinations of features. Use of different features has illustrated their complementary nature towards dialect processing. Performance comparison of both categories of classification algorithms has shown that ensemble algorithms perform better over single classifier based algorithms. Further, the intuition to use rhythm based aspects of dialects through chroma and spectral-shape features has shown better performance over state-of-the-art i-vector features. Moreover, this feature set has shown the noise robustness over the conventional MFCCs. In this work, we also have proposed intonation and intensity features to capture dialectal information from words and sentences for effective classification of dialects. In continuation, the role of duration, energy, pitch, three formants, and spectral features is also found to be evidential in Kannada dialect classification.

Keywords: Kannada dialect identification, Spectral features, Prosodic features, Excitation source features, Spectro-temporal features, Chroma features, Spectral-shaped features, Dynamic and static features, *Cases*, Support vector machine, Random forest, Extreme random forest, Extreme gradient boosting.

Contents

List of Figures

List of Tables

Abbreviations

AANN	Auto Associative Neural Network
ABI	Accents of British English
ADI	Automatic Dialect Identification
ANN	Artificial Neural Network
ANOVA	Analysis and Variance
ASR	Automatic Speech Recognition
CENK	Central Kannada
CSTK	Coastal Kannada
CNN	Convolutional Neural Network
COCOSDA	Committee for the Coordination and Standardization of Speech Databases and Assessment Techniques
CV	Cross Validation
DCT	Discrete Cosine Transfer
DNN	Deep Neural Networks
ERF	Extreme Random Forest
ESVM	Ensemble Support Vector Machine
F0	Fundamental Frequency
F1	First Formant Frequency
F2	Second Formant Frequency
F3	Third Formant Frequency
F4	Fourth Formant Frequency
FFNN	Feed Forward Neural Network
FFT	Fast Fourier Transfer
GMM	Gaussian Mixture Model

GVV	Glottal Volume Velocity
GFD	Glottal Flow Derivative
HCI	Human Computer Interaction
HMM	Hidden Markov Model
HYDK	Hyderabad Kannada
IF	Instantaneous Frequency
IIITH	International Institute of Information Technology, Hyderabad
IViE	Intonational Variations in English
JFA	Joint Factor Analysis
LDA	Linear Discriminant Analysis
LID	Language Identification
LP	Linear Prediction
LPCC	Linear Prediction Cepstral Coefficients
LSP	Line Spectral Paris
LSTM	Long Short Term Memory
KDSC	Kannada Dialect Speech Corpus
MEB	Minimal Enclosing Ball
MFCC	Mel Frequency Cepstral Coefficients
MLP	Multilayer Perceptron
MUBK	Mumbai Kannada
NPS	Nationwide Speech Project
NTIMIT	Network Texas Instruments-Massachusetts Institute of Technology
OED	Oxford English Dictionary
PLPC	Perceptual Linear Prediction Coefficients
PMVDR	Perceptual Minimum Variance Distortion-less Response
PPR	Parallel Phone Recognition
PPRLM	Parallel Phone Recognition-Language Modeling
PRLM	Phone Recognition-Language Modeling
QCRI	Qatar Computing Research Institute
RBF	Radial Basis Function

RASTA	Relative Spectra
RF	Random Forest
RM1	Resource Management
RMFCC	Residual Mel Frequency Cepstral Coefficients
RNN	Recurrent Neural Network
SARA	Spoken Arabic Regional Archive corpus
SDC	Shifted Delta Coefficients
SNR	Signal-to-Noise-Ratio
SOE	Strength of Epochs
SSOE	Strength of Slope of Epochs
STHK	Southern Kannada
SVM	Support Vector Machine
VOT	Vocal Onset Time
XGB	Extreme Random Forest
TIMIT	Texas Instruments-Massachusetts Institute of Technology
TV	Total Variability
UBM	Universal Background Model
ZCR	Zero Crossing Rate
ZFF	Zero Frequency Filter
ZFFS	Zero Frequency Filtered Signal

CHAPTER 1

Introduction

This chapter in brief, contains the significance of dialects from linguistic and speech signal perspectives. General approaches used in the research community for addressing the issues in dialect processing are fairly introduced. Importance of prevailing datasets is also touched upon. An attempt has been made to throw some light to take up this task as a research problem, important challenges associated, scope and contributions of the present work. Further, there are seven chapters in this thesis that deal with different approaches which have been proposed for dialect processing. And, thesis concludes with summary, learning outcome and specific future research directives.

Speech can be one of the natural modalities for human-machine interaction. Now-a-days, due to the extensive use of voice-controlled interface devices, there is an increased demand for developing an efficient and robust automatic speech processing system. However, natural speech is a complex signal combined with an abundance of diversified information such as emotions, gender, age, speaker related evidence, accents, dialect details and so on. Among these, next to gender factor, dialectal differences are the primary cause of speech variabilities that cause heavy degradation in the performance of automatic speech processing systems. (Biadsy et al., 2011). Now-a-days, Automatic Dialect Identification (ADI) systems are gaining more popularity and research attention among active speech researchers. Dialects commonly represent the different speaking patterns observed among groups of native speakers belonging to some particular geographical region. Qualitative speech traits such as tonality, loudness, durations, and nasality also contribute to dialectal variations. These variations exist basically because of differences in geographical locations, socioeconomic status, cultural and education background, influence of neighboring languages, caste, ethnicity, mother tongue and so on (Chambers and Trudgill, 1998).

1.1 Dialects: Linguistic Perspective

The Oxford English Dictionary describes dialect as the subordinate forms or varieties of a language arising from local peculiarities of vocabulary, pronunciation, and idioms. Study of linguistic properties of dialect, along with geographic and social distribution, is known as *Dialectology* (Trousdale, 2010). Linguistic aspects of dialects primarily focus on considering language-specific variations that exist across dialects. In this case, characterization and identification of dialects would be beneficial if linguistic knowledge is available in advance. Dialects of a language can be distinguished based on lexical variations (difference in vocabulary), grammatical variations (difference in the construction of phrases and sentences) and phonological variations (difference in pronunciation). Dialectal differences may be observed from several levels of linguistic hierarchy (e.g., acoustics, phonetics, prosody, vocabulary, syntax) (Chen et al., 2010; Rao and Koolagudi, 2011). Figure 1.1 illustrates various language/dialect specific cues and their levels of manifestation in speech. Every language has its own lexicon set or character set equivalent to alphabets. The lexical grammar suggests the rules and syntactics for usage of character sequence. Each language has its own vocabulary with its own way of forming words. Usually the vocabulary is developed with person's age and it represents the set of familiar words within the person's language. Even though two languages share a common word, the set of words that may follow or precede a word may be different. Proper identification of dialects is possible using speaker-specific cues existing at a lower level and from the higher-level speech/language specific features. Low-level cues are directly derivable from a given speech signal whereas, high-level cues need to be derived from the textual content (transcriptions). Therefore, transcription of a speech is essential to extract high-level cues.

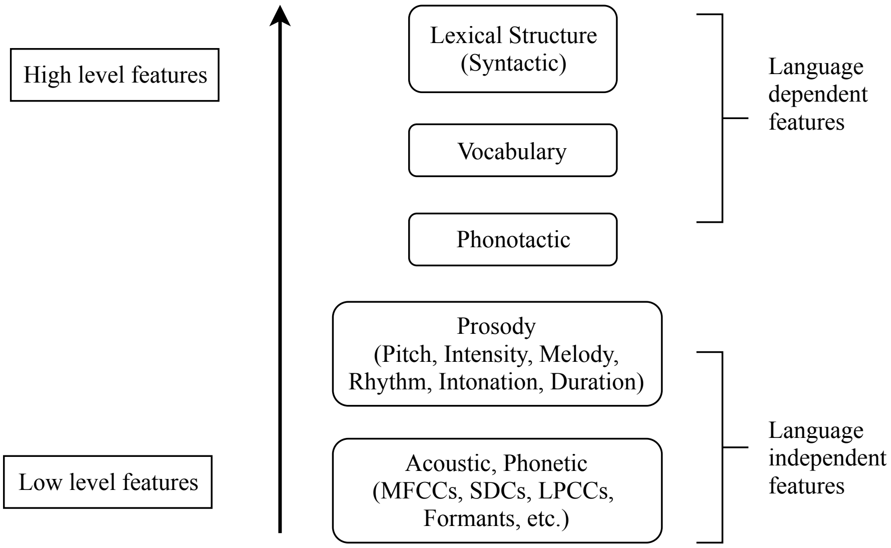

High level features
Low level features
Lexical Structure (Syntactic)
Vocabulary
Phonotactic
Prosody (Pitch, Intensity, Melody, Rhythm, Intonation, Duration)
Acoustic, Phonetic (MFCCs, SDCs, LPCCs, Formants, etc.)
Language dependent features
Language independent features

Figure 1.1: Various language/dialectal cues and their levels of manifestation

Nevertheless, dialect variations are prominently observed in phonetics, acoustics, prosodic level and with a minor deviation in use of standard vocabulary across dialects. At higher levels, the sentence pattern and grammar are generally different between languages also between dialects (Curzan, 2013).

Dialects of a language include variations even at the level of phonological features. That is, the existence of dialectal differences may be observed at various phonetic units such as phoneme (vowels and consonants), syllable, word, sentence, and phrases. Majority of the times, dialectal differences may exist only in pronunciation patterns with addition, deletion or replacement of phonemes or syllables. During transformation of a phone into the other, the phenomena like elimination and/or addition of new phones are observed among English dialects. Pronunciation of ”r”, when followed by a consonant is observed to be different; in some dialects as ”park” (”r” is clearly pronounced) and some other times as ”park” (”r” is silent) (Chen et al., 2010). Many similar such phenomena happen in Kannada language also which are unexplored. For instance the word ”banniri” (means come) will be pronounced as ”barri” or ”banri” in few dialects indicating the elimination of middle phone ”i” in the pronunciation and replacement of phone ”n” with phone ”r”. Many such differences in pronunciation can be observed among dialects.

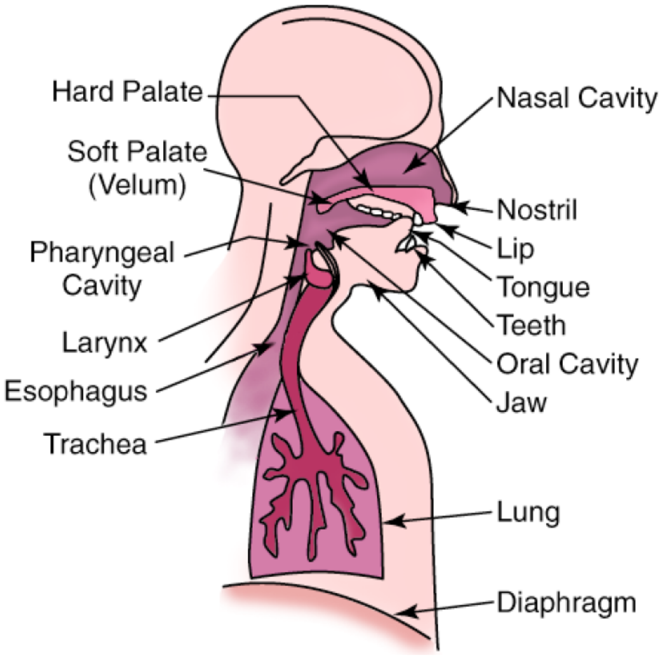


Figure 1.2: Schematic view of human speech production system (Flanagan, 2013)

Some times, certain words, quite often used and vary across dialects can model dialects effectively. It is observed that the written form remains unchanged across dialects of almost all languages except few (Hansen and Liu, 2016). Generally, dialects of a language are said to be mutually intelligible where speakers of the different dialects of the same language can understand and communicate to a larger extent. In this regard, a systematic study of dialects of any language requires a better understanding of linguistic properties of that language (Ma et al., 2006; Purnell and Magdon-Ismail, 2009).

1.2 Dialects: Speech Perspective

Production of speech units involves unique articulatory activity in the vocal tract, along with the excitation. The air pressure created in the sub-glottal region passes through the vocal tract and nasal cavity in the form of air puffs causing vocal fold vibration also called as glottal activity (Kodukula, 2009).

Figure 1.2 shows the schematic view of the speech production system of human beings. The air pressure created in the trachea (sub-glottal region) gets released into the vocal tract and nasal cavity, in the form of air puffs. The pressure created beneath the vocal folds causes their vibration, known as glottal activity. Sequence of quasi-periodic air pulses are resulted due to the chopping of air flow when passing through vibration of vocal folds. These impulse-like excitation sequences get transformed into varying frequency components when passing through the oral and nasal cavities. Here, vibration of vocal folds are the major source for excitation to the vocal tract system. The oral cavity acts as a time-varying resonator with simultaneous controlled movements of articulatory organs in the vocal tract (lips, tongue, jaws, and velum). Movement of quasi-periodic sequence of air puffs through vocal and nasal tract produces the speech which gets radiated from the lips. Voiced (involving vocal folds vibration) and un-voiced (constriction caused at different places in the vocal tract) sound units are produced due to physiological characteristics and the co-articulation of different articulatory organs (Rabiner and Juang, 1993). The characteristics of vocal tract shapes and glottal activity play a major role in modulating different languages and dialects during the production of speech.

Speech is primarily intended to convey some message. A message is carried through a sequence of legal sound units with different combinations across languages. Dialectal differences may also exist due to acquired habits. These are characteristics that are learned over a period of time, mostly influenced by the social environment and also by the characteristics of the first or native language in the critical period (lasting from infancy to puberty) of learning (Curzan, 2013). Existing studies have revealed the existence of dialectal differences across spoken units in terms phenomenon like pitch, duration, stress patterns, intensity, voice onset time, voiced stop release times, acoustic phoneme space, etc. These features play a significant role in identification of dialects (Arslan and Hansen, 1996). However, many times it may be often difficult for an individual to describe attributes used in dialect recognition (Etman and Beex, 2015). The language, dialect and speaker-specific information, contained in the speech signal is inferred using features at several levels. Literature shows that dialect-specific speech features may be extracted from excitation source patterns, vocal tract shapes, prosodic (supra-

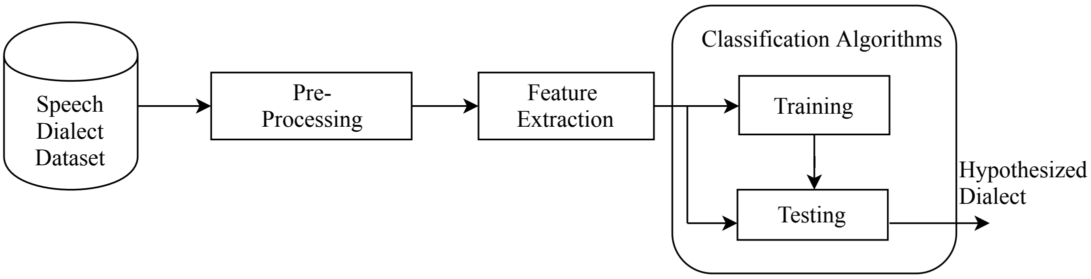


Figure 1.3: Basic model for dialect recognition

segmental) behavior, spectro-temporal attributes, linguistic features and so on.

1.3 Overview of Automatic Dialect Identification (ADI) Systems

Automatic dialect identification task addresses the dialect recognition problem from the uttered speech of a particular language. Majority of research reports on ADI have concentrated mostly on languages such as English, Chinese, Japanese, Dutch, Arabic, Spanish and so on (Lei and Hansen, 2011; Zhang and Hansen, 2018). In Indian context, we observe a multicultural and multilingual society. As per the reports in the literature, dialect recognition research activities are found with only a few widely spoken languages like, Hindi, Assamese, Tamil, Telugu, Punjabi, Kannada, Bengali, etc. (Rao and Koolagudi, 2011; Sarma and Sarma, 2016; Soorajkumar et al., 2017) implemented with very basic ADI systems. However, advanced research initiatives are still in their nascent stage for many Indian languages. Many of the existing approaches reported on ADI are language-specific (Biadsy, 2011). They may not be valid to be used for other languages due to many language-dependent dialectal cues. Hence, it is necessary to develop an individual system based on the identification of significant dialectal cues for each language. In this thesis, the main focus is given on the development of speaker independent, text-independent, language independent ADI system especially for South Indian language 'Kannada'. Performance of the proposed ADI system is also tested on the internationally available English dataset. A workflow diagram of a standard ADI system is given in Figure 1.3. The phases in sequence are explained in the following subsections.

1.3.1 Speech Dialect Datasets

Majority of the times, quality, type, environment, and size of the datasets used in the evaluation of speech based systems, have larger implications on the assessment of the performance of a developed system. The procedure followed, type, and size of a dataset highly vary based on the need of the speech systems to be developed. Some of the important ways of the speech corpora collection for different speech tasks are through spontaneous recording, reading, telephonic conversation, TV shows, interviews formats etc. (Zissman et al., 1996).

In general, dialects of a language carry different dialectal cues namely, variations in pitch (intonation), intensity, stress patterns, speaking rate and speaking style, duration, and so on. However, these cues are clearly observed in speech only when a person speaks spontaneously without any written material or prior preparation (Ali et al., 2015). Most of the available speech datasets are collected in a read mode preferably with prior training, which may not be significantly useful in dialect processing (Clopper and Pisoni, 2006). Adequate sized speech corpora with a variety of samples and a good amount of naturalness are essential in measuring the performance of the dialect recognition systems. The literature presents a large number of speech corpora of different variety proposed for dialect processing (Etman and Beex, 2015). Majority of them are collected for the dialects of widely spoken languages. However, a very few datasets with natural and spontaneous speech are available in the context of Indian languages.

1.3.2 Pre-Processing

In the development of any speech processing system, pre-processing is the preliminary step carried out for preparation of speech data for further processing. In general, silence removal for voice activity detection and noise removal for reducing background and ambient noise are two major tasks performed during preprocessing. Pre-emphasis will be performed for boosting the high frequencies, so that higher formants are clearly noticeable. Sometimes, recorded raw speech samples are first converted from stereo to mono sounds for easier processing. This stage also includes steps to remove unnecessarily prolonged long pauses, where, short and relevant pauses are retained to ensure necessary naturalness and intelligibility of speech.

1.3.3 Features Used

The characteristics of both, vocal tract system and excitation source are embedded in the speech signal. The features used in general, for majority of the speech tasks are briefly discussed in the subsections to follow.

A Excitation Source Features

During the production of speech, vibration of vocal folds provides impulse-like excitation to the vocal tract system in the form of air chunks. Voiced speech is produced due to these quasi-periodic air pulses which act as a primary source of excitation. Similarly, unvoiced speech production involves the continuous exhalation of air getting restrictions later at various places in the vocal tract causing different categories of sounds (Kodukula, 2009). Speech is produced by the convolution of excitation of source and vocal tract system response. Excitation source information alone may be estimated through inverse filtering (Mary and Yegnanarayana, 2008). The obtained noise like signal is also known as Linear Prediction (LP) residual. This random noise like signal contains mainly fundamental frequency details of a speaker. Vocal tract system mainly exhibits lower order relations present in the speech signal which can be observed from adjacent and nearly adjacent samples of speech signal. Noise like LP residual exhibits higher ordered relations which may not be prevalent in adjacent samples and are difficult to obtain using normal available signal processing techniques. Glottal activity leading to vibration of glottal folds, concepts of epochs (Glottal Closure Instants(GCIs)), Glottal Volume Velocity (GVV), Instantaneous Frequency (IF), Strength of Epochs (SOE), slope of epochs are commonly used excitation source features reported in the literature. These features are said to carry complementary information compared to vocal tract features and hence can be supportively used for identifying the language, speakers, emotions, and dialects (Choudhury et al., 2018; Nandi et al., 2017; Mary and Yegnanarayana, 2008).

B System Features

The sequence of impulse-like excitation resulted due to vocal fold's vibration acts as a primary stimulus to the vocal tract system. The production of different sound units involves a distinguishable sequence of vocal tract shapes since the vocal tract system is a cascade of cavities of different cross-sections. These shapes can be captured in terms of spectral envelope resembling the vocal tract system characteristics for different sound units (Kodukula, 2009). Similarly, while speaking different vocal tract shapes are observed due to co-articulation effects especially in different dialects (Rao and Koolagudi, 2011; Chittaragi et al., 2018b; Huang and Hansen, 2007).

Formants are nothing but the resonances of the vocal tract that are unique to each sound units. To obtain them, frequency domain analysis of the speech signal is carried out with the units of size 20-30 ms, with a shift of 10 ms known as frames. Average LP spectra are of the vowel /e/ uttered by a randomly selected speaker in five Kannada dialects are shown in Figure 1.4. It may be observed from spectra that vowel /e/ exhibits distinct spectral properties indicating the significant dialectal differences. Variations are observed in energy levels, spectral peaks, spectral sharpness and positions of formant frequency values (F1-F4) across five dialects of Kannada language. The length, shape and dynamics of the vocal tract system vary while speaking in different dialects. Parameterization techniques namely, LPCCs, MFCCs , PLPCs and their derivatives are used to capture these dialect specific vocal tract characteristics (Rabiner and Juang, 1993).

C Prosodic Features

Prosodic features impose auditory qualities and naturalness to the spoken units than merely conveying a textual message. In fact, the imposition of prosodic variations such as intensity and pitch patterns, intonations (rise and fall of pitch), duration, rhythm, melody, and loudness features make speech more natural(Mary and Yegnanarayana, 2008; Rouas, 2007). These features are captured from longer

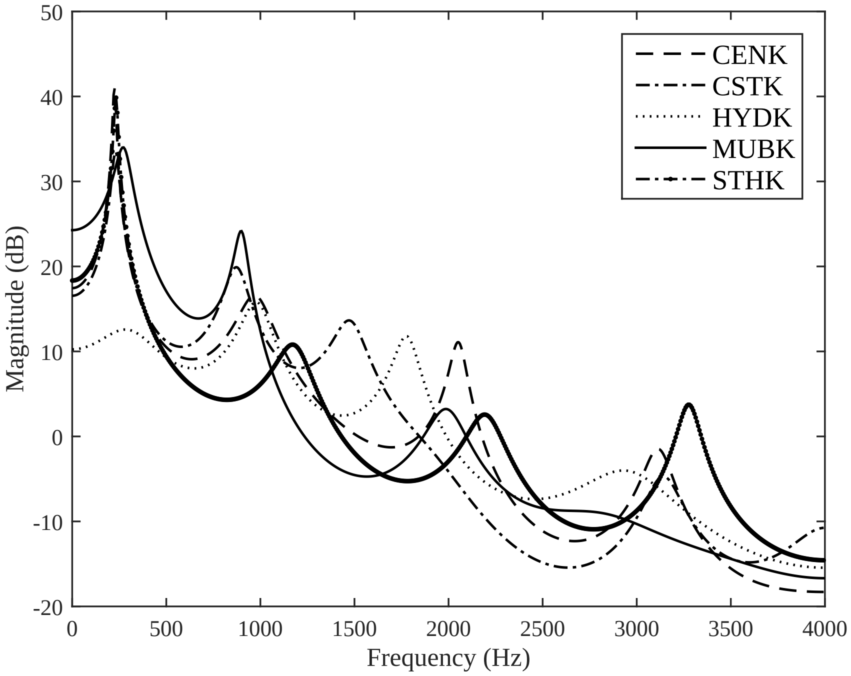


Figure 1.4: Average LP spectra of vowel /e/ in five Kannada dialects, CENK-Central Kannada, CSTK-Coastal Kannada, HYDK-Hyderabad Kannada, MUBK-Mumbai Kannada and STHK-Southern Kannada

speech segments such as, phonemes, syllables, words and sometimes from entire sentences. In the literature, pitch, duration, and energy are identified as primary prosodic attributes. These features are also called as supra-segmental features and acoustically correlate to intonation, rhythm and stress respectively (Ramus and Mehler, 1999).

It is observed from the dataset that different intonation and rhythmic patterns, along with syllable and word level stress practices distinguish Kannada dialects. Overall these three prosodic attributes and their derivatives render naturalness to speech and assist in characterizing the language, dialects, emotions and so on (Chittaragi et al., 2018b; Chittaragi and Koolagudi, 2018).

1.3.4 Classification Algorithms

In the recent past, wide variety of machine learning algorithms have been employed by different researchers for developing of automatic speech-based systems like speech recognition; speaker identification; language, dialect, & emotion processing and so on. In the literature, 'dialect classification' problem is addressed by using a few standard pattern recognition methods like, Gaussian mixture models(GMM), hidden Markov model (HMM), linear discriminant analysis (LDA), support vector machines (SVM) and artificial neural network (ANN) to name a few (Rao and Koolagudi, 2011; Biadsy et al., 2011). In general, these are single classifier based methods that use statistical (probabilistic) or rule-based approaches for classification. Recently, ensemble methods designed by combining the predictions of several individual classifiers have shown a new research direction to classification problems (Dietterich, 2000a). Ensemble algorithms are expected to improve the predictive performance of classifiers since performance relies on the decisions made by multiple individual classification algorithms. Rotation forest, AdaBoost, random forest, gradient boosting methods are a few among widely used ensemble systems (Huang et al., 2007; Chittaragi et al., 2018b).

1.4 Applications

The systems that are capable of characterizing and identifying dialects would supply valuable inputs to the process of improving the performance of interactive speech systems. Dialectal traits are essential factors in degrading the performance of ASR and Human-Computer Interaction (HCI) systems (Ferragne and Pellegrino, 2007). ADI can be useful in modeling subsystems of ASR such as pronunciation modeling, acoustic, phonetic and language models including lexicons adaptation. The subsystems with the component of dialect processing can considerably improve the performance of ASR and HCI systems that use natural speech as data (Najafian et al., 2014).

Natural language recognition systems quipped with the module on ADI responds to interactive telephone response systems more effectively. In recent years, due to a remarkable escalation in Internet usage, there is an increased demand for interactive voice response based interfaces which are yet to be more realistic in terms of human-computer communication. ADI may be used as an interpreter in call centers for an active region based customer call attention (Zissman et al., 1996). Retrieving and processing of historical spoken documents can be assisted with proper dialect recognition system, to make them more useful and relevant. Apart from these, dialect identification benefits nativity identification, medical applications, entertainment media, and so on (Gray and Hansen, 2005). Extensive research findings are available in the literature for speaker profiling, speaker recognition and verification techniques applied to forensic applications. However, they are not supported by suitable dialect processing module (Brown, 2015). As such, speaker profiling deals with capturing linguistic and paralinguistic cues from unknown speakers. Information such as age, gender, language, dialect, emotional state, ethnicity, geographical, and socio-economic status of the speaker, is drawn from the speech input (Kulshreshtha et al., 2012).

Further, dialect identification systems are applicable to practical speech-to-text conversions, spoken document retrieval, spoken language translation, and in dialogue processing systems (Li et al., 2013). In the context of immigration screening, it may also be useful in nativity verification. ADI systems are also beneficial in real life applications such as entertainment, telemedicine, e-health providing medical assistance for old-age people, e-learning, etc.

1.5 Issues and Challenges in Development of Dialect Identification Systems

In this section, a few of the important research issues, pertaining dialect processing through speech, have been discussed.

- The word '*dialect*' has an uncertain meaning and interpretation being interchangeably is used with an '*accent*'. However, *accent* represents variations in only speaking styles, whereas, dialects show variations in pronunciation styles, vocabulary usage, and grammatical constructs. Hence, the word '*dialect*' should not be confused with an '*accent*' or the colloquial forms of a language where every individual has his/her speaking style.
- Identification of dialects from speech signal without knowing the phonotactic rules, the syntax of phoneme and syllable structures and morphological rules, is a challenging task.

- There are no standard dialect-specific speech corpora available for majority of the Indian languages collected with spontaneous speech. Most of the existing datasets that have been recorded have read speech with prior training from a limited number of speakers. However, spontaneous speech recorded from native speakers have better and clear dialectal cues. Sometimes dialectal cues are even gender specific.
- Dialect processing is more challenging than language processing as dialects are perceptually, syntactically, and structurally more similar due to similar linguistic properties. Also, drawing a clear, boundary between dialects is a challenging issue for any given language as the phenomenon is not clearly defined by linguists (Hansen and Liu, 2016).
- Pseudo-dialectal variations can also be observed because of the text spoken, prevailing emotions, and contextual information apart from speaking styles. Hence, it is necessary to explore the evidence that is independent of these parameters while developing dialect recognition systems.
- Dialectal variations may subsist at several levels among spoken units, namely phonemes, syllables, words, and sentences. For majority of the languages, segmentation of continuous or read speech into separate speech units is a challenging and time consuming tedious task.
- Identification of appropriate dialect-specific features those that efficiently discriminate different dialects of the language is not so easy and straight forward task. Based on the properties of data suitable classification approach is also to be chosen.
- Sophisticated ADI systems are expected to demonstrate robustness in noisy conditions which are true in most of the real-life situations.

1.6 Objectives and Scope of the Work

The primary objective of the present work is to develop a robust and sophisticated dialect recognition system taking Kannada language as the case study. This work focuses on deriving various dialect-specific features from sub-segmental, segmental, and supra-segmental level information from the speech signal. Each level carries a distinct set of dialect-specific cues. Excitation source information is used to capture dialectal cues from the glottal activity and exhalation aspects. Spectral features are extracted through frame-wise processing of vocal tract system. In this work, spectro-temporal variations are captured through 2D Gabor features for classification of dialects. Apart from these, chroma features which are associated with rhythm relevant aspects of speech are used for differentiation of dialects. This is due to an assumption that dialects of a language are correlated with musical aspects in terms of melody, rhythm, intonation, and intensity variations. Prosodic features such as pitch, intensity, duration, and intonation variations are extracted to capture dialectal differences as literature shows that prosodic features effectively correlates to dialects. The dialect-specific study has been conducted to analyze the contribution of global and local features. It is observed that the dialectal difference may also be significantly extracted from shorter speech units such as words, sentences, vowels, consonants, etc. This work proposes an ADI system that uses word and sentences to distinguish dialects. This also explores phonemes particular to Kannada vowels and consonants to identify dialects using spectral and prosodic features.

To classify dialects, multiple classifier based ensemble-based classification algorithms are employed. The work presented in this thesis, confines its scope to explore various speech features and classification models to capture the dialect-specific cues. Language dependent phonotactic approaches are not processed due to unavailability of transcriptions for the newly recorded Kannada dialect dataset.

1.7 Contributions of the Present Work

Contributions made in this thesis are two-folds and are here. To begin with language independent ADI systems have been developed using general dialect-specific features. Further, efforts are made towards Kannada dialect identification, using language dependent features.

- This thesis includes a comprehensive analysis of the existing literature on speech dialect identification from source, system, and prosodic aspects. Including a review on dialect datasets and various classification methods employed for ADI.

- A new text-independent spontaneous dialect speech corpus for five dialects of Kannada language has been developed.
- Regular spectral, prosody, and source level excitation features have been extracted from speech signal for classification of the dialects.
- Spectro-temporal variations among the dialects have been captured through biologically-inspired 2D Gabor features for efficient dialect identification.
- Sometimes specific styles seem to be more rhythmic and musical. Chroma features combined with spectral shape based features are extracted for dialects characterization.
- Dialectal information that subsists at the phoneme, word, and sentence level utterances is explored through dialect-specific spectral and prosodic features and are extracted from the specific speech units.
- Single classifier based support vector machines and multiple classifier based ensemble techniques are used to study various classification approaches during dialect identification.
- This work includes the study of the contribution of vowel and consonantal cues in identifying Kannada dialects. Acoustic-phonetic features are extracted from vowels to evaluate dynamic and static behavior of vowels across dialects. Further, significances of each feature in dialect discrimination is analyzed. Statistical analysis of spectral attributes of vowels is performed using Single Factor-ANOVA (Analysis of Variances) tests.
- Special morphological operations those exist in the Kannada language in terms of various *cases* (Vibhakthi pratyayas) commonly called as a grammatical function of a noun or pronoun are considered for classification of Kannada dialects.

1.8 Organization of the Thesis

The contents of the thesis are organized into seven chapters. The brief inclusions are discussed below.

- **Chapter 1: Introduction:** This chapter deals with the need and importance of dialect identification. In general, characteristics of dialects of any language from both linguistic and speech perspective are discussed. Necessary basic steps required in the development of ADI are briefly mentioned. Some important challenges that are generally face during the development of ADI along with day-to-day applications are touched upon. Main objectives and scope of this thesis are also discussed in brief. The chapter ends with a mention of significant research contributions of the thesis and chapter-wise organization.
- **Chapter 2: Dialect Identification: A Review:** This chapter comprises of a detailed review of the existing research works related to the dialect identification. The chapter contains a critical review of various dialect-specific features reported in the literature, different types of dialect corpora used in the literature, machine learning algorithms employed for dialect recognition and their suitability. Motivation for the present work along with research gaps are explored from the literature. To conclude the chapter, problem statement is formed for the present thesis. In the end, the dialectal speech corpora used in the current research work is discussed in detail.
- **Chapter 3: Dialect Identification using Speech based Features:** This chapter presents details of the existing and proposed features from the excitation source, vocal tract system, and prosodic aspects of speech signal for dialect recognition. Also, the spectro-temporal feature extraction procedure is presented. Training of classifiers that have been used in automatic dialect recognition systems are discussed. Experimental details are provided with the tabulation of results. Analysis and a comparative discussion of results are given at the end of the chapter followed by summary and conclusions.

- **Chapter 4: Dialect Identification using Chroma-Spectral based Shape Features:** This chapter contains the explanation of the proposed chroma feature extraction and their use for dialect identification. Details of eight spectral shape based features and their significances during dialect recognition are included. Single and ensemble based classification methods are employed and the performance is compared. Comparative analysis of obtained results with state-of-the-art methods is conducted along with a discussion on results.
- **Chapter 5: Dialect Identification from Word and Sentence Level Properties:** This chapter presents the analysis on the existence of dialectal cues at shorter spoken units namely words and sentences. Spectral and prosodic extracted features that have been extracted from words and sentences are presented for dialect classification. Comparative analysis of dialectal cues across different dialects from these spoken units is done.
- **Chapter 6: Characterization and Identification of Kannada Dialects:** This chapter elaborates on the three different dialect recognition systems designed for Kannada language. Language dependent dialect-specific cues are extracted from vowels, consonants and *case* utterances. Details of dynamic and static properties of vowels modeled through spectral and prosodic features are discussed. ANOVA based statistical analysis details are elaborated along with dialectal studies conducted using Kannada consonants and *cases.*
- **Chapter 7: Summary and Conclusions:** Significant contributions made to the development of dialect recognition systems are briefly summarized in this chapter along with important conclusions. Immediate future research directions are provided at the end of the chapter.

At the end of the thesis, relevant references and publications out of the thesis are listed.

CHAPTER 2

Literature Review

In Chapter 1, general background on automatic dialect processing from speech perspective is briefed. Basic stages in the development of dialect recognition systems are also discussed. This chapter covers compendious reviews about the various datasets, speech features and classification techniques used for dialect processing. Research gaps derived from the available research literature and problem statement is formulated for the current thesis work are provided at the end of the chapter.

2.1 Introduction

Dialects of any language are because of the phonological, lexical and grammatical variations exhibited in the usage of language with very minor and subtle differences. In the recent past, dialect identification from speech is emerging as one of the prominent areas in speech research. This is mainly due to the extensive increase in the use of interactive voice-based systems. In this scenario, it is important to address speech variabilities caused due to dialectal differences in order to achieve effective, realistic man-machine interaction. Logically these ADI systems are essential in the development of speech based systems for regional, and resource-constrained languages. In the last few decades, considerable work is being carried out in the area of dialect processing. Majority of the work reported in the literature on dialect processing have concentrated mainly on the use of acoustic, phonetic and phonotactic approaches. This chapter provides a critical review of the existing research used in dialect processing, in the context of international and Indian scenarios. Review on important characteristics of dialect datasets and different speech corpora used for the dialect recognition purpose is given in Section 2.2. Details of distinguishable features that have been extracted from different aspects of speech for dialect processing are discussed in Section 2.3. A brief review of various machine learning algorithms employed for dialect processing is given in Section 2.4. Research gaps which have been identified from critical review of the existing work and the motivation for the present work are discussed in Section 2.5. Problem has been formulated for the current thesis work and is presented in Section 2.6. Further, this chapter provides the details of dialect datasets that are collected and used in the current research work in Section 2.7. This chapter ends with conclusions drawn out of the review work along with a brief summary in Section 2.8.

2.2 Dialect Speech Corpora: A Review

Dialect processing is the complex aspect of speech processing, since it also involves language and speaker characteristics of a specific geographical region. In the literature, limited research activities are reported in the field of dialect processing considering many languages being spoken across the world. This is mainly because of unavailability of standard datasets and lack of knowledge of clear boundaries between different dialects within a language (Curzan, 2013). Dialects vary with the variations in age, place and geographical boundaries. Dialects also evolve due to adaptations in usage patterns of the languages over long period of time (Chambers and Trudgill, 1998). Dialect related tasks, such as characterization, identification or synthesis of dialects demand suitable dialect specific datasets. The design, type, and collection of dialect datasets are mainly dependent on the dialect related task to be addressed. In the literature, two types of speech datasets are commonly found, one is generally known as read speech either recorded in the studio or in a clean, controlled atmosphere, the other one is spontaneous speech recorded in normal surroundings. The speech collected may be text dependent or read from

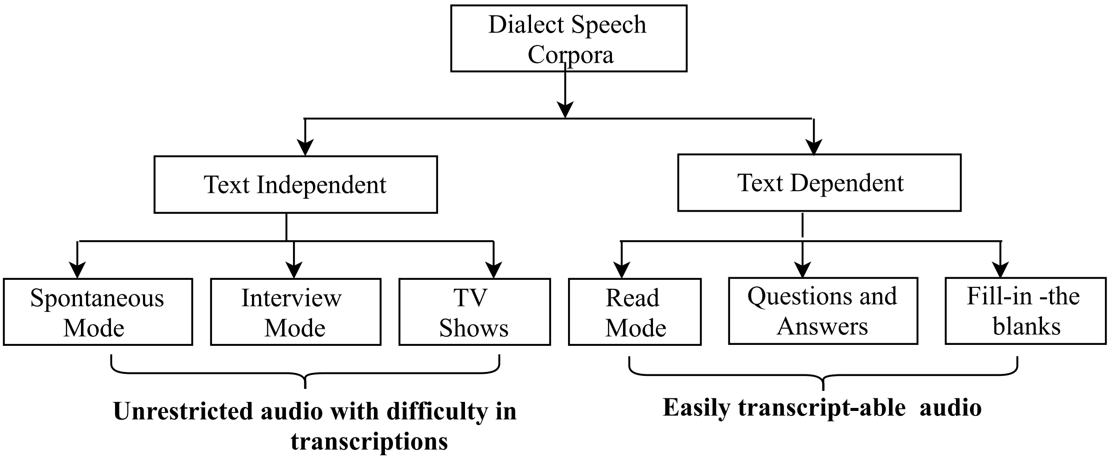


Figure 2.1: Taxonomy of datasets used for dialect processing

the pre-typed text. In the other case recording is done on-site which is realistic; available as a natural conversation and a spontaneous one. This is mostly used to be text independent. This type of recording through natural conversational and interviews generally captures background noise as well. This is one of the more comfortable and reliable methods of speech data collection.

It is obvious and has also been reported in the literature that, dialectal cues are naturally present to a great extent in spontaneous speech rather than in text based read speech (Rouas, 2007). Spontaneous speech has obvious prosodic cues such as different speaking rates, pauses, intensity variations, intonation patterns, hesitations, repetitions, and partially spoken words (Liu et al., 2010). Broad categorization of type of standard datasets is presented in Figure 2.1. Majority of the existing datasets are collected through the read mode or fill-in-the questionnaire, possibly with prior training, which may not be significantly useful in processing the dialects (D'Arcy et al., 2004). Very few text independent and spontaneous speech datasets are recorded and made publicly available. Generally, these are collected through telephonic speech, TV/Radio reality talk shows, natural conversations, or interviews (Huang et al., 2007). Table 2.1 presents the significant merits and demerits associated with these speech datasets.

Table 2.1: Types of speech corpora for dialect processing

Type of a Dataset, Examples Mode	Merits	Demerits
Text Dependent. Collection Mode: Reading, Questions and answers, Fill-in-the blanks, (Clopper and Pisoni, 2006), (Garofolo et al., 1993)	• Easily Transcriptable • Prior preparation can be done before recording to record only the intended information • Recording is effective to the purpose • Based on the purpose of the task, questionnaire may be prepared • Studio recording is possible with controlled environment so that noise free speech data is available	• Lack of presence of natural dialectal cues such as speaking rate, pauses, corrections, hesitations, repetitions, partial words, which play a role in dialect identification • Time consuming and tedious recording especially when number of speakers is large. • Many times, effective size of the data is limited • This speech data many a times may not be suitable for real time system development.
Text Independent (Unrestricted). Collection Mode: Spontaneous recording, TV/Radio reality shows, Telephonic Conversations, Interview, (Canavan and Zipperlen, 1996), (Khurana et al., 2017)	• Unrestricted audio is most suitable for dialect processing. The collection does not need much preparation in advance. Large sized dataset can easily be recorded. • Speech data is more natural and rich with prosodic information such as speaking rate, intonation, stressed patterns, pauses, partial words, disfluencies etc., • Telephonic conversations and TV/Radio programs may be used as data. • This speech data is suitable for speech system development that is more realistic.	• Difficultly in transcribing the spontaneous speech as no automatic transcribers are available • Due to the presence of overlapping utterances, it is difficult to segment the speech into smaller units • Presence of background noise

Many research papers on datasets have also discussed the general aspects to be considered while preparing, recording and using them. Important once among them are listed below.

1. During the design and collection of new dialect speech corpus, potential influences of several factors such as social differences between speakers, demographics of the speaker, etc. are needed to be carefully addressed (Clopper and Pisoni, 2006).
2. Selection of size (in duration), quality or type (read or spontaneous) of dataset generally depends on applications (Huang et al., 2007; Huang and Hansen, 2007). Contextual information is essential and to be recorded as semantic and linguistic context; both play a prominent role in many applications.
3. Number, gender and age of speakers, contributing to the recording, have to be taken care of, to avoid the development of biased systems.
4. In the era of deep learning, size of the dataset plays a vital role in majority of the machine learning applications. Generally deep learning architecture does not work well with small sized datasets.
5. Dialect information is paralinguistic in nature and it is observed that such information lies in higher frequency ranges; poor quality recording devices may hinder the performances of the systems.
6. Sufficient support of the linguists is to be taken to identify different dialects and the boundaries between them.

Brief details of the standard datasets widely used for discrimination and identification of dialects can be found in Table 2.2. From this Table, it is observed that many of the datasets are collected in English and Arabic languages. Similarly, few datasets are collected in Chinese, Spanish and Japanese languages. The number of datasets reported in Indian languages is comparatively less.

Table 2.2: List of widely used speech corpora for dialect discrimination

SL. No.	Name of Database	No. of partici-pants	Language	No. of Dialects	Recording Environment and mode	References
1.	Texas Instruments and Massachusetts Institute of Technology (TIMIT)	630	American English	8	Broadband (Clean room), Read	(Garofolo et al., 1993)
1.	CALLFRIEND	60	American English	2	Telephone, Spontaneous	(Canavan and Zipperlen, 1996)
3.	Network Texas Instruments and Massachusetts Institute of Technology (NTIMIT)	630	American English	8	Telephone, Spontaneous	(Jankowski et al., 1990)
2.	Miami	219	Spanish	3	Quiet Room, Read	(Zissman et al., 1996)
5.	CALLHOME	-	Arabic	-	Telephone, Spontaneous	
3.	Resource Managament (RM1 2.0)	160	English	2	Quiet Room, Read	(P. Price et al., 1993)

Table 2.2: Details of commonly used speech corpora for dialect discrimination

4.	Accents of the British Isles corpus (ABI)	300	British English	15	Quiet Room, Read	(D'Arcy et al., 2004)
5.	Intonational Variations in English (IViE)	84	British English	9	Quiet Room, Read	(Grabe and Post, 2002)
6.	Nationwide Speech Project (NPS)	60	American English	6	Read, Interview	(Clopper and Pisoni, 2006)
7.	Santa Barbara (Part 1-4)	100	American English	-	Spontaneous, Class Room	(Du Bois et al., 2005)
11.	Switchboard corpus	543	American English	5	Spontaneous, Telephone	John Godfrey and Edward Holliman (1993)
8.	QCRI Corpus		American English	5	Spontaneous	(Khurana et al., 2017)
9.	Spoken Arabic Regional Archive (SARA) Corpus		Arabic	3	Spontaneous	(Zaidan and Callison-Burch, 2014)
13.	Annotated Al Jazeera Dialectal Speech Corpus		Arabic	4	Spontaneous, Class Room	(Wray and Ali, 2015)
14.	UT-Podcast		English	3	Spontaneous	(Hansen and Liu, 2016)
10.	Chinese dialect dataset	–	Chinese	4	Spontaneous	Lei and Hansen (2011)
11.	Pan Arabic dialect dataset	100	Arabic	5	Spontaneous	(Lei and Hansen, 2011)

Table 2.2: Details of commonly used speech corpora for dialect discrimination

12.	Multi-dialect multi-genre evaluation corpus (MGB-3)	–	Arabic	5	Spontaneous	(Bahari et al., 2014)
13.	Arabic CTS corpora	–	Arabic	4	Spontaneous	(Bořil et al., 2012)
14.	Hindi Dialect dataset	30	Hindi	4	Read, Spontaneous	(Sinha et al., 2015b)
15.	Telugu Dialect Dataset	–	Telugu	3	Read	(Mannepalli et al., 2016)
16.	Kannada Dialect Dataset	115	Kannada	5	Spontaneous	(Chittaragi et al., 2019)
17.	Assamese Dialect Dataset	12	Assamese	4	Read	(Sarma and Sarma, 2016)
23.	Kannada Dialect Dataset	115	Kannada	5	Spontaneous	Chittaragi et al. (2019)

In literature, apart from these listed datasets various other dialect datasets of several regional languages can be seen. Majority of them are collected, purely with the intention of dialect processing and dialect-related studies of respective languages. They are mostly small in size with limited number of speakers. It is also true that many of the datasets listed above are not available publicly and are small.

From the context of Indian languages, several studies have been reported for dialect identification from different dialects of Hindi language (Rao and Koolagudi, 2011; Agrawal et al., 2016). At IIIT Hyderabad, few research groups are working on the development of a standard dataset for Indian languages including Kannada for speech recognition purposes. A team from IIT Guwahati is working on the dialects spoken in North-East Indian states and characterization of tonal languages (Konnerth et al., 2015). The International Committee for the Coordination and Standardization of Speech Databases and Assessment Techniques called as COCOSDA has been established, to promote and encourage international collaboration and information exchange in the areas of spoken language and dialect processing (Sinha et al., 2015a). From the literature review, it is observed that no significant works are reported on dialect processing of many Indian languages, due to unavailability of standard datasets. There is a need for the dialect specific speech datasets for many Indian regional languages to conduct full-fledged research activities. Also, these datasets must be sufficiently large with more number of speakers and with proper gender ratio.

2.3 Features: A review

Selection of suitable features for any speech task is an important aspect in developing an efficient speech system. Specifically, features are nothing but the low dimensional representation of the entire input speech signal chosen for a task. It is necessary to understand the various aspects of speech signal available to human perception. Features identified have to be suitable to capture the intended information for the target task. From the existing literature, it is found that acoustic-phonetic and phonotactic are the two commonly used feature types for dialect recognition [1] (Biadsy, 2011). Acoustic-phonetic models include the extraction of excitation source, spectral, and prosodic features using speech signal processing techniques. Phonotactic models involve the use of PRLM including phone recognition followed by language modeling for dialect identification. However, these are supposed to be language dependent models with the heavy constraint of transcribed speech (Zissman et al., 1996; Barkat et al., 1999; Biadsy and Hirschberg, 2009). Present section focuses on providing a brief review of the different types of features proposed in the literature for dialect recognition.

2.3.1 Acoustic-Phonetic Features

In general, the speech signal is processed at three different levels namely, sub-segmental, segmental and supra-segmental. The sub-segmental level approach extracts feature by processing a signal within a glottal cycle in terms of the shape of the glottal pulse, durations of open and close phases of vocal folds (Kodukula, 2009) etc.. Segmental level processing involves the observation of the unique sequence of the vocal tract shape and size. Each of these sound units, produced may be characterized with the spectral envelope. Finally, the supra-segmental level analysis of signal includes the processing of longer duration signal in oder to capture the prosodic information such as, duration and energy patterns, the intonation, pitch features from syllable and/or word sequences.

A Excitation Source Features

Every spoken unit is produced as a result of the unique articulatory configuration of the excitation source and the vocal tract system (Kodukula, 2009). The quasi-periodicity of air pulses generated by the vibration of the vocal fold is the primary source of excitation during the production of voiced speech. Features derived from these excitations are named as excitation source features. These features are extracted from speech through inverse filtering of Linear Prediction Coefficients

[1] In this thesis, the classification, identification, and recognition words are used interchangeably conveying similar meaning with a standard machine learning goal.

(LPCs), where, inverse filtering separates both excitation source and vocal tract information (Makhoul, 1975). The obtained signal is called an LP residual. Subsegmental analysis of LP residual signal includes extraction of epoch locations, the slope of the epoch, strength of epoch, the energy of excitation, instantaneous frequency features and so on. Characteristics of glottal pulse activity, open and closed phases of the glottis, and Glottal Volume Velocity (GVV) properties have also been computed in literature as the correlates of excitation source information (Kodukula, 2009). Accurate estimation of epoch locations can be obtained from Zero Frequency Filtered Signal (ZFFS) and LP residual signal (Nandi et al., 2017). Epoch location extraction from ZFFS is preferred over that from the LP residual signal. Since, LP residual-based method generally computes the peaks of random polarity around the epochs locations resulting with ambiguous values (Kodukula, 2009).

The attributes extracted from and around identified epoch locations are said to contain substantial language-specific information (Rao and Nandi, 2015). Therefore, accurate estimation of the epoch locations from the signal, plays a prominent role in the analysis of excitation source (Nandi et al., 2017). Apart from these, language recognition studies proposed (Nandi et al., 2015; Murty, K Sri Rama and Yegnanarayana, B, 2008) have suggested the existence of language-specific cues, at regions immediately following the epoch locations. They have also demonstrated that these regions are relatively more robust to the external degradations than the other regions. Experiments have been conducted between a pair of speakers, to identify the speaker and dialect-related cues through estimation of glottal flow derivatives on English and Spanish dialects. This study has reported the existence of dialectal differences in these languages (Yanguas and Quatieri, 1999).

From the existing literature, it may be noticed that the excitation source level information can also be used as supplementary to develop dialect recognition systems along with vocal tract and prosodic features. However, studies reported on dialect identification using source information are substantially less exploratory when compared to other features. Excitation source information has to be systematically studied to extract dialect specific cues present in the form of higher order relations from LP residual signal. Identification of instants of excitation, glottal closure strength and slope of excitation, and parameters of glottal pulse and so on are helpful to capture source information.

B Spectral Features

Acoustic modeling mainly uses the features directly extracted from the speech signal. Majority of the existing dialect processing studies reported in literature have considered acoustic features. These include; analysis of waveform for temporal variations, frequency domain analysis, cepstral analysis, LP analysis, processing of Voice Onset Time (VOT), formant frequency measurements, jitter, shimmer and so on. Frequency domain analysis gives considerably better vocal tract features compared to their time domain counterparts. The Fast Fourier Transform (FFT) of a speech frame of size 20-30 ms well establishes a short time spectrum for vocal tract characterization (Rabiner and Juang, 1993).

Generally, spectral features extracted from a speech segment of size 20-30 ms have been treated as the strong correlates of varying shapes of the vocal tract (Benesty et al., 2007). Spectral acoustic differences existing among dialects have been studied extensively from the cepstral domain through MFCCs, SDCs, and LPCC features. MFCC features are used widely in literature since they resemble and imitate human auditory system and follow a nonlinear model to process speech signal similar to the way humans process. A nonlinear Mel scale filter is used to measure lower frequency components as they carry most of the phonetic information suitable for dialect processing (Tsai, Wuei He and Chang, Wen Whei, 2002). Several studies are available in literature where, MFCC features are used extensively for implementation of dialect recognition systems (Torres-Carrasquillo et al., 2004; Campbell et al., 2006; Mehrabani and Hansen, 2015). MFCC feature vector alone describes the power spectral envelope in a frame. Besides, most of the time instead of only 13 MFCC features, delta (differential), and delta-delta (acceleration) features, representing SDC features are also extracted to capture trajectory information over time (Liu and Hansen, 2011).

Table 2.3: Important research outcomes on the use of spectral features for dialect recognition

Sl.No.	Features	Dataset, Purpose and Approach	References
1	PMVDR-SDC	Three Latin American Spanish dialects Cuba, Peru and Puerto-Rica are used for development of dialect recognition using tUBM (traditional UBM) and hUBM (hierarchical UBM). Achieved performance is comparatively better over GMM model.	Liu and Hansen (2011)
2	MFCC-SDC	Call Friend corpus with two dialects of Spanish, English and Mandarin, Miami corpus with three dialects-Cuba, Peru and Puerto-Rica are used. Dialect recognition system is proposed by using GMM-SDC and UBM method. Better performances was found with Miami over Call Friend corpus	Torres-Carrasquillo et al. (2004)
3	MFCC-SDC	Kannada dataset with five and IViE British English dataset with nine dialects are used. Dialect recognition systems are proposed by using SVM and neural network models with hyper parameter tuning method using grid search algorithm. Cepstral features have shown better performance with both SVM and ANN approaches.	Chittaragi et al. (2019)
4	MFCC-SDC	TIMIT dataset with eight American English dialects is used. Dialect identification using SVM and extreme learning algorithm is proposed and extreme learning algorithm has shown better performance	Rizwan and Anderson (2018)
5	MFCC-SDC, PLP- Perceptual linear prediction coefficients, MF-PLP	Hindi dataset with five dialects is used for implementation of dialect identification system. 5 layered auto associative neural networks (AANN) with 3 hidden layers	Sinha et al. (2015b)
6	13 MFCC-SDC	Three dialects of Telugu are used for development of dialect identification system by using GMM and layered auto associative neural networks (AANN) with 3 hidden layers	Mannepalli et al. (2016)
7	F0+ Slope of F0+ 39 MFCC features in the tones	Tibeto-Burman Ao language with two dialects Changki and Mongsen is used for development of dialect recognition system with both tonal, spectral features and combination. Semi-supervised GMM classification model is used for dialect recognition system.	Tzudir et al. (2018)
8	Formants, LSP (Line Spectral Pairs), and MEPZ (MFCCs+ Energy+Pitch)	Three dialects of Spanish language are used. Dialect identification is done using individual and combination of features explored using SVM-GMM hybrid algorithm.	Chitturi and Hansen (2007)

Existing literature has shown that majority of the dialect processing systems have explored the cepstral coefficients mostly through RASTA MFCCs, and sometimes with RASTA PLP features. Majority of the authors have used short time processing of the whole audio signal while parameterizing various spectral features. Whereas, few have considered the vowel, consonants and syllable level utterances for processing. Dialect recognition systems are also modeled by capturing the vocal tract shape related spectral characteristics, and they have significantly performed well. Features like spectral flux, slope, entropy, centroid, resonance frequencies (formants) have also been extracted from the spectrum (Fourier transform of a speech frame) (Huang et al., 2007; Chittaragi and Koolagudi, 2017). In this work, the systems developed using MFCCs are considered as the baseline models (Etman and Beex, 2015).

Identification of dialects from noisy speech have been addressed through the measure of perceptual minimum variance distortion-less response (PMVDR), using SDC features. This combination has shown robustness over MFCC features, along with an improvement in accuracy (Liu and Hansen, 2011). A study has been reported to measure the spectral differences from volume space analysis in a 3D model with the use of log-likelihood score distributions, derived from MFCCs and is modeled by using GMM approach (Mehrabani and Hansen, 2015). However, addition of text-independent prosody features, with MFCCs, have performed well in the case of Arabic dialects and South Indian languages (Mehrabani and Hansen, 2015). In the literature, a few systems have been reported to use spectral features such as formants frequencies, LSP (Line Spectral Pairs), and MEPZ (MFCCs + energy + pitch) features for classification of Spanish dialects. In addition of these, new features have shown the dialect recognition improvement using GMM-SVM hybrid models (Chitturi and Hansen, 2007). A few of the dialect identification systems developed by analyzing of spectral artifacts on different datasets are listed in Table 2.3.

C Spectro-Temporal Features

Existing systems have extensively evaluated spectral aspects of speech for dialect processing. However, there are few dialectal cues that vary in both spectral and temporal domains known as spectro-temporal features. In the literature, temporal aspects of speech are mainly captured using 2D Gabor features. Gabor features are said to have biological-inspiration derived from 2D Gabor filter bank. These features are found to decompose the spectro-temporal power density. Indeed, these features are popularly used in image processing applications (Schädler et al., 2012). Few attempts are also found in the literature on using Gabor features for speech and speaker recognition applications (Lei et al., 2012). Gabor features are found to be successful, as these features try to model specific stimuli to which the neurons of the mammalian auditory cortex are sensitive (Meyer, Bernd T and Kollmeier, Birger, 2011; Lei et al., 2012). Both spectral and temporal modulation frequencies do exist in these stimuli. Gabor filters representing spectro-temporal modulations are believed to emulate the human auditory system, using signal processing strategies, leading to the recognition of language and dialectal aspects from speech (Meyer et al., 2011).

D Prosodic Features

Human beings embed certain cues of naturalness, such as intonation, duration, and energy patterns on the sequence of sound units during speech production, known as prosodic features (Ramus and Mehler, 1999). Prosody captured through intonation, intensity variations, stress patterns, rhythmic style loudness, melody and etc. are considered as para-linguistic features of speech, contributing to the aspects like emotions, accents, dialects, gender etc. (Mary and Yegnanarayana, 2008). In general, the prosody is observed over a prolonged duration of spoken units (supra-segmental) namely; phonemes, pseudo syllables, syllables, words, phrases, sentences or discourse (Biadsy and Hirschberg, 2009). Pseudo syllables are shorter spoken units of the form CV, that includes a cluster of optional consonants, followed by a single vowel segment. It is reported that dialect specific cues are observed in spontaneous speech in the form of speaking rate, filled pauses, phonetic repetitions, intonation, rhythm, melody and stress characteristics during pronunciation (Liu et al., 2010). Contrastingly, read speech is said to be rich in spectral information instead of prosody (Nakamura et al., 2008). It is popularly known and understood that prosody makes speech more realistic and intelligible along with carrying an intended message. Intonation shows the variations (rise and fall) in temporal pitch dynamics. Due to differences in speaking styles, the length of a spoken unit varies with unique rhythmic pattern and is observed with a unique stress imposed. Stress is the emphasis given to certain syllables or words in a sentence. Profile of a fundamental frequency of a speaker, phoneme duration, and energy features are acoustic correlates corresponding to intonation, rhythm and stress feature respectively (Rouas, 2007; Bougrine et al., 2017). In the literature, there are several references on the use of prosodic features for dialect recognition. Table 2.4 presents some of such important works.

It is observed from the literature that most of the Arabic dialects have demonstrated a significant prosodic difference among them. Indeed, a study carried out by (Barkat et al., 1999) has shown that prosodic features alone are sufficient to classify majority of the Arabic dialects. Few studies have reported that, in the cases of Indian languages, energy, duration, pitch, and their derivatives are considered as the quality attributes for classification of dialects (Sinha et al., 2015a; Biadsy et al., 2011).

Majority of the studies conducted on dialect identification have concentrated on extracting prosodic features from syllable and pseudo-syllabic structures (Rouas, 2007; Rao and Koolagudi, 2011; Etman and Louis, 2015). A study has been reported on the use of global and local prosodic features extracted from Z-normalized pitch and intensity contours along with the duration values of each pseudo-syllables. Further, intonation and rhythmic features have been extracted to classify four Arabic dialects using, HMM classifiers. However, overall dialect recognition accuracy has improved with the addition of Phonotactic PRLM models (Biadsy and Hirschberg, 2009). A study has proposed to use, four pitch values ($f0_{min}$, $f0_{max}$,

Table 2.4: Important literature on the use of prosodic feature for dialect recognition

Sl.No.	Features	Dataset, Purpose, and Approach	References
1	Intonation and rhythm, Global prosodic features	Four dialects of Arabic language are used. Dialect recognition using global prosodic features from pseudo syllables and modeled through HMM based GMM method	Biadsy and Hirschberg (2009)
2	Rhythm and intonation in terms of pitch, tonal information, and pitch trajectory features	six dialects of Algerian Arabic language are used. Dialect recognition using SVM using syllable information. A study with same features using Hierarchical classification approach for spoken Arabic Algerian Dialect Identification (HADID, DNN)	Bougrine et al. (2017, 2018)
3	Pitch statistics, energy statistics and duration models	ARABER dialect corpus with three dialectal regions Dialect recognition from pseudo-syllable units using GMM method	Rouas (2007)
4	Pitch and Energy statistical features	Five dialects of Kannada language. Dialect recognition from sentences level utterances using SVM and XGB algorithms.	Chittaragi and Koolagudi (2018)
5	Tone related parameters are extracted from pitch flux	Chinese dataset with three dialects is used. Dialect classification using GMM method	Ma et al. (2006)
6	Formant frequencies, pitch, pitch slope, intensity, duration	Hindi dataset with four dialects is used. Dialect identification using SVM and GMM methods from linguistic and paralinguistic analysis of vowel sounds	Sinha et al. (2017)
7	Pitch statistics, rhythmic features, pitch slope, pitch peak alignment, RMS intensity, and duration	TIMIT dataset with eight American dialects is used. Dialect classification between pairs of dialects is using SVM	Etman and Louis (2015)

$f0_{mean}$, $Df0(f0_{max} - f0_{min})$), frame energy and duration features for classification of four dialects of Hindi language using AANN. Score level fusion of prosodic features along with 12 MFCCs and 24 Delta features, have shown better accuracy (Sinha et al., 2015a). Very few works are reported in the literature on processing speech for longer duration, such as, words or sentences for dialect processing. Fundamental prosodic features, like energy, pitch and duration are extracted from sentences and are seen to perform better than frame level spectral features, during Kannada dialect recognition (Chittaragi et al., 2019; Purnell et al., 1999; Huang et al., 2007). There are also references in the literature on extracting dynamic and static behavior of the pitch and intensity values for classification of dialects. These have also demonstrated significant contributions in the classification of dialects (Lim et al., 2005; Biadsy and Hirschberg, 2009).

E The i-vector Features

The i-vector features are a state-of-the-art discriminative latent features. These are primarily proposed for speaker and language identification tasks (Dehak et al., 2011a,b). A text-independent approach of deriving i-vectors is basically initiated through Joint Factor Analysis (JFA). JFA represents the speaker and channel variabilities in two different subspaces. However, the i-vector approach defines only one space to include both variabilities (with speaker and channel normalization). This space is called as *total variability* space and is represented by the matrix called *'total variability matrix'* (T). This is the reason, i-vector features have performed well in the case of speaker identification and verification systems (Dehak et al., 2011a; Sadjadi et al., 2013). These features are also considered as the alternative feature set over the conventional spectral and prosodic features. As the performance of the i-vectors is better for language recognition tasks, several attempts have been reported in literature on the application of i-vectors, for dialect identification (Dehak et al., 2011b). GMM-UBM, Joint factor analysis, and total variability space-based approaches are important among them and they are proposed for dialect and foreign accent recognition tasks (Dehak et al., 2011b; Hansen and Liu, 2016; Behravan et al., 2015). The i-vectors are represented by a low ranked matrix T that captures relevant variabilities concerning total variability matrix. Further, a slight increase in performance of dialect processing system is reported with i-vectors compared to conventional features even with the reduction in dimensions (Zaidan and Callison-Burch, 2014).

2.3.2 Phonotactic Features

Phonotactics is a branch of phonology that deals with the study of permissible usage and combination of phonemes in a given language. Phonotactics also sets rules for syllable structures, consonants, and vowel sequences. Generally, phonotactic constraints are highly language dependent. A language-specific phonotactic modeling technique primarily includes a phone recognizer followed by language modeling. Basically, phone recognizer tokenizes the speech into phonemes along with phonetic transcripts. Phonetic transcriptions describe the word with the sequence of known phonemes (Zissman et al., 1996; Chen et al., 2010). Intuitively, it is quite reasonable to use phonotactic features for dialect recognition as it is basically a sub-task of language identification (LID) (Zissman et al., 1996). Hence, we observe many references in literature on the application of LID techniques for developing ADI systems. Familiar LID models such as Phone Recognition followed by Language Modeling (PRLM), parallel PRLM (PPRLM), and Parallel Phone Recognition (PPR) models are widely used in dialect identification as well. However, these are highly language dependent models with the basic constraint of transcribed speech (Zissman et al., 1996; Barkat et al., 1999; Biadsy and Hirschberg, 2009). It is also challenging to apply language based phonotactic approaches to dialect recognition if there is no transcribed speech available. Moreover, producing either the phonetic transcriptions or the orthographic transcriptions for each training utterance is an expensive task (Zissman and Berkling, 2001). It is also a time-consuming process that takes a lot of time and expertise of highly skilled linguist, fluent in the language of interest. It is also true that dialect recognition task has to face many ambiguities since there are few linguistic rules to demarcate the dialectal boundaries. Normally, dialectal boundaries are highly overlapped and confused since dialects are being derived from the same language by sharing common phoneme set and grammatical rules. Apart from these, phonotactic models demand a transcribed speech for processing.

2.3.3 Combination of Features

Recent research trends in the area of speech processing are to use a combination of various features extracted from several aspects of speech such as production, perception, linguistics etc. Vocal tract system, prosodic features, and excitation source features discussed earlier in this chapter are known to constitute the complementary, mostly mutually exclusive information (Etman and Beex, 2015). Hence, we generally observe that in the literature, many studies reported on dialect recognition using various combinations of available features. In addition, some works on dialect identification have reported the combination of both acoustic and phonotactic models representing speech and language models. These combined feature models have shown better performance when compared to the systems that is developed using individual features (Biadsy et al., 2009; Mehrabani and Hansen, 2015; Chittaragi et al., 2018b; Hansen and Liu, 2016).

2.4 Machine Learning Algorithms: A Review

Several machine learning algorithms have been adopted for developing automatic speech based systems such as speech recognition, speaker recognition, language & dialect recognition, emotion classification, speaker verification and so on. A wide variety of machine learning algorithms is available now a days to solve classification and predictions problems due to advancement in hardware and software technologies (Pedregosa et al., 2011; Giannakopoulos and Pikrakis, 2014). However, there is no systematic and standard approach to choose an appropriate one. Majority of the times specific algorithms are chosen either based on heuristics (existing references for similar problems) or in random. Sometimes a particular classifier is selected among available options by performing alternative experimentation or running pre-test experiments. Selection of a suitable classification approach, while developing efficient systems is considered as the challenging problem (Etman and Beex, 2015).

In general, there are three broad categories of classification models namely: supervised, semi-supervised, and unsupervised. Audio/speech based systems are generally developed, using supervised learning approaches. The audio input is highly non-linear and hence most of the time dialect recognition problems are assumed as a classification problem and therefore supervised approaches are used. Classification algorithms may also be classified into either generative or discriminative models. Generative models use probability density functions, prior probabilities and use an unsupervised learning approach. A few important generative classification methods are GMM, hidden Markov models (HMM), Sigmoidal belief networks, Bayesian networks, Markov random fields and so on (Kotsiantis et al., 2007). Discriminative classifiers are known as non-probabilistic binary classification models and usually adopt supervised learning. These models directly estimate the posterior probabilities, without considering the underlying probability distributions. Few important models are Logistic regression, artificial neural networks, SVMs, k-Nearest neighbor, Conditional Random Fields and so on. However, these models attempt to learn either hard or soft boundary between the classes. Generative models exploit the data distribution of individual classes. These models can be clearly understood by taking a simple analogy. During the task of identifying language from speech, discriminative models try to determine and explore the linguistic differences among the large number of different languages, without learning any language, which is said to be much more straightforward in approach. Whereas, generative models try to learn each language, with the knowledge of probability distributions of features and determine to which language the speech belongs to (Jebara, 2012).

In recent time, there is a trend of using combination of output of multiple classifiers for final decisions. Based on the number of classifiers used one can categorize the classification into two groups namely; single classifier based algorithms and ensemble algorithms. Generally, single classifier based methods use statistical (probabilistic) or rule-based approaches. In these algorithms, classification performance relies only on a single classification model. Models such as GMM, HMM (Hidden Markov Model), LDA (Linear discriminant analysis), SVM (Support vector machines), and neural network are few examples of this kind (Rao and Koolagudi, 2011; Chittaragi et al., 2019; Biadsy et al., 2011). Whereas, ensemble methods are known as meta-algorithms since they are designed by combining the predictions of several individual classification techniques. This approach of combination is expected to reduce variance (bagging), bias (boosting), and improve predictions (stacking). Recently, ensemble methods have shown a new research direction while solving classification problems (Dietterich, 2000a). Ensemble algo-

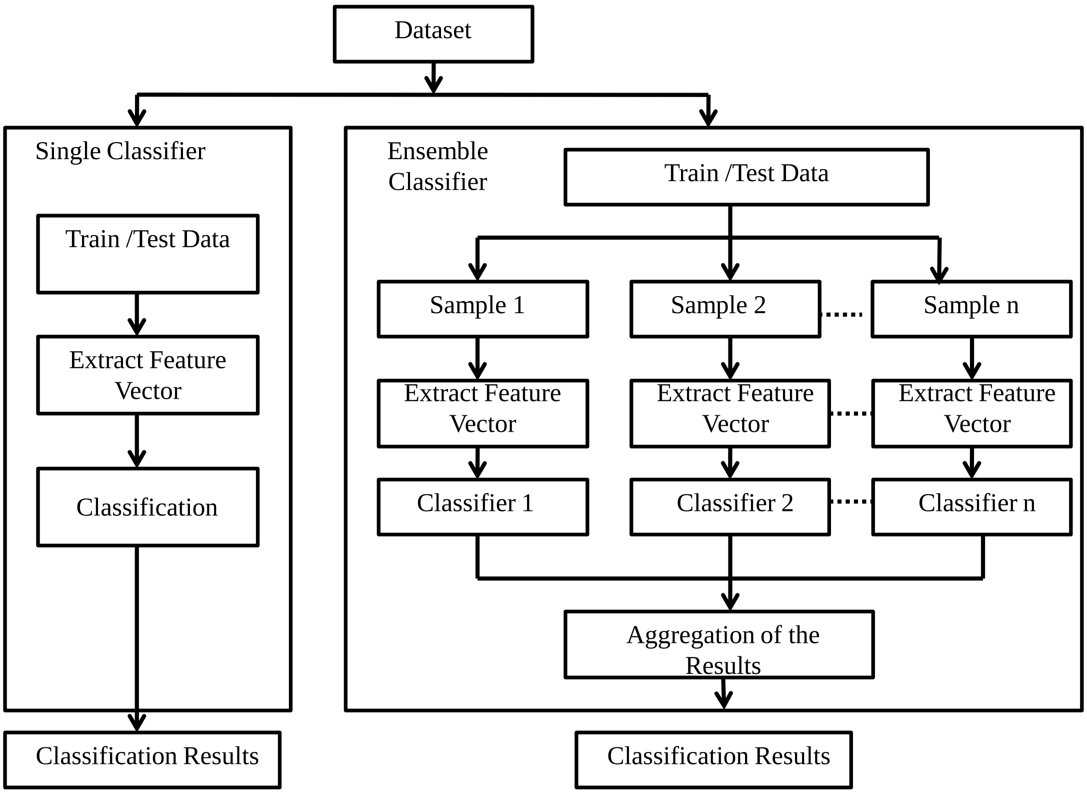


Figure 2.2: Working of single and ensemble classifiers

rithms are expected to improve the predictive performance of classification since end performance is relying on decisions made by multiple classification algorithms. Also, these ensemble algorithms work with an analogy that commonly used in human predictions such as *"the wisdom of the crowd"*, over individual prediction, democratic decisions over individual ones. Fig. 2.2 shows the schematic working block diagram of single and ensemble classification methods. It is also suggested in the literature that, in majority of the situations, ensemble algorithms perform better over the single classifiers (Lessmann et al., 2015).

Various types of classification methods are used for development of ADI systems. Few popular ones among them are, GMM, HMM, SVM, ANN, DNN, and ensemble algorithms. The list of different classification algorithms used for ADI development is given in Table 2.5 along with the other useful information.

2.4.1 Single Classifier based Algorithms

Out of several single classifier based algorithms, GMM based dialect recognition systems are widely used and reported in literature. GMMs efficiently model any set of uncorrelated normally distributed data (Mehrabani and Hansen, 2015; Biadsy, 2011; Rouas, 2007; Chen et al., 2001). Majority of the baseline systems are implemented using GMMs (Huang et al., 2007; Soorajkumar et al., 2017). Later, several studies have been proposed, where, GMMs are combined with the universal background model (UBM) resulting in the improvement of performance (Liu and Hansen, 2011; Hansen and Liu, 2016). HMMs and GMMs have been employed as classifiers to classify four Arabic dialects, using local and global prosodic features.

SVMs are found to be very powerful prediction and classification models, designed for handling high dimensional input spaces. SVM method demonstrates generalization performance across several applications of speech. SVMs try to capture the discriminating parameters across the feature vectors for identification of dialects (Utami, Iut Tri et al., 2014; Chittaragi et al., 2019; Pedersen and Diederich, 2007; Biadsy et al., 2011). SVM method has been used for classification of Hindi dialects using spectral MFCC features combined with prosodic features. SVM hyperplanes are employed for classification of dialects of American English from traditional MFCC features (Pedersen and Diederich, 2007). GMM-SVM hybrid classifiers are have been proposed for classifying three dialects of Spanish. Experiments have been conducted on individual and combinations of few features such as line spectral pairs (LSP), MFCC, Energy, Pitch, and Zero-crossing rate (MEPZ) attributes along with formants frequencies features (Chitturi and Hansen, 2007). SVMs sometimes end up with an increased computational cost during training if the dataset is too large. This problem is being addressed by using minimal enclosing ball (MEB) technique (Lachachi and Adla, 2016).

2.4.2 Ensemble Classifier based Algorithms

Recently, ensemble of multiple classifiers is gaining attention of the researchers due to its improved performance in different speech tasks. These are proven to be better classification approaches since they are designed by combining the prediction outcomes of several individual classifiers (Dietterich, 2000a; Utami, Iut Tri et al., 2014). Bagging and boosting are the two most popular techniques used to generate multiple ensembles of classifiers through manipulation of training data (Dietterich, 2000b). In bagging (bootstrap aggregation), the training set is created using random sub-sampling with replacement. Every sub-sample is solved independently using base learners, and majority voting is followed for classification problems and weighted averaging is followed for prediction problems. However, the boosting technique follows a slightly different approach. The training set is created through random sub-sampling with replacement approach. Boosting maintains a set of weights over the original training set and adjusts these weights after the base learning algorithm learns each classifier. The adjustment mechanism increases the weights of examples that are misclassified by the base learning algorithm and decreases the weights of examples that are correctly classified (Kim et al., 2002; Friedman, 2001; Friedman et al., 2001). In the literature, very few references are found using ensemble techniques for developing ADI systems (Huang et al., 2007; Darwish et al., 2014). Rotation forest, AdaBoost, random forest, gradient boosting methods are the few among widely used ensemble models (Liu and Hansen, 2011; Huang et al., 2007). Majority of the reported systems have used decision trees and SVMs as the base classifiers.

2.4.3 Artificial Neural Network based Algorithms

Artificial neural networks (ANN) are widely used, in the case of highly uncorrelated non linear data, for classification. ANNs are said to be effective in capturing the complex non-linear relations present among data. An auto-associative neural network (AANN) is a feed-forward neural network that captures the distribution of the input and is used for dimensionality reduction of the input. A five-layered AANN model is proposed using MFCCs, duration, pitch and energy features for identification of four dialects of Hindi (Rao and Koolagudi, 2011). A simple 2-layered FFNN is used to classify four dialects of Hindi using spectral and prosodic features (Sinha et al., 2014). Further, this work is extended to improve performance by using AANN using PLP and prosodic features (Sinha et al., 2015b). Recently, a study is proposed for identification of five dialects of Kannada language using MFCC and SDC features on ANN model with multilayer perceptron (MLP). A better performance of 91.9% is achieved by choosing hyper-parameters through Grid search algorithm. Optimized network with ReLU activation function, the Adam solver optimization algorithm with four hidden layers with 200 neurons each has produced better results (Chittaragi et al., 2019).

Due to adequate technological growth, many variants of neural networks such as Deep Neural Network (DNN), Convolutional Neural Network (CNN), Recursive neural network (RNN), Recurrent neural network (RNN), Long short-term memory (LSTM), Sequence-to-sequence models, models with more significant number of hidden layers and various activation functions are currently proposed in literature for different machine learning applications. These models are usually built by considering the entire speech signal instead of feature vectors.

Recently, few systems have incorporated i-vector features with DNN and have shown comparatively better dialect recognition performance. However, DNNs are found to be well suited for large datasets with inherent complexities (Zhang and Hansen, 2018; Snyder et al., 2017). CNN based models which are familiar with image processing tasks, are also used for dialect classification (Shon et al., 2018; Jiao et al., 2016). Currently, DNNs and CNNs are the widely used classifiers, since they are found to perform well over the existing ones. The main concern of these models is that they perform poorly with smaller datasets. Table 2.5 provides the details of important classification methods employed for characterization and identification of dialects from a speech signal.

Table 2.5: Important research publications on the use of various classification algorithms for dialect recognition

Sl.No.	Classification Algorithms.	Features	**References**
1	Hidden Markov models	Phonotactic Modeling and Spectral Features (MFCCs)	(Alorifi, 2008; Tsai and Chang, 1999; Biadsy et al., 2009; Huang et al., 2007; Ranjan and Dubey, 2016)
2	Gaussian Mixture Models	MFCCs, spectral and prosodic features	(Torres-Carrasquillo et al., 2004; Mehrabani and Hansen, 2015; Soorajkumar et al., 2017; Huang et al., 2007)
3	Support Vector Machines	Spectral, Global and static prosodic features, Phonotactic modeling	(Rao and Koolagudi, 2011; Pedersen and Diederich, 2007; Biadsy et al., 2011)
4	Ensemble Algorithms	British English dialects Kannada Dialects	(Liu and Hansen, 2011; Chittaragi and Koolagudi, 2017; Chittaragi et al., 2018b; Chittaragi and Koolagudi, 2018; Chittaragi et al., 2018a)
5	Artificial Neural Networks	Spectral and Prosodic features	(Chittaragi et al., 2019; Sinha et al., 2014, 2015b; Sarma and Sarma, 2016; Hassani and Hamid, 2017; Chan et al., 1994)
6	Deep Learning Networks	MFCCs, SDCs, i-vector, features, Bottleneck features	(Jiao et al., 2016; Shon et al., 2018; Chittaragi et al., 2019; Zhang and Hansen, 2018; Siddhant et al., 2017; Yang et al., 2018; Belinkov and Glass, 2016)

2.5 Research Gaps for Future Research and Motivation for Present Work

The following research gaps are identified after going through the available literature from datasets, features, and classification approaches used for dialect processing. The work carried out in the context of Indian languages is comparatively less. Specifically, with respect to the South Indian language, Kannada. Further, Kannada has a phonological system that can be used for observing dialectal differences; such as existence of retroflex consonants, presence of long and short vowels, exhaustive use of *cases*, substitution, deletion or modification of phonemes and so on. As dialectal variations are highly language dependent, the existing approaches of dialect processing are used for other languages and may not be effective for Kannada language.

Some of the important research gaps identified from the literature are listed as follows:

1. Characterization and identification of dialects using language-specific dialectal cues.
2. Complete and full-fledged dialect dataset for Kannada and other regional languages.
3. Use of excitation source features for effective dialect classification.
4. Characterization of dialects by extracting spectro-temporal variations across dialects.
5. Use of dialect related prosodic features in terms of intonation, rhythmic and melodic patterns
6. Use of phonemes, syllables, words, and sentences as units for dialect identification
7. Analysis of spectral and prosodic information of vowels and consonants among different dialects is an unexplored issue.
8. Applying multiple classifier based ensemble algorithms for dialect classification. Comparison of performance with that of traditional single classifier based approach.

9. Dialect processing with noisy and small data corpus.

There is a need for addressing general research issues related to the identification of prominent features, suitable classification models and various techniques for identification of dialects in speaker independent, unrestricted and independent text scenarios.

2.6 Problem Definition

Based on the research gaps identified from the literature review, the problem statement for the current research work is formulated as follows.

To propose an automatic dialect identification system by using speech-based features for English language. Extending the research to Kannada language (one of the Dravidian languages of India) to process five dialects of Kannada language representing five geographically diversified regions of Karnataka.

This problem is further elaborated into the following objectives.

1. Classification of dialects based on conventional speech-based features.
2. Classification of dialects based on non-conventional (dialect-specific) features.
3. Characterization and identification of Kannada dialects using language specific features.

The above-defined research objectives may be understood with the following insights. Characterization and identification of dialects demand a complete, full-fledged dataset in Kannada language with appropriate speech recordings collected from native speakers. Hence, the foremost task is to collect the text-independent spontaneous speech from people of the five identified Kannada dialectal regions. Various speech features extracted from different aspects of speech such as excitation source, vocal tract system, and prosodic features have demonstrated their importance in dialect processing. Usually, individual features provide a more affluent base of information and robustness across dialect and also sometimes exhibit limited knowledge, whereas, a combination of features can capture the non-overlapping and complementary dialectal cues. The first objective aims at using various conventional speech features for identification of dialects. Conventional speech features such as excitation source, spectral and prosodic ones are used for characterization of dialects. Language-independent ADI systems are developed by exploiting dialectal variations irrespective of gender, text, and speaker information. Every language has its unique characteristics and has its own linguistic and phonological variations. In this regard, it is very much essential to recognize dialect specific non conventional cues from each language. Dialect-specific features, such as intonation, rhythmic patterns, intensity, and stress variations, need to be extracted from each language.

In this research work, some non-conventional (dialect-specific) features, which are not used in any of the speech tasks, are considered for dialect identification. An investigation to know the significances of chroma based audio features is carried out for dialect identification. This study is aimed at investigating musical characteristics of speaking styles (dialects), especially of rhythm and intonation patterns. Chroma features are well known in the music processing domain. As every dialect has shown variations in pitch and energy parameters, chroma features are expected to capture musical pattern in speech. The second objective of this thesis work aims at developing dialect recognition systems by using non-conventional features extracted from different spoken units. Individual and combination of features are used for improving performance with the use of single classifier based systems and ensemble based systems.

Further, in this thesis, word and sentence level utterances are considered to recognize dialects. Words and sentences are said to carry significant dialect specific distinct information. This study has proposed extraction of intonation and intensity variations from pitch and energy contours to capture dynamic and static prosodic characteristics across dialects.

A language-dependent ADI system is proposed in this study for Kannada language using basic phonetic units like, vowels and consonant. However, research activities are still in their nascent stage as far as dialectal studies are considered. Vowel-based ADI system is proposed by analyzing dialectal variations in ten monophthong vowels which are manually segmented from a continuous speech of KDSC dataset. Similar study is performed with consonants by extracting spectral features, namely, MFCCs, flux, centroid, rolloff, and first two formant frequencies. Prosodic features are not considered while classifying dialects from consonants as they have shorter duration. Further, the *cases* (Vibhakthi Prathyays) in Kannada are used to observe morphological peculiarities among the dialects. In this thesis, an attempt is made to propose ADI systems by using words with different *case* information. Dynamic local and global changes in pitch and energy features are extracted from the *case* based words from Kannada datasets. Kannada language has a unique use of special *cases*. These features are said to carry dialectal variations. Hence, a dialect recognition system is proposed by making use of these distinct features of Kannada language.

2.7 General Dialect Datasets used in this thesis

This section provides details of newly proposed Kannada dialectal dataset and procedure employed for its collection. A brief introduction of standard IViE English dataset is also given.

2.7.1 Kannada Dialect Speech Corpus (KDSC)

Present work focuses on development of new dialect dataset, with five prodigious dialects representing five diversified geographical regions with unique speaking styles of Karnataka. A new dataset proposed consists of five dialects; namely central Kannada (CENK), coastal Kannada (Karavali, CSTK), Hyderabad Kannada (HYDK), Mumbai Kannada (MUBK) and southern Kannada (STHK) dialects.

Kannada is a highly agglutinative (concatenative) and morphologically rich language with the influence of Sanskrit on it. Similar to other Dravidian languages, the agglutinative property includes the creation of new words with suffixes and prefixes to the root word. Hence, complex words are formed by adding morphemes together (meaningful word elements), without changing them in spelling or phonetics. Morpho-syntax is determined by the order in which suffixes get attached

Table 2.6: Details of Kannada Dialect Speech Corpus (KDSC) (Spontaneous Speech)

Sl.No.	Dialect Name	Age (Years)	Number of participants					Total Duration (Minutes)
			Male		Female		Total No. of Speakers	
			No.	Dur.	No.	Dur.		
1.	CENK Central Kannada	20-85	18	65	12	47	30	112
2.	CSTK Coastal Kannada	15-70	19	64	15	68	34	132
3.	HYDK Hyderabad Kannada	14-90	25	75	12	45	37	120
4.	MUBK Mumbai Kannada	25-80	12	85	14	45	26	130
5.	STHK Southern Kannada	21-76	16	78	13	50	29	128

to the root word. Kannada is an example of a verb-final inflectional language including relatively free word order (Rajapurohit, 1982). Kannada language has 49 phones of which, 14 are vowels (long and short) and 35 are consonants. Vowels may appear individually, whereas, individual consonants only appear at the end of the words; known as dead consonants, otherwise consonants always appear in combination with vowels. Vowels are often observed as carriers of dialectal variations in Kannada than consonants (Arslan and Hansen, 1996; Zhenhao, 2015). Kannada language includes the existence of retroflex consonants, the presence of long and short vowels, exclusive use of cases (Vibhakthi's), excessive presence of vowel harmony, etc. Kannada language also exhibits consonantal contrasts borrowed from both Sanskrit and Indo-Aryan languages (Prahallad et al., 2012).

The text-independent Kannada dialect speech dataset is recorded from various parts of the identified dialectal regions, mostly from the rural and interior places of the state of Karnataka in India. It is obvious that dialectal cues are naturally present to a great extent in spontaneous speech than in text-read speech (Rouas, 2007). Spontaneous speech has obvious prosodic cues such as different speaking rates, filled pauses, intensity and intonation patterns, hesitations, repetitions, and partially spoken words, etc. (Liu et al., 2010). Normally, these features heavily convey dialectal variations. Even the native dialectal cues are found to be varied because of a change in socioeconomic status and educational background of the speakers. Hence, while recording, due care has been taken to ensure that the speakers chosen have less education and are from rural areas, to effectively reduce the influence of pseudo accents of city-bred; the recordings are kept as pristine as possible. Details of dataset recorded are given in Table 2.6.

During the design and collection of a new Kannada dialect speech corpus, potential influences of several factors such as social differences between speakers, demographics of the speaker, etc. need to be controlled (Clopper and Pisoni, 2006). In this study, the demographics of the speakers are ensured by considering the following aspects during recording: 1. Age above 20 years, 2. Balanced gender ratio, 3. Minimum education, 4. Resident in the same place for more than ten years, 5. Native speakers of the dialect and with parental history. Along with these aspects, the recording equipment used and the environment in which the recording is done, does influence the quality of the recordings and speaking styles. The recording is done in a relatively quiet outdoor environment using a Sony recording device with a sampling rate of 44.1 kHz. Pre-processing is done to remove unnecessarily prolonged long pauses. Short and relevant pauses are retained to ensure necessary naturalness and intelligibility. In this work, recorded Kannada dialect dataset is used for implementation of dialect identification system. Both conventional and non-conventional features are extracted to analyze the significances of these features on Kannada dialects. Further, this dataset is used for evaluating the dialectal differences from Kannada vowels, consonants, words, sentences, and *cases* utterances.

2.7.2 IViE English Dialect Dataset

Intonational Variation in English (IViE) speech corpus consists of nine dialects of British English, spoken across various regions of the British Isles. The speech dataset has been collected with the intention of investigating cross-varietal and stylistic variations in English intonations across nine dialects. The nine dialec-

Table 2.7: Details of IViE English dialect speech corpus (Semi-spontaneous speech)

Sl. No.	Region	Dialects	Number of Speakers (Male+Female)	Read mode (Dur. in Min)	Semi-read Mode (Dur. in Min)
1	Belfest	BELF	12(6+6)	52	32
2	Bradford	BRDF	12(6+6)	49	31
3	Cardiff	CRDF	12(6+6)	49	35
4	Cambridge	CAMB	12(6+6)	51	37
5	Dublin	DUBL	12(6+6)	48	33
6	Leeds	LEDS	12(6+6)	51	31
7	Liverpool	LVRP	12(6+6)	48	26
8	London	LOND	12(6+6)	50	38
9	Newcastle	NECL	12(6+6)	53	31
Total duration:				∼ 8 hours	∼5 hours

tal regions included are: Belfast (BELF), Bradford (BRDF), Cardiff (CRDF), Cambridge (CAMB), Dublin (DUBL), Leeds (LEDS), Liverpool (LVRP), London (LOND), and Newcastle (NECL). The complete process of recording has been introduced by the experimenters, to the speakers before the start of recording. The 'Cinderella' story is readout, from a printed script, by female and male speakers, and has been recorded in a studio environment. Recording has been done, in both read and semi-spontaneous mode, from 12 subjects (6F + 6M adolescents) representing each dialect (Grabe and Post, 2002). Both, read and semi-spontaneous datasets are available separately. Details of IViE corpus are given in Table 2.7. Sizes of the read and semi-spontaneous dataset is approximately 8 and 5 hours respectively. In this research work, standard internationally known English IViE dataset is used for implementation of various dialect recognition systems from longer and shorter speech units. This dataset is used for comparison of dialect identification performance achieved with Kannada dialects. Further, this dataset is used for evaluating the dialectal differences from English words and sentence utterances.

2.8 Summary

In this chapter, critical review of the existing works, related to different dialect processing systems from the perception of datasets, speech features, and classification algorithms has been discussed. Various dialect datasets that are proposed in the literature have been listed along with details and typical properties. At the end of the chapter, problem statement is formulated with identification of clear objectives for the present thesis.

Existing dialect related studies reported in the literature using excitation source, spectral, prosodic, spectro-temporal features are discussed. Various classification algorithms that are employed for development of dialect recognition systems have been briefed along with discussion on single and ensemble classification methods. It has been observed from overall review of the literature, the excitation source and the dialect-specific prosodic attributes of speech, have not been explored much for dialect identification task. It is also noticed from the literature that, majority of studies have used single classifier based algorithms for dialect classification where, the studies proposed on multiple classifier based ensemble algorithms are lacking. In order to address these issues, the next chapter presents the details of the dialect recognition systems proposed, by using the conventional speech based features, along with use of ensemble classification algorithms.

CHAPTER 3

Dialect Identification using Conventional Speech based Features

An introduction to the dialect identification system along with the description of datasets, features and classification models, is provided in the previous two chapters. This chapter includes the use of various conventional acoustic-phonetic speech features for dialect identification. Details of features extracted from sub-segmental, supra-segmental, and segmental levels of speech in order to capture the dialectal cues are briefly discussed. Implications of spectro-temporal variations across dialects are discussed. Different classification methods employed in the present work are briefly covered. Experimental details of ADI systems developed on Kannada and English dialect datasets are provided. Analysis of, dialect identification results obtained using individual and combinations of features with different classification methods is done.

3.1 Introduction

Intrinsic factors such as dialects, gender and vocal tract system contribute significantly to the speech variabilities. In the present chapter, dialect discriminating characteristics of speech are discussed. An attempt is made to use conventional speech features such as source, system, and prosodic features to identify language dialects. Dialect identification task primarily deals with recognition of dialects of a presumed language from the uttered speech. This chapter focuses on identification of dialects from unconstrained or unrestricted audio signal, that contains an unknown text, gender, and speaker. In this work, four varieties of dialectal cues extracted through conventional excitation source, spectral, prosodic, and spectro-temporal features are discussed.

Production of speech units by humans involves a unique articulatory configuration of the vocal tract system and excitation source (Nandi et al., 2015). This hints that contribution of both vocal tract system and excitation source are significant in the production of speech. Hence, sub-segmental level information such as epoch locations corresponding to instances of excitation are extracted along with the slope of the epoch, the strength of epoch locations and instantaneous frequency features. Further, traditional LP residual of the original speech signal is processed to extract MFCCs and are named as Residual Mel Frequency Cepstral Coefficient (RMFCC) features (Prasanna et al., 2006). Spectral analysis of speech is performed to capture dynamics of the vocal tract system through Mel Frequency Cepstral Coefficients (MFCC), Shifted Delta Coefficients (SDC), spectral flux, and entropy features. Dialectal prosodic features, namely pitch and energy are extracted from longer units of speech. This work also investigates the use of Gabor features; also known as biologically inspired features. These are well known features to exploit spectro-temporal variations. Gabor filter based features are popularly used in image processing applications. These features are derived from a filter bank of two dimensional Gabor functions and are used for the classification of Kannada and English dialects.

From the literature, it is observed that, a systematic study has not been conducted on the use of excitation source information and spectro-temporal features for dialect processing. Hence, this work, has developed a system using excitation source and 2D Gabor features for dialect identification. Experiments have been carried out using single classifier based SVM, decision trees and SVM based multi-classifier based classification techniques. Performance evaluation of the above mentioned features and classification techniques is done on Kannada KDSC and English IViE datasets.

The present chapter is organized in the following order. In section 3.2 various conventional acoustic-phonetic features extracted and used for dialect identification are discussed. Section 3.3 provides the detailed information of the various classification models that are employed in this work. Section 3.4, includes the discussions on the results. Section 3.5 summarizes the contents of the chapter.

3.2 Conventional Acoustic-Phonetic Speech Features

Acoustic-phonetics features attempt to investigate the time and frequency domain parameters of speech signal. Features such as duration, pitch (fundamental frequency), mean squared amplitude, frequency spectrum, combined spectro-temporal features, excitation source features and other articulatory features are few representatives of the acoustic-phonetic information (Arslan and Hansen, 1996). Four varieties of acoustic-phonetic features extracted from the speech for classification of different dialects are discussed in following sub sections.

3.2.1 Excitation Source Features

Quasi-periodicity of air pulses generated through the vibration of vocal folds is the major source of excitation during the production of voiced speech. Epoch is the impulse-like signal caused due to vocal folds' closure within a pitch period during the speech production. Glottal pulse, the signal representing the vocal activity is said to be a source for the vocal tract excitation. Among these activities the significant excitations are supposed to take place at epoch locations. The attributes extracted from and around identified epochs contain substantial language-specific information (Rao and Nandi, 2015). Hence, accurate detection of epoch locations plays a prominent role in the estimation of the correct information (Nandi et al., 2017). LP residual and ZFF signals represent the source information fairly well (Prasanna et al., 2006). However, epoch extraction from LP residual signal is not preferred as it contains peaks of random polarity around the epochs resulting ambiguous locations (Kodukula, 2009). Hence, in this work, ZFF based method is used for estimation of the epochs. ZFF approach is also said to be robust to degradations caused due to white, vehicle and babble noise.

Excitation Source Features extracted using ZFF: The algorithm proposed in (Murty, K Sri Rama and Yegnanarayana, B, 2008) is used to generate the ZFF signal from the given input signal. ZFF signal is obtained by passing the original speech signal through the zero-frequency resonator twice. This resonator attempts to mitigate the effects of higher frequency resonances. The resulting signal is termed as ZFF signal, which is a filtered signal varying as a polynomial function of time. Positive zero crossings found in the ZFFS represent the Glottal Closure Instants (GCI) locations nothing but the epoch locations in the signal (Yegnanarayana and Murty, 2009).

The steps followed in ZFF extraction are as follows:

1. Bias of low frequency that is varying with time is removed by taking difference of input signal.
$$x(n) = s(n) - s(n-1) \tag{3.1}$$
2. Signal is passed through a cascade of two ideal zero-frequency (digital) resonators at 0 Hz
$$y(n) = \sum_{k=1}^{4} a_k y(n-k) + x(n) \tag{3.2}$$
where $a_1 = +4$, $a_2 = -6$, $a_3 = +4$, $a_4 = -1$.
3. Trend is removed by subtracting the local mean computed over the average pitch period, at each sample.
$$\hat{y}(n) = y(n) - \frac{1}{2N+1} \sum_{n=-N}^{N} y(n) \tag{3.3}$$
where, $2n + 1$ is the average pitch period of a long speech segment. It can also be considered as the size of the window which is used to calculate the local mean. $\hat{y}(n)$ is the ZFF signal.

The ZFF signal is processed further to extract the following features.

Epoch Locations: There are moments during the production of speech which contain most of the information regarding the excitation of the vocal tract; these moments are called epoch locations. These moments present themselves in the form of glottal closures (Kodukula, 2009).

Strength of Epoch (SOE): The regions of vocal fold vibration commonly known as glottal activity is measured by using the strengths of excitation of epochs. The difference between the consecutive frames, right before and after the epoch locations is used as the strength of epoch (Kodukula, 2009).

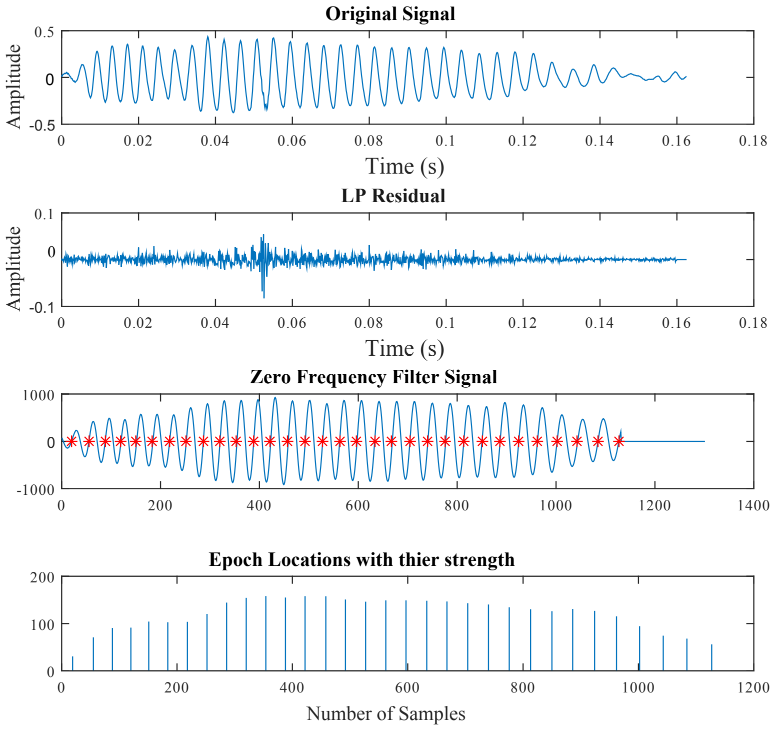


Figure 3.1: Extraction of epoch locations and their strength from utterance ”Cinderella” uttered in the dialect.(a) Original speech signal, (b) LP residual, (c) ZFF with epoch locations and (d) Strength of epochs

Instantaneous Frequency (IF): Instantaneous frequency is computed as the reciprocal for the absolute difference between two following epoch locations. It represents the number of times the glottal folds vibrate in the given unit time.

Slope of Strength of Epoch (SSOE): This feature is extracted to explore the dynamic behavior of epoch parameters for differentiating the dialects. The running slope of the strength of epoch is computed to study the impulsive behaviors of the identified epochs. The slope of the strength of epoch, in the windows of 5 succeeding epochs locations is considered to obtain strength. The difference of

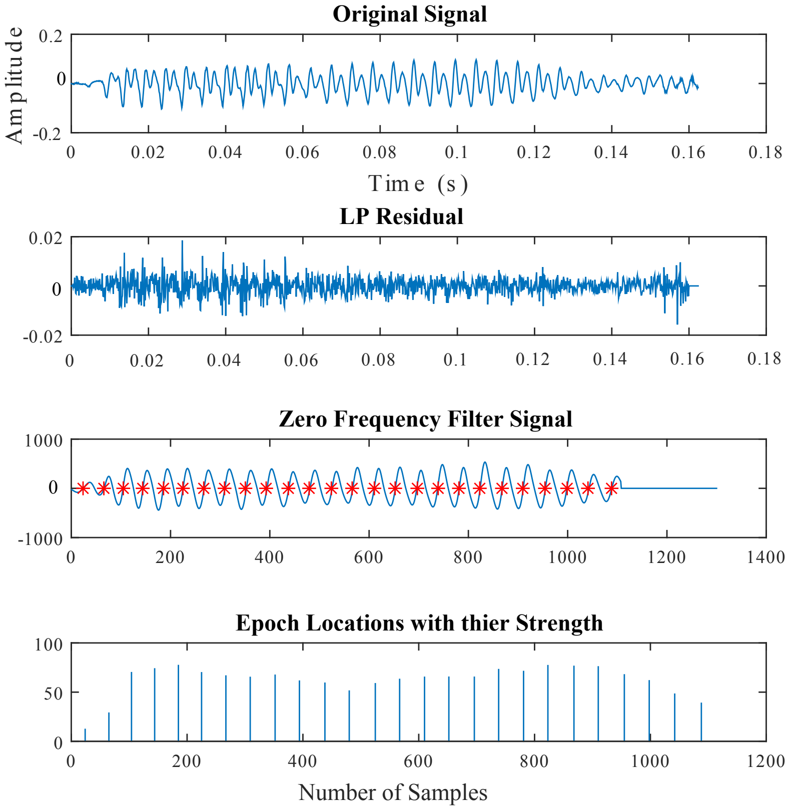


Figure 3.2: Extraction of epoch locations and their strength from utterance ”Cinderella” uttered in the dialect. (a) Original speech signal, (b) LP residual, (c) ZFF with epoch locations and (d) strength of epochs

strength of epoch of the first and the last epoch locations in the window is taken. Then to get the next slope, the window is shifted to the right, by one epoch location. Figure 3.1 and Figure 3.2 present the variations observed in the speech signals of two English dialects. Epoch locations and strength of an epoch location extracted from LP residual and ZFF of the speech signal are shown in the figure.

Excitation Source Features Extracted from LP Residual: The production characteristics of the sound unit can vary across dialects because of slight differences in the co-articulation effects. Similarly, the characteristics of vocal folds' vibration can also contain some dialect specific information. This is the motivation to use excitation source information for dialect processing (Makhoul, 1975).

3.2.2 Spectral Features

Vocal tract system behaves as a time varying resonator or as a filter during the process of speech production. Time varying characteristics are modeled from the variations in the vocal tract shape manifested in the form of resonances and anti-resonances. Generally, behavior of vocal tract resonator is captured through spectral analysis of a speech signal.

MFCCs: Majority of the time, vocal tract information is modeled through MFCC features. The human auditory system follows the nonlinear way of mapping of speech frequency components onto areas on cochlear membrane. According to existing literature, the lower frequency components of speech signal carry useful phonetic information. A nonlinear mel scale filter is used to weigh a lower frequency components more (Tsai, Wuei He and Chang, Wen Whei, 2002). Majority of dialect recognition systems in the literature have used these MFCC features. In speech processing, The mel frequency cepstrum represents the frame wise short term power spectrum of a speech signal. Linear cosine transform of the power spectrum to a nonlinear mel frequency scale is done using the following formulation, where, frequency f is converted into mel frequency m.

$$m = 2,595 \log_{10}\left(\frac{f}{700} + 1\right) \tag{3.4}$$

The steps used for obtaining MFCC features from the speech signal are as follows:

1. Pre-emphasis of the speech signal is carried out. This filtering operation emphasizes the higher frequency components of speech signal.
2. Speech signal is divided into a sequence of short frames of size 20 ms and a shift of 10 ms. each time articulatory movement of vocal tract during that some amount of time is zeroed. Also each frame is Hamming windowed to reduce the edge effect.
3. Discrete Fourier transform is performed on each frame, to compute the magnitude spectrum.
4. Obtained DFT signal is passed through a Mel filter bank, to compute the Mel spectrum. Mel frequency coefficients are computed by multiplication of

the magnitude spectrum with each of the triangular mel weighting filters.

5. Finally, log operation is performed and Discrete Cosine Transform (DCT) is applied on the log Mel spectrum to compute the cepstral coefficients. These are known as Mel Frequency Cepstral Coefficients (MFCCs).

In this work, RASTA (Relative Spectra) processed MFCC features (12+1 frame energy) are extracted from 20 ms frame with 10 ms overlap from 40 filter banks (Hermansky and Morgan, 1994). MFCCs are trusted to be the best available representations of the gross characteristics of the vocal tract system. These spectral features are widely used in many speech-based applications since they try to imitate the human auditory system (Liu and Hansen, 2011). RASTA filter based processing of speech for finding MFCCs, helps in suppressing noisy portions of the speech. The RASTA filter is a band-pass filter that uses the transfer function H_{RASTA},

$$H_{RASTA}(z) = 0.1z^4 \left[\frac{2z+z^{-1}-z^{-3}-2z^{-4}}{1-0.98z^{-1}}\right] \tag{3.5}$$

In this filter, each log filter-bank magnitude component f[m,i], where i takes values from 1,...N (No. of channels) is filtered by using $H_{RASTA}(z)$ function and which produces the RASTA filtered log filter bank magnitudes $f_{RASTA}[m, i]$. This approach attenuates all frequency components less than 1 Hz and above 10 Hz. As a result of analysis of artifacts, the low-pass filtering assists in smoothening of spectral changes that exist in adjoining frames (Kotnik et al., 2002). Later, speaker normalization is done by using the cepstral mean subtraction method. In addition to MFCCs, SDC features are also derived to capture temporal dynamics existing among dialects (Liu and Hansen, 2011). SDC features are computed through stacking up of delta cepstra combined across an aggregation of frames. The basic set of coefficients, $c_j(t)$ is calculated, where j=1,2..N-1, at frame t and j is the dimension index and N is the number of cepstral coefficients explored. The SDC features are associated with four parameters, N, d, P, k and can be expressed as shown in the following equation:

$$S_{iN+j}(t) = c_j(t + iP + d) - c_j(t + iP - d), i = 0, 1, 2, ...k - 1 \tag{3.6}$$

Where, d represents the time delay and advancement used for computing delta parameters which is the time difference between frames. P corresponds to the time shift between consecutively computed delta blocks. k gives the number of blocks for which delta coefficients are concatenated (Hermansky and Morgan, 1994; Liu and Hansen, 2011). For every audio clip, SDC feature vector, is estimated for each frame of time t, considering the context across $P * k$ frames based on chosen values. From this, 13 MFCCs and 26 SDC features are obtained (Liu and Hansen, 2011).

Spectral Flux (SF): Timbre is a speaker specific information available in speech utterances. Spectral flux is one of the ways of extracting timbre information. Spectral flux usually corresponds to a perceptual roughness of a sound. In this work, flux feature is computed and used to measure the spectral changes existing between two successive frames. It is computed by extracting the power spectra of one frame against the same of the previous one (Giannakopoulos and Pikrakis, 2014). Square of the values is taken to avoid negative sign in summation. The following equation is used,

$$Fl_{(i,i-1)} = \sum_{k=1}^{Wf_L} (EN_i(k)) - (EN_{i-1}(k))^2 \tag{3.7}$$

Where $EN_i(k) = \frac{X_i(k)}{\sum_{l=1}^{Wf_L} X_i(l)}$, here $EN_i(k)$ is the k^{th} normalized DFT coefficient at the i^{th} frame, Wf_L is the frame size.

Spectral Entropy (SE): Spectral entropy of a signal measures the distribution of spectral power. It is computed using Shannon entropy concept. Spectral entropy also captures abrupt changes occurring in the energy level of an audio signal (Giannakopoulos and Pikrakis, 2014). While computing spectral entropy of a frame, corresponding spectrum is divided into L sub-bands (bins). The energy E_f of the f^{th} sub-band, for f = 0, . . ., L-1 is calculated using equation (3.8). Then, energies of all bins are normalized by dividing them with the spectral energy of the whole frame, i.e., $ef = \frac{E_f}{\sum_{f=0}^{L-1} E_f}$, the entropy of each normalized energy value is calculated using the equation (3.9)

$$E(i) = \frac{1}{Wf_L} \sum_{k=1}^{w_L} |x_i(k)|^2 \tag{3.8}$$

$$H = -\sum_{f=0}^{L-1} ef. \log_2(ef) \tag{3.9}$$

In this work, the value of L is set to 10 indicating that each frame is divided into 10 bins. The entropy value can be lower if there are abrupt changes in the energy envelope in the frame. These variations in energy may be useful to discriminate dialects.

3.2.3 Prosodic Features

Prosodic features attribute few characteristics to speech in order to make it more natural and legible. Prosodic features such as pitch, controlled modulation of pitch known as intonation, compaction or prolongation of few speech units, imposing stress on few sound units while pronouncing them, melody, and so on are responsible for naturalness of the language/dialect (Ramus and Mehler, 1999). Rhythm, stress and intonation features are complex perceptual entities usually expressed through duration, energy, and pitch respectively. Generally, prosody features play a prominent role in the perception of different dialects, since majority of the dialects demonstrate variations in both pitch and energy values. In this work, prosodic features are extracted from longer frames of size 50 ms.

Pitch: This is also known as fundamental frequency (F0), one of the perceptual properties of speech. Higher the pitch, sharper is the speech. F0 is a physical correlate of the pitch and is the rate of vibration of the vocal folds. Different F0 ranges are observed among the speakers of different dialects. The auditory pitch perception influenced by the harmonic structure and amplitude plays a role in distinguishing various pronunciation styles across dialectal regions (Ramus and Mehler, 1999; Wightman, 1992). Subharmonic-to-Harmonic ratio based pitch estimation algorithm is used in this work to obtain frame-wise pitch values. This algorithm is based on perception of pitch from alternate pseudo cycles in speech. The algorithm employs a logarithmic frequency scale and a spectrum shifting technique to obtain amplitude summation of harmonics and subharmonics respectively (Sun, 2000).

Energy: Frame-wise energy values associated with speech signal vary with respect to time. Energy corresponds to the loudness of speech. Normally, short time energy is computed from phonetic units for assessing loudness. Energy plays a prominent role in human aural perception. Energy variations with time are measured concerning to each samples' amplitudes within a frame. Often, stressed speech is attributed to high energy values. In both dialect speech corpora, reflected variations in energy profile are used since dialects follow varying stress patterns of pronunciation. Hence, frame level energy is considered as a feature for identification of dialects. Short-term energy is calculated as per the formulation given in equation (3.10).

$$E(i) = \sum_{n=1}^{w_L} |x_i(n)|^2 \tag{3.10}$$

here $x_i(n)$, n = 1, . . . , W_L is the n^{th} audio sample in the i^{th} frame, where W_L is the length of the frame. Average energy is obtained by 3.11.

$$E(i) = \frac{1}{W_L} \sum_{n=1}^{w_L} |x_i(n)|^2 \tag{3.11}$$

3.2.4 Spectro-Temporal Gabor Features

Spectro-temporal aspects of speech are effectively modeled through the use of 2D Gabor features. Gabor features are commonly known as biologically-inspired features and are familiar in image processing applications. These features successfully decompose the spectro-temporal power density into components of spectral, temporal and joint spectro-temporal modulation patterns (Lei et al., 2012). Gabor features are said to model specific stimuli and which are sensitive to the neurons of the mammalian auditory cortex (Meyer, Bernd T and Kollmeier, Birger, 2011).

Various stages involved in the process of 2D Gabor feature extraction are given in Figure 3.3. 2D Gabor features are computed through convolution of the log-Mel spectrogram of the speech signal (input) with a set of 2D Gabor filters. Logarithmic compression and Mel-frequency scale are considered, since they are popular and are straightforward approach in audio processing. A 2D Gabor filter bank is designed to provide approximately uniform coverage of the spectro-temporal frequency modulations. Each Gabor filter is mathematically represented as a product of complex sinusoidal carrier function with corresponding envelope function. Generally, Gaussian or Hann functions are used as envelope functions. In the present work, Hann envelop function is used for designing Gabor filter. Eq. 3.12 gives the Gabor filter.

$$g(m, q, m0, q0, \omega_m, \omega_q, \nu_m, \nu_q) = s(m, q).h(m, q); \tag{3.12}$$

Where, ω_m and ω_q are frequencies of spectral and temporal modulation, ν_m and ν_q are the number of semi-cycles of envelope function in both spectral and temporal dimensions. Carrier function and envelope function are computed by using equation 3.13 and 3.14.

$$S(m, q) = s_{\omega_m}(m - m_0).s_{\omega_q}(q - q_0) \tag{3.13}$$

$$h(m, q) = \frac{h_{\nu_m}}{2\omega_m}(m - m_0).\frac{h_{\nu_q}}{2\omega_q}(q - q_0) \tag{3.14}$$

The complex Sinusoidal carrier function and Hann envelope function is given in equation 3.15 and 3.16.

$$S_{\omega_i}(x) = exp(j\omega_i x) \tag{3.15}$$

$$h_a(x) = \begin{Bmatrix} 0.5 - 0.5\cos(\frac{2\pi x}{a}) & ; \frac{-a}{2} < x < \frac{a}{2} \\ 0 & Otherwise \end{Bmatrix} \tag{3.16}$$

In this work, 23 frequency channels and 40 spectro-temporal filters are used resulting in 920 features. Normally, the filters placed adjacently, exhibit minimal/no features changes, thus the redundant features may be avoided by selecting the specific feature channels at appropriate distances. This approach reduced the dimensionality of the feature vector to 311. Hence, the center frequency channel is selected for designing each modulation filter and the channels with overlapping Gabor filters are messed to get the reduced feature dimension (Meyer et al., 2011). Real components of the 2D complex on Gabor filters are used as features.

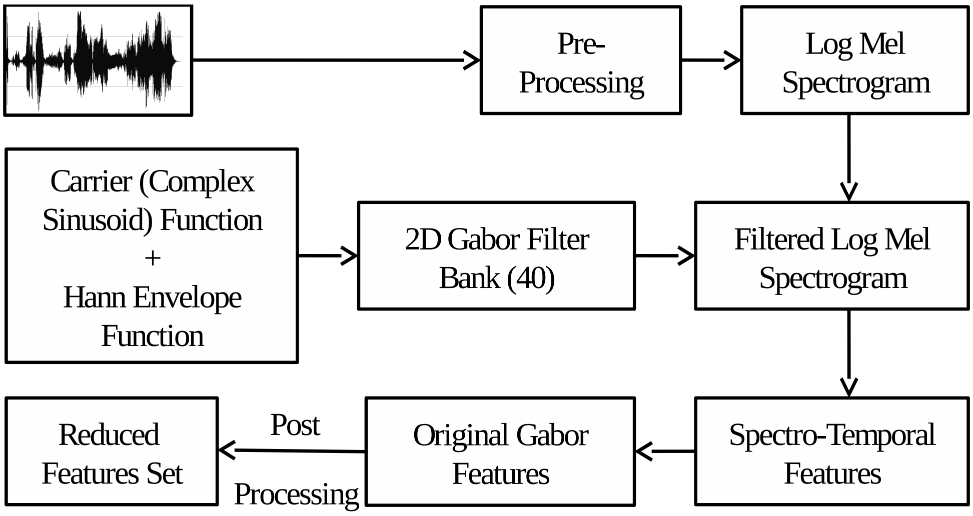


Figure 3.3: Steps in Gabor Filter based feature extraction

3.2.5 The i-vector Features

The i-vectors, which are being commonly used in several state-of-the-art dialect identification systems are extracted to compare the results. These i-vectors have been quite successful in speaker and language identification systems (Dehak et al., 2011b). They are derived from traditional 39 RASTA processed MFCC+SDC features. The i-vectors are represented as a low ranked matrix T that captures variabilities concerning total variability matrix and are implemented using UBM-GMM model.

GMM-UBM Model:

The basic idea is to adapt Universal Background Model (UBM)[2] to a set of given speech frames. Estimation of utterance dependent GMM is done through the Eigenvoice adaptation technique (Dehak et al., 2011b). The GMM super-vector is obtained by stacking all mean vectors from the Gaussians of a given utterance. The i-vector is modeled as follows (Sadjadi et al., 2013):

$$M = m + T\omega + \epsilon \tag{3.17}$$

where M is speaker and session dependent UBM super-vector is derived from either MFCCs or MFCC-SDC features and m is speaker and session independent super-vector taken from UBM. T is a rectangular lower rank matrix called total

[2] A UBM is a large GMM trained to represent the speaker-independent distribution of features

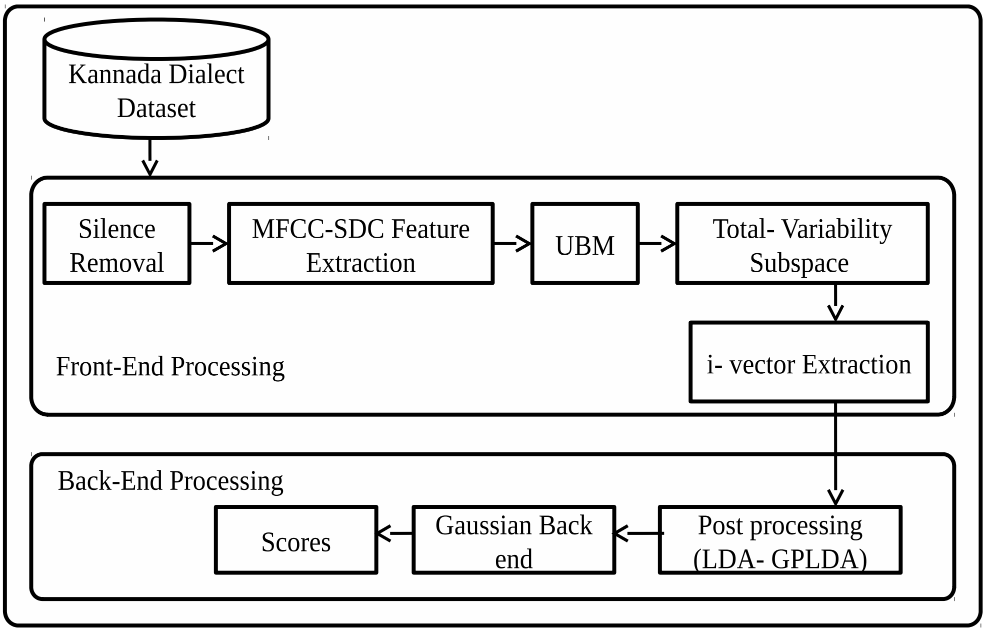


Figure 3.4: Steps involved in GMM-UBM based i-vector feature extraction

variability matrix and ω and are total vectors or i-vector features. The residual noise term is shown as ϵ. Conceptually, i-vector is said to capture the sequence summary of a given utterance in terms of both speaker and session variabilities. However, it follows a computationally intensive procedure. Various stages, followed during the extraction of GMM-UBM based i-vectors, are presented in Figure 3.4.

Initially, the i-vector feature extraction procedure requires to choose the number of mixtures and iterations for building UBM model. In this work, 512 mixtures for UBM building with 10 iterations for every mixture are selected heuristically. From the trained UBM, a Total Variability (TV) matrix is also trained with the same data. This, i-vectors of 300 dimension are computed from variability matrix to obtain the fixed length feature vectors from the varied length audio files. Expectation-Maximization algorithm is employed to train both UBM and total variability matrix models. Further, the i-Vectors for both train and test sets are extracted and Gaussian Back-end algorithm is used for five class dialect modeling. In the present work, the MFCC-SDC and i-Vector feature based ADI system is designed using GMM classifier.

3.3 Classification Algorithms

In the present work, a single classifier based SVM, and multi-classifier based ensemble algorithms are used for dialect classification. Two different types of ensembles algorithms are used by combining decision tree and SVM as base classifiers. Three decision tree based ensemble algorithms are used in this work. They are Random Forests (RF), Extreme Random Forests (ERF), and Extreme Gradient Boosting (XGB). The fourth one is ensemble based SVM (ESVM) algorithm. A single classifier based SVM method is also used since it has been reported in the literature to show better generalization performance across several applications of speech processing. SVM tries to capture the discriminating parameters across the feature vectors for identification of dialects. Literature gives a few SVM based systems for dialect recognition tasks (Pedersen and Diederich, 2007; Toohill et al., 2012). This section covers details of four different classification methods used in this work.

3.3.1 Single Classifier based Algorithms

In this work, dialect identification system is developed by using single classifier based SVM on Kannada dataset. SVM is modeled to capture dialect specific cues through multiple speech features. Generally, prediction or classification performance is dependent on a single and individually trained model.

Computational complexity of SVM is independent of the dimensionality of the feature space and it is known to exhibit better generalization performance over other classifiers. Five different SVMs are separately trained on five dialects of Kannada using one-versus-rest approach to solve the five class problem. Radial basis function (RBF) is used as a kernel function for separating the examples of each class with the maximal margin in a high-dimensional feature space (Chang and Lin, 2011). Five separate SVM models are trained with individual and combination of features. Training inputs from all five classes are of the form $\left\{\{(x_i,k)\}_{i=1}^{N_k}\right\}_{k=1}^{n}$, where N_k is the total speech inputs belonging to k^{th} dialect class, and k takes five labels k=1,2,3,4,5. SVM for each class k is constructed by using the set of training inputs and the desired outputs as, $\left\{ \{(x_i, y_i)\}_{i=1}^{N_k} \right\}_{k=1}^{n}$, y_i respectively. For training example x_i, the output y_i takes a value $+1$ if $x_i \in k^{th}$ class represents positive example, else -1 represents the negative example. After training, evaluation of the system is performed by giving feature vectors from test speech clips, as inputs to all trained SVM models. For instance, for a test pattern x, the evidence vector $TS_k(x)$, for $k = 1, 2, ..5$ is obtained from all five SVM models. The class label k associated with SVM that gives highest evidence, is hypothesized as dialect class C of the test pattern, i.e. $C(x) = argmax_k(TS_k(x))$.

3.3.2 Ensemble based Classification Algorithms

Ensemble algorithms consisting of a multiple individually trained classifiers and final predictions obtained, are the combination of the results of these individual classifiers. Due to this collective decision, ensemble methods show better predictive performance compared to that of individual classifier (Utami, Iut Tri et al., 2014). In this work, SVM and decision trees algorithms are employed as base learners to derive ensembled decision.

A Decision Tree based Ensemble Algorithms

Three decision tree based ensemble algorithms that use bagging (RF, ERF) and boosting (XGB) techniques are employed in this work. Bagging (bootstrap aggregation) algorithms combine predictions from independent base models derived from bootstrap samples by sub-sampling through replacement in the original data (Breiman, 2001). Boosting follows a dependent fashion in the growth of ensembles. Base models are improved iteratively, to reduce the errors of the ensemble. The logic is that subsequent predictors, learn from the mistakes of the already chosen predictors (Freund and Schapire, 1999).

Decision tree based RF classifier uses a combination of randomized tree predictors with bootstrapping technique used during training. Predictions of many decision trees in this case (empirically chosen 2048 trees), are aggregated (such as majority voting) for final prediction. $\sqrt{n}$ features are used to split a node, where n is the size of the feature vector. Sizes of feature vectors vary with the number of features selected for evaluation. Splitting of a tree node is controlled through the best split decided by Gini criterion among the random subset of features. After construction of different decision trees (forest) for different set of input data, classification is done by taking majority voting from the predictions of these trees (Breiman, 2001).

ERF is an extreme case of RF, where 2048 randomized trees are built by subsampling the input with replacement. Forest structures are independent of the output values of the learning sample. Hyper-parameters such as max_depth, max_features, and n_estimators are empirically assigned to 6, $\sqrt(n)$ (n is the number of features) and 2048 respectively. These hyper-parameter values are chosen after performing several experiments by assigning different values for each parameter (Pedregosa et al., 2011; Geurts et al., 2006). Whereas, this combination of values has resulted in better performance.

XGB uses boosting, that improves the base learner prediction iteratively following a greedy fashion. Here, each additional base learner improves the accuracy by further reducing the loss (error) function. Multi class *logloss* function, shown in equation (3.18), is used in this work.

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j}\log(p_{i,j}) \tag{3.18}$$

Where, N is the size of the feature vector, M is the number of class labels, $y_{i,j}$ of base learner is 1, if observation i belongs to class j, 0 if not. $p_{i,j}$ represents the predicted probability if observation i is in class j. Decision tree classifier is used as the base learner. Construction of trees is done as follows: η a learning rate is assigned to 0.2, which controls the shrinking of feature weight to make the boosting process more conservative. Maximum depth of a tree is limited to 6, subsample ratio of the training inputs is limited, such that 60 percent of data instances are used to grow trees. Objective function softmax is used for handling five classes. These parameters are fine-tuned and selected empirically to get better prediction accuracy. Implementation is done using XGBoost library (Friedman, 2001; Chen and Guestrin, 2016).

B Ensemble based SVM Algorithms (ESVM)

In addition to tree based algorithms, ensemble model with SVM as a base learner is done for designing dialect identification system using bagging technique. Using this, training set $TR = (x_i, ; y_i)|i = 1, 2, 3, 4, 5$ is divided into K training sets to construct K independent SVMs. Here K is chosen to be 2048, empirically. Hence, 2048 training sets are created from TR through random sampling with replacement. Each of these is used with SVM, individually for training. Trained SVMs independently are later aggregated either using majority voting or the probability sum. Since it is a classification problem, after training, the majority voting method is followed in this work, to combine results of individual SVMs. Binary classification based SVM is extended by using one-vs-rest method to handle the multi-class problem (five dialect classes) using libsvm implementation (Pedregosa et al., 2011). RBF kernel is chosen among available kernel functions, that empirically results in better performance. Using ESVM, classification accuracy is improved since combinations of several SVMs incrementally tries to expand the correctly classified area.

3.4 Experimental Results and Discussions

In this work, experiments are carried out to evaluate the performances of ADI systems, implemented using traditional MFCC-SDC, i-vector, prosodic, excitation and spectro-temporal features for dialect identification. The influences of the features are evaluated, individually and in combination to understand dialect classification using late equal fusion mechanism on the KDSC dataset. Different features extracted generally represent different aspects of speech. Hence, extracted acoustic-phonetic features contain non-overlapping properties. In this regard, even dialect-specific cues carried by these features are assumed to be non-overlapping. With this intuition, all four varieties of features obtained are combined.

Experiments are carried out to evaluate the performance of features individually and in combination. In addition to this, the performance comparison, of both single SVM and ESVM is done with the mentioned feature vectors. In the beginning, details of the statistical parameters of the prosodic features namely, pitch and energy on five Kannada dialects are discussed. Later part of the chapter includes, analysis of results obtained with single and ensemble classification methods followed by a comparison study with existing state-of-the-art systems. General work flow diagram of the proposed dialect identification system is given in Figure 3.5. Metric used for performance evaluation is accuracy and it is computed using the following formula:

$$Accuracy = \frac{Total\, number\, segments\, correctly\, classified}{Total\, number\, of\, segments\, used\, for\, testing} * 100 \tag{3.19}$$

The present work, considers the newly proposed Kannada dialect Speech Corpus (KDSC) (Chittaragi et al., 2019) and an internationally known standard Intonational Variations in English (IViE) dialect dataset (Grabe and Post, 2002) for evaluation of the proposed features. A detailed description of both dialect datasets is provided in section 2.7.

3.4.1 Statistical Analysis of Prosodic Features with Kannada Dialects

This section provides a detailed statistical study of three prosodic features, namely, stress, rhythm, and intonation whose direct correlation are pitch, energy, and duration. These parameters obtained from within speech convey important dialect specific information. Few studies on language identification have shown that pitch and energy information may be used if acoustics and phonotactics information in speech is degraded (Mary and Yegnanarayana, 2008). In order to understand the influence of pitch and energy features on discriminating five Kannada dialects, mean values of pitch and energy features from male and female speakers are considered. Box plots are chosen to represent detailed statistics of the prosodic features. Box plots include five statistical parameters namely: median, minimum, maximum, two intervals: one between 25% to 75%, and other between 9% to 91%. Statical parameters obtained with Kannada dialects are presented in the form of box plots as shown in Figure 3.6.

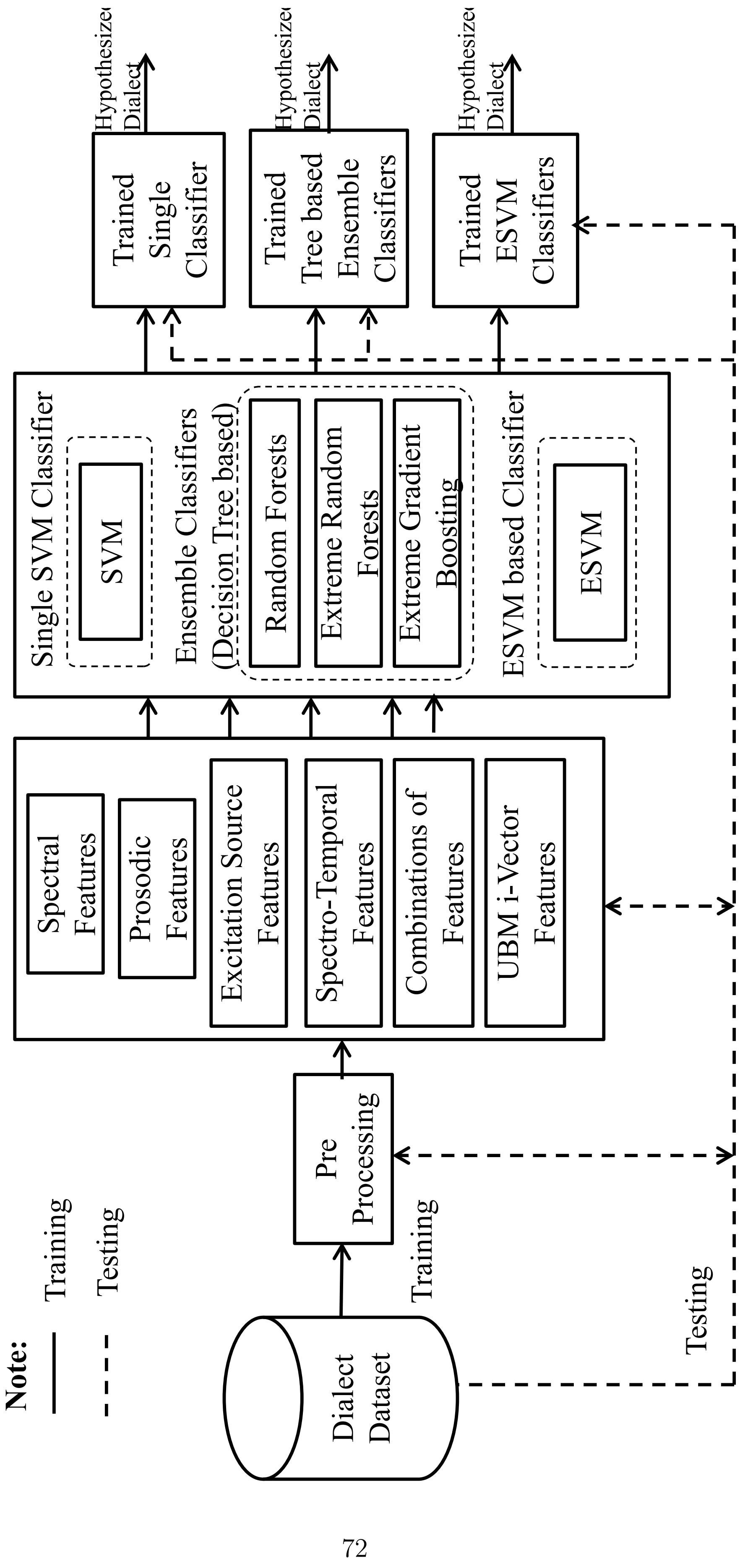


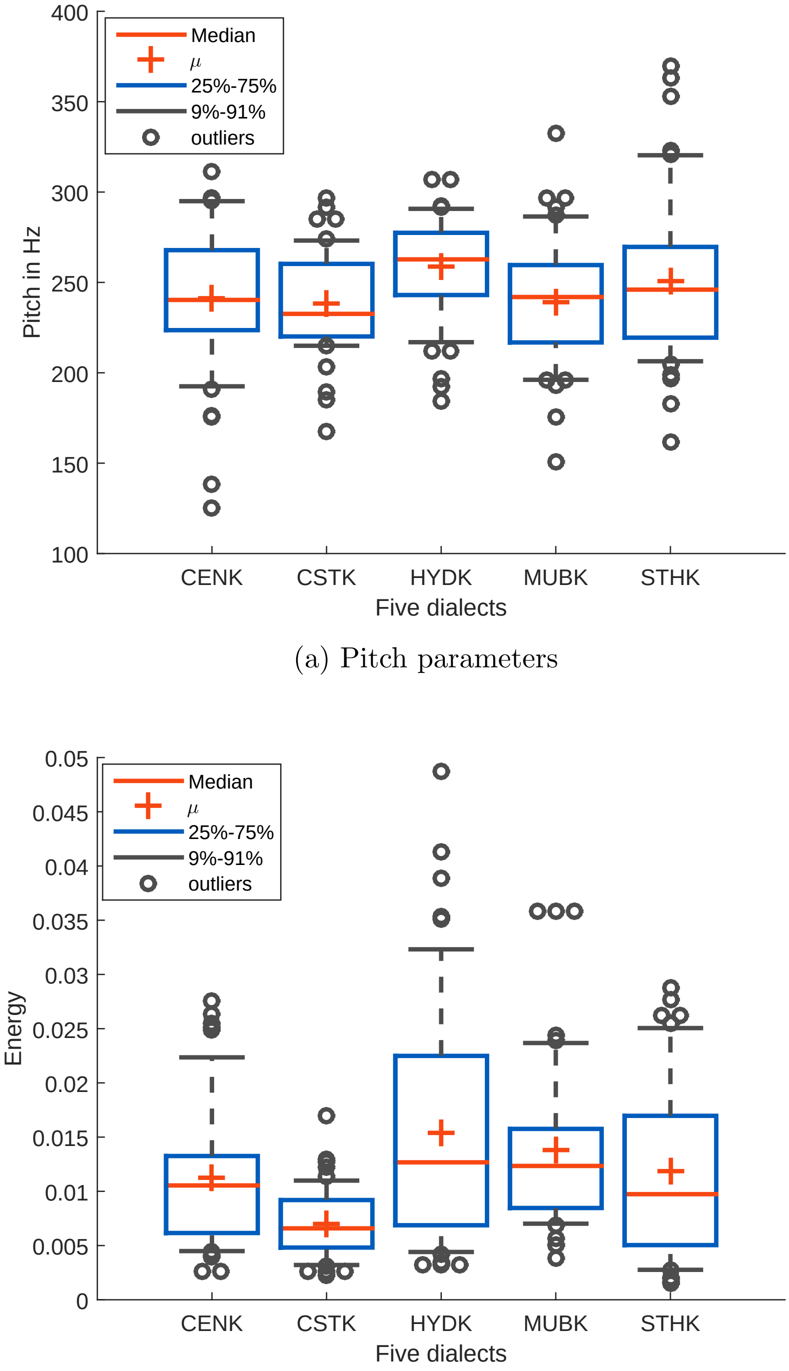
Figure 3.5: Automatic dialect recognition system with different features and classification algorithms

Table 3.1: Mean values of the prosodic features for the speech samples collected from five Kannada dialectal regions

Dialects Regions	Name	Male Speakers			Female Speakers		
		Pitch Hz	Standard deviation	Energy	Pitch Hz	Standard deviation	Energy
Central Kannada	CENK	238	25	0.0121	247	29	0.0092
Coastal Kannada	CSTK	225	22	0.0082	244	36	0.0077
Hyderabad Kannada	HYDK	255	37	0.0186	268	25	0.0109
Mumbai Kannada	MUBK	232	25	0.0149	246	27	0.0101
South Kannada	STHK	248	33	0.0116	265	42	0.0096

In addition to this, the mean and standard deviation of pitch and energy features extracted from female and male speakers of five dialects are presented in Table 3.1. From Figure. 3.6 and Table 3.1 following observations are made. Both pitch and energy parameters are found to be comparatively low in the case of CSTK dialect. The span of both energy and pitch features is found to be smaller for CSTK dialect which is spoken in the coastal region. CSTK speakers are found to follow uniform flat pitch pattern and low energy profile while speaking. This is mainly because Kannada is the second language in coastal region speakers (Tulu is the first language for majority of speakers). Kannada is an acquired language for them and spoken mostly on formal occasions. They are likely to pronounce Kannada words with much consciousness. This results in minor variation in acoustic properties of vowels and consonants across different phonetic contexts, in both read and spontaneous speech.

Speakers of HYDK Karnataka dialect spoken in Hyderabad region have demonstrated higher values of pitch and energy profiles when compared to all the other dialects. The span of energy range is observed to be broader in this region. Also, the speaking rate followed in this dialect is found to be higher and in turn has resulted in deletion of syllables in the final words. Speakers of this region are found to use highly stressed words and syllables. Even though high pitch values are noticed; their span has been observed to be smaller (Soorajkumar et al., 2017). It is also observed from these speaking styles that the people follow unique pronun-



(a) Pitch parameters

(b) Energy parameters

Figure 3.6: Basic statistics for pitch and energy values for speech data collected from five Kannada dialects, CENK:Central Kannada, CSTK:Coastal Kannada, HYDK:Hyderabad Kannada, MUBK: Mumbai Kannada, STHK: Southern Kannada

ciation patterns very different from other Kannada dialects. These variations are mainly due to the influence of two languages Marathi and Telugu spoken in the two neighboring states. This is mainly due to its proximity to Maharashtra and Andhra Pradesh states respectively. Apart from these reasons, Kannada spoken in HYDK dialect is slightly different due to the influence of the former Hyderabad Nizam Kingdom as a fair chunk of Urdu vocabulary is used (Rajapurohit, 1982). The speakers of the Mumbai Karnataka region (MUBK) also have an influence of Marathi language and also have a fair use of Urdu vocabulary while speaking. Even though this region is adjacent to the Hyderabad Karnataka region, unlike them, speaking rate is slow; the pitch values are lesser and energy values are slightly lesser (Soorajkumar et al., 2017). Interesting observations are also made with CENK and STHK dialect speakers as they follow slightly similar speaking patterns. For these regions, Kannada is a primary language and the speaking styles are closer to the written form of language. In specific, CENK dialect is very close to the written/standard form of Kannada where, every utterance is phonetically and clearly pronounced. However, STHK dialect speakers have shown slightly faster speaking styles similar to that of HYDK dialect with higher energy values. Nevertheless, a difference exists with respect to pitch variations between them. Outliers are represented with ”o” and are found in all dialects in the cases of both pitch and energy features. Outliers are slightly more among CENK and STHK dialects.

In general, prosodic, spectral, excitation source features extracted in this work are expected to carry non-overlapping dialect specific features (Rao and Koolagudi, 2011). This is because, spectral features represent dialectal cues from vocal tract system (speech production), excitation source features represent vibration pattern of vocal folds, and prosodic features represent paralinguistic information For instance, higher energy values generally indicate higher pitch values and slower speaking rate. For example, dialects like HYDK and STHK have shown higher energy and pitch range with lesser speaking rate. From this observation, it may be said that prosodic features influence the task of identification of dialects; however

Table 3.2: Dialect recognition performance on KDSC using GMM-UBM and UBM-i-vector features

Feature Vectors (Size)	Models	Accuracy in %
MFCC (13*2)	GMM-UBM	75.5
MFCC-SDC (39*2)		74.9
MFCC (13*2)	UBM - i-vector based system	81.4
MFCC-SDC (39*2)		78.54

these alone are not sufficient to discriminate dialects effectively. In the following subsections, several experiments and their results are reported using spectral, prosodic and source level features. Features are considered individually and in combination for implementation of different ADI systems.

3.4.2 Post-Processing of Features

The raw features, extracted from speech are further processed to derive new feature vectors. The procedure followed is as follows: Speech signal is segmented into M frames and features discussed earlier are extracted from each frame resulting into M*N dimensional matrix, where N is the size of feature vector. These features are named as raw features. Further, statistical processing of these matrices is done to obtain a new feature set called as derived feature set containing statistical parameters namely, mean and standard deviation of original features. These are computed by taking the statistical mean of feature vectors F_i and F_{i+1} obtained from two consecutive frames, to get the first N features. Later, features from N+1 to 2N represent the standard deviation of the values of the same frames. Hence, the dimensionality of the feature vector of each frame is doubled to 2N from N (Giannakopoulos and Pikrakis, 2014; Chittaragi et al., 2018b). The steps involved in feature extraction and post-processing of features are shown in Figure 3.7. These two statistics extracted from two consecutive frames, exhibit similar behavior and a few temporal variations are discarded due to averaging of two consecutive frames. Thus matrix size M*2N is obtained. At this stage, averaging of all M frames is done to get the final derived feature vector of size 2N. However, the N raw features, computed earlier from each frame are noticed to be different

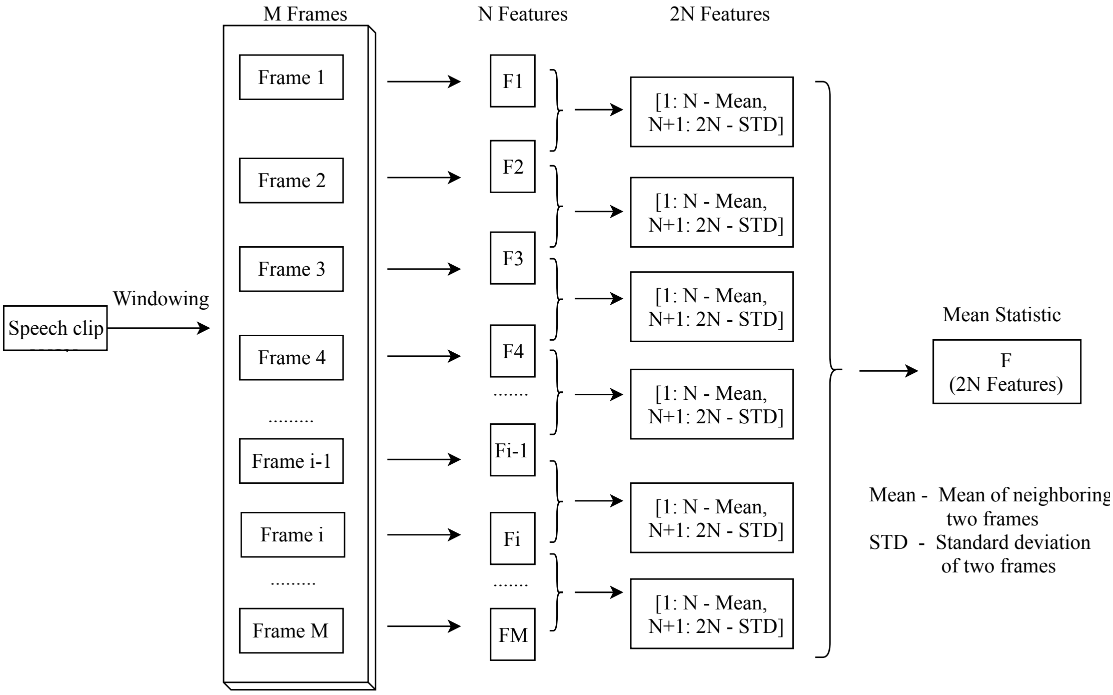


Figure 3.7: Schematic for deriving statistical features from speech signal

from the present 2N derived features.

3.4.3 Dialect Identification Using GMM-UBM Model on KDSC

Initially, experimental results obtained with existing traditional GMM-UBM based model and state-of-the-art UBM based i-vector features are discussed. Results obtained using these are given in Table 3.2. It is observed from the table, that i-vector features with UBM have shown a better performance of 81.4% over traditional GMM-UBM based systems. The combination of SDC features with MFCCs has not enhanced the dialect identification performance. This hints that, significances of temporal variations captured through SDC, may not be much useful when identifying Kannada dialects. In further experiments, only 26(13+13) MFCC derived features are considered.

3.4.4 Dialect Identification using Different Conventional Speech Features on KDSC

In this subsection, the results obtained from a series of experiments conducted using a single SVM classification method on proposed KDSC are reported. Text independent, spontaneous speech recorded from native speakers is used as dataset. To begin with, experiments are carried out using spectral and prosodic features, since these features have demonstrated significant contribution in dialect processing. Later, less explored excitation source cues are used. Finally, bio-inspired Gabor filter based features are used to develop ADI systems.

A Dialect Identification using Spectral and Prosodic Features

These features are known to capture the vocal tract and prosodic information across dialects. Four different kinds of spectral and prosodic features are used for implementation of ADI systems. They are: (1) 13 MFCCs, (2) 13 MFCCs, spectral flux and spectral entropy, (3) pitch and energy features, and (4) The combination of all. The experiments are carried out cross-validation approach. Total data samples are split into k folds, then training is performed on k-1 folds and testing on the one left out fold. This is repeated for all combinations and average of the result on each instance is taken as the final result. In this work, we considered 5-fold cross-validation approach.

Dialect recognition performance obtained from all experiments are tabulated in Table 3.3. Size of the feature vectors is mentioned in the brackets. In addition, the experimentation is conducted with raw (original features extracted from speech signal) and derived (post-processed to derive two statistics from the original feature vectors) feature vectors in all four ways. Further, few experiments are also carried out to analyze the significance of SDC and MFCC features on dialect classification.

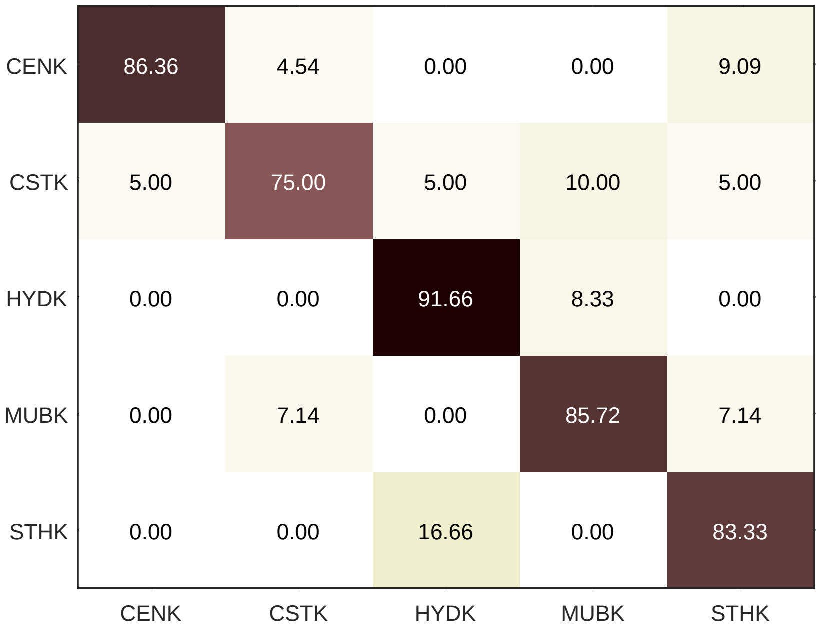
From Table 3.3, it is observed that, 78.50% of recognition accuracy has been obtained with 13 MFCCs as features. Addition of spectral flux and entropy features to MFCCs has given a performance of 79.29% . From these results, it may be noticed that spectral features alone are relatively successful in the classification

Table 3.3: Dialect recognition performance with spectral and prosodic features using single SVM on KDSC. Legend: SF-Spectral flux, SE-Spectral entropy

Sl. No.	Features	MFCC (13)		MFCC+SDC (39)	
		Raw Features	Derived Features	Raw Features	Derived Features
1	MFCC	78.50 (13)	81.25 (26)	60.50 (39)	72.25 (78)
2	MFCC+SF+SE	79.25 (15)	83.02 (30)	61.75 (41)	72.52 (82)
3	Pitch+Energy	37.50 (02)	47.75 (04)	37.50 (02)	47.75 (04)
4	MFCC+SF+SE+ Pitch+Energy	80.75 (17)	**84.41** (34)	64.25 (43)	77.53 (86)

of Kannada dialects, where, MFCC features have successfully characterized the unique vocal tract shape variations followed across five Kannada dialects. Compared to MFFCs, slightly lesser performance is seen with prosodic features on KDSC. Even though prosodic features play a significant role in discriminations of dialects, pitch and energy prosodic features alone cannot sufficiently distinguish Kannada dialects. It is noticed from the results that, the combination of both the features has shown that, there is a slight increase in accuracy of about 80.75%. This is because, spectral features effectively recognize dialect specific cues from vocal tract system and prosodic features encapsulate dialect related perceptual pitch and intensity patterns. Further, from all experiments, it is noticed that the use of derived features has demonstrated an improved dialect performance over raw features. Where highest recognition of 84.41% is observed with a combined feature vector. Derived features have performed better consistently with all four combinations of features. Also, 10.25% increment is noticed with the use of proposed derived features over raw features.

Generally, SDC features are expected to contain additional temporal information (Liu and Hansen, 2011). Further, to verify the influence of SDC features on Kannada dialect recognition, similar experiments are conducted by adding a 26-dimensional feature vector (13 delta and 13 delta-delta) to 13 MFCCs (total 39). Performance of dialect classification with addition of SDC features is presented in Table 3.3. After addition of SDCs, some interesting observations are



	CENK	CSTK	HYDK	MUBK	STHK
CENK	86.36	4.54	0.00	0.00	9.09
CSTK	5.00	75.00	5.00	10.00	5.00
HYDK	0.00	0.00	91.66	8.33	0.00
MUBK	0.00	7.14	0.00	85.72	7.14
STHK	0.00	0.00	16.66	0.00	83.33

Figure 3.8: Confusion matrix obtained on spectral and prosodic features using SVM for Kannada dialect identification

accomplished, where decrease in performance is noticed. In all four approaches, the performance obtained are lesser than the use of 13 MFCCs alone, in Kannada language. From these results, it hints that temporal dynamics, captured through SDCs, may not be much useful for Kannada dialect classification.

The confusion matrix of Kannada dialect classification obtained for the combination of two features is given in Figure. 3.8. It is drawn for the ADI system that gives highest/average accuracy of 84.41% obtained by using derived feature set with SVM classification method. It is worth noting that, HYDK is classified with the highest accuracy of 91.66%. This is because, Kannada spoken in this region is with an unique pronunciation style with higher pitch and energy values than the remaining four dialects. It is more interesting to notice that, misclassification is observed with MUBK (8.33%). As mentioned earlier, these two dialects have more similar speaking patterns and they are also spoken in neighboring regions of Karnataka state. Least performance of 75% is observed for coastal dialect and remaining 25% are misclassified with all four dialects. This could be due to the use of Kannada as secondary acquired language for communication. Also, dataset also may contain the recordings of these migrating from other regions of Karnataka.

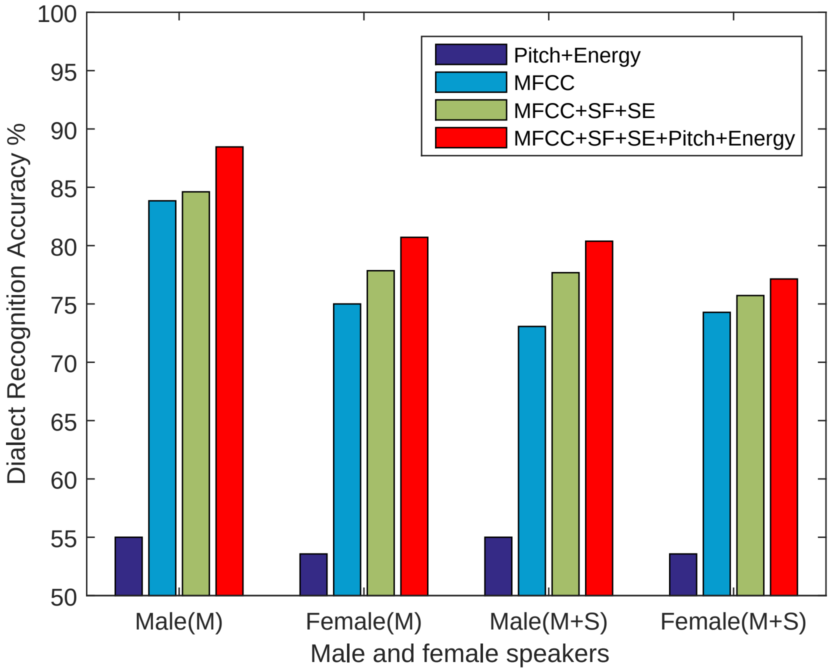


Figure 3.9: Comparison of dialect recognition performance of male and female speakers using spectral features, Legend: SF-Spectral Flux, SE-Spectral Energy, M-13 MFCC and M+S- MFCC+SDC

Hence, speakers from this region with higher pitch and energy values are misclassified with other dialects, slightly more with Mumbai dialect. Mumbai Karnataka and coastal Karnataka are adjacent regions. 83.33% of STHK dialect samples are classified correctly and 16.66% are misclassified with Hyderabad region. This is primarily due to the use of similar speaking styles, with high pitch and energy properties. It is also demographically observed that many migrates from MUBK and HYDK regions to CENK and STHK regions.

We intuitively thought and observed during speech data recording that there would be gender specific dialectal cues due to various influences and social exposers. Hence, gender specific analysis of speech samples is performed with female and male speakers separately on five Kannada dialects. An analysis performed on spontaneously produced speech in a large corpus has revealed that female speakers tend to follow different pronunciations compared to their counterparts. Male speakers normally exhibit filled pauses and repetitions (Clopper and Smiljanic, 2011). Kannada dialects are classified with higher accuracy in the case of male speakers over female speakers in every scenarios considered in this study. Compar-

Table 3.4: Dialect recognition performance using source features with SVM on KDSC. Legend: RMFCC-Residual MFCCs, SOE-Strength of Epoch, SSOE-Slope of SOE, IF-Instantaneous Frequency, SF-Spectral flux, SE-Spectral entropy

Sl.No.	Features	Accuracy in %
1	RMFCCs	78.75
2	SOE+SSOE+IF	39.06
3	RMFCCs+SOE+SSOE+IF	81.25
4	MFCC+SF+SE+Pitch+Energy	84.41
5	4+3	86.74

ison of dialect recognition performance with derived features on male and female subjects is presented in Figure 3.9. An interesting observation is possible that, a slight reduction in performance is seen in the case of female speakers when 39 MFCCs are used over 13 MFCCs features. However, a significant reduction is noticed, in the case of male speakers when SDC features are added.

B Dialect Identification using Excitation Source Features

Excitation source and vocal tract system features are said to be almost complementary in nature from both speech production and feature extraction points of view. It can be seen from Table 3.4 that the recognition accuracy obtained is about 84%, using both spectral and prosodic features. In order to study the dialectal significance of source information, LP residual samples chosen around epoch locations are used along with cepstral coefficients of LP residual signal. Results obtained, using source level information and its combination with spectral and prosodic features are given in Table 3.4. From the table it is clearly observed that, source level features alone have classified dialects with an accuracy of 39.06%. RMFCCs (cepstral coefficients extracted from residual signal) have shown slightly lesser accuracy than traditional speech MFCCs. Whereas, combination of RMFCCs and speech features has resulted in an accuracy of about 81.25% indicating the existence of non-overlapping information. Further, it is noticed from the results that Kannada dialect classification performance is slightly improved to 86.74% when spectral, prosodic and source level features are combined together. From these

Table 3.5: Dialect recognition performance using Gabor features with single SVM classifier on KDSC. Legend: SF-Spectral flux, SE-Spectral entropy

Sl.No.	Features	Accuracy in %
1	MFCCs	81.25
2	Gabor features	84.50
3	MFCCs+Gabor Features	87.50
4	MFCCs+SF+SE+Pitch+Energy SOE+SSOE+IF+Gabor Features	89.25

results it may be understood that, source, prosodic, and system level features are different in nature and carry complementary dialect-specific attributes.

C Dialect Identification using Spectro-Temporal (Gabor) Features

Though, source and system level features characterize dialects in a fairly good way, few dialectal forms show clear variations in the spectro-temporal aspects (Chittaragi et al., 2018a). These aspects can be effectively modeled through the use of biologically-inspired 2D Gabor features derived from a 2D Gabor filter bank (Meyer, Bernd T and Kollmeier, Birger, 2011; Lei et al., 2012). Several experiments are conducted to evaluate performance of dialect recognition with individual and combined feature vectors. The obtained results are given in Table 3.5. Gabor features have performed well over popular MFCC features and combination of both have shown a slight increase in the performance on Kannada dialects. However, addition of spectral, prosodic and excitation source features has further improved the accuracy up to 89.25%. Gabor features are said to model specific stimuli to which the neurons of the mammalian auditory cortex are sensitive (Schröder et al., 2015). Both spectral and temporal modulation frequencies do exist in these stimuli. Gabor filters, representing spectro-temporal modulations attempt in emulation of the human auditory system, through signal processing strategies (Meyer et al., 2011).

Table 3.6: Dialect recognition performance with spectral, prosodic, source, Gabor and combination of these features using single SVM classifier method on KDSC

Sl.No.	Features	SVM	RF	ERF	XGB	ESVM
1	MFCCs	81.25	68.12	70.93	66.25	**81.36**
2	MFCCs+SF+SE	83.02	79.58	81.06	78.55	**83.12**
3	Pitch+Energy	**47.75**	43.12	42.18	36.87	44.52
4	RMFCCs	**78.75**	60.93	68.12	65.63	75.31
5	SOE+SSOE+IF	39.06	39.68	**42.18**	32.83	38.02
6	Gabor Features	84.50	83.87	**88.25**	83.16	86.24
7	MFCCs+SF+SE+Pitch+Energy	84.41	75.00	75.63	77.5	**86.25**
8	4+5+6+7	**89.25**	88.45	88.51	84.26	**92.50**

SF-Spectral flux, SE-Spectral entropy, RMFCCs-Residual MFCCs, SOE-Strength of Epoch, SSOE-Slope of SOE, IF-Instantaneous Frequency, SVM-Support Vector Machine, RF-Random Forest, ERF-Extreme Random Forest, XGB-Extreme Gradient Boosting, ESVM-Ensemble SVM

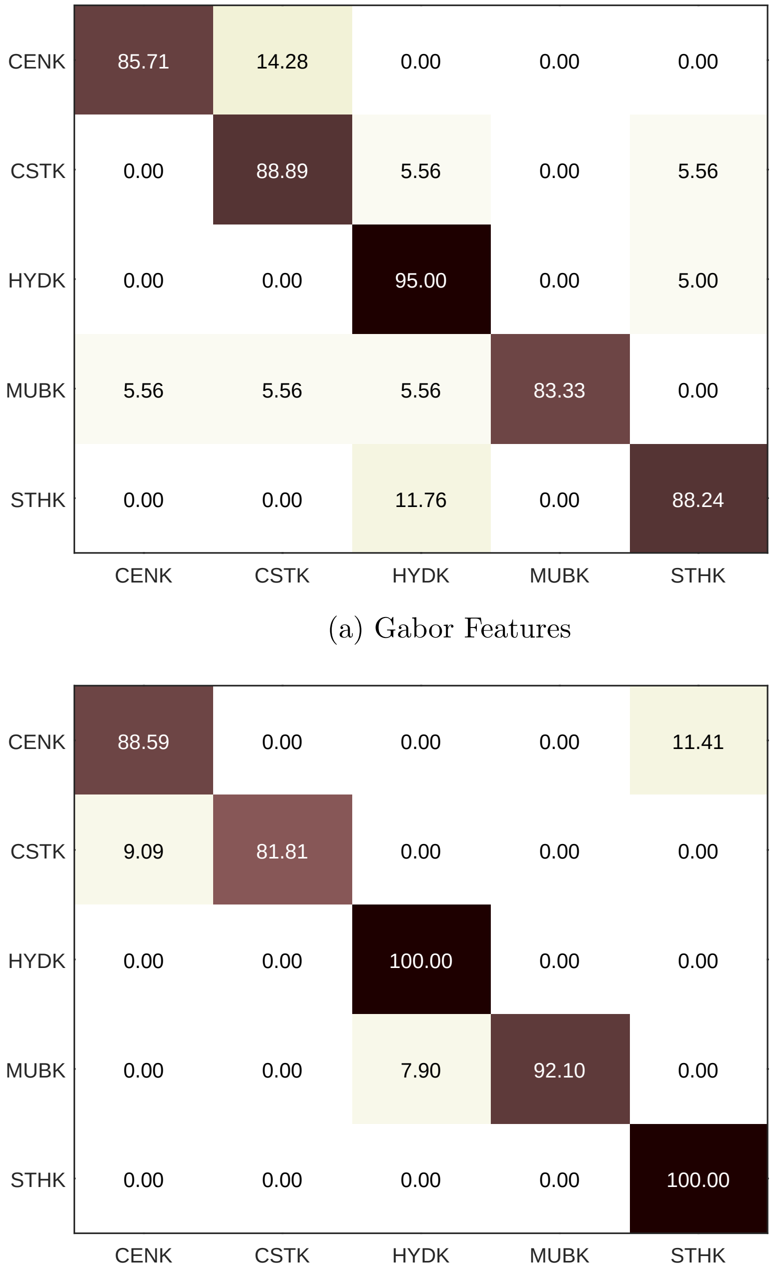
D Dialect Identification using Combination of Features with Single and Ensemble Classifications Algorithms

It is proposed in the literature that, an ensemble of various classifiers works better over traditional single classifier based algorithms (Dietterich, 2000a). Several experiments are conducted with single SVM and ensemble methods using various extracted speech features.

Table 3.6 presents, dialect recognition results, that have obtained using single SVM and ensemble algorithms with 13 MFCCs and other features without addition of SDCs. It is noticed from the results that, with majority of the features, SVM based ensemble classifiers has resulted in better performance over decision tree based algorithms. SVM and ESVM have performed comparatively better over other classifiers. It is observed from the table that, the ADI system with highest accuracy of 89.25%, is obtained with a combination of all feature set and with SVM classifier. Among decision tree based methods and ESVM, ERF classifier has shown better performance of 88.25% with Gabor features when compared to 84.5% using SVM classifier. Both prosodic and source features have demonstrated poor performances with all classification algorithms attempted in this work. The confusion matrices obtained with the Gabor features (row no. 6 of Table 3.6) and combination of four different features (row no. 8 of Table 3.6) are given in Figure. 3.10. It is observed that, the HYDK dialect is more accurately classified with Gabor features with 95% of accuracy and combination has resulted with 100% of classification performance. It is worth noting from Figure 3.10b that, both HYDK and STHK dialects are more accurately classified with a combination of four features that have been considered in the present work. Indeed, slight misclassification are also observed with dialects spoken in the neighboring regions (CENK & CSTK, HYDK & MUBK, STHK & CENK). Comparison of dialect identification performance, obtained using excitation source, spectral, prosodic, and spectro-temporal features, is presented in the form of bar graph in Figure 3.11.

3.4.5 Performance Comparison of Dialect Identification Systems with IViE Speech Corpus

In this subsection the above proposed approaches are evaluated using the standard IViE dialect speech corpus. An average dialect recognition performance obtained using derived spectral and prosodic features is presented in Table 3.7. Overall results obtained with both single SVM and ESVM classifiers are comparatively better. Meanwhile, it can be noted that spectral features have shown 87.85% recognition rate with ESVM classifier. Also, prosodic and source features have shown lower performance in the case of IViE dataset when compared to Kannada dataset. IViE dataset includes nine dialects which generally are clearly distinguishable. Nine dialects represent different parts of British Isles. Gabor features and combination of other features (refer 8 of Table 3.6) have resulted in better accuracies of about 96.74% and 99.21% with ESVM classifiers. Increased recognition rate may be attributed to the fact that, IViE is a clean, studio recorded semi-spontaneous corpus with least speaker variabilities. Majority of the nine dialects of IViE dataset are clearly distinguishable by statistical properties. Whereas, the five Kannada dialects presented, are very closely spaced overlapping statistical properties with regard to different features. Hence, there is a lot of similarity among five dialects. Moreover, proposed KDSC is with larger speaker variabilities



	CENK	CSTK	HYDK	MUBK	STHK
CENK	85.71	14.28	0.00	0.00	0.00
CSTK	0.00	88.89	5.56	0.00	5.56
HYDK	0.00	0.00	95.00	0.00	5.00
MUBK	5.56	5.56	5.56	83.33	0.00
STHK	0.00	0.00	11.76	0.00	88.24

(a) Gabor Features

	CENK	CSTK	HYDK	MUBK	STHK
CENK	88.59	0.00	0.00	0.00	11.41
CSTK	9.09	81.81	0.00	0.00	0.00
HYDK	0.00	0.00	100.00	0.00	0.00
MUBK	0.00	0.00	7.90	92.10	0.00
STHK	0.00	0.00	0.00	0.00	100.00

(b) Source+Spectral+Prosodic+Gabor

Figure 3.10: Confusion matrices of dialect recognition with Kannada dataset. Legend: CENK:Central Kannada, CSTK:Coastal (Karavali) Kannada, HYDK:Hyderabad Kannada, MUBK: Mumbai Kannada, STHK: Southern Kannada (a) Using Gabor features, (b) Using Source, spectral, prosodic and Gabor features

and is recorded in an open environment.

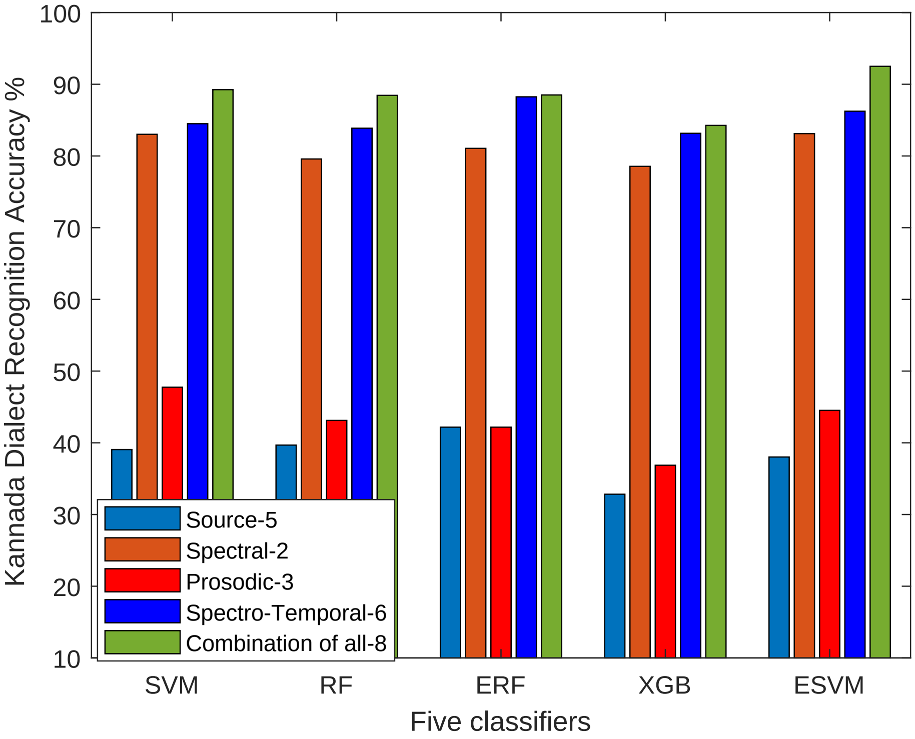


Figure 3.11: Comparison of dialect recognition performance using Source, Spectral, Prosodic, and Spectro-Temporal features on five classification methods. Numbers in brackets indicates the Sl.No. from Table 3.6

3.5 Summary

This chapter has presented an ADI system proposed by exploring four different speech features namely; (i) Excitation source, (ii) Spectral, (iii) Prosodic, and (iV) Spectro-temporal domain. Kannada dialect speech corpus (KDSC), is a newly proposed text independent and spontaneous speech corpus; which is collected exclusively from native speakers belonging to five prominent dialectal regions of Karnataka. In this chapter, an attempt has been made to characterize and classify five dialects of Kannada, by exploiting conventional speech features from segmental, supra-segmental and sub-segmental levels. Traditional MFCCs along with SDCs, spectral flux and entropy features are used as representatives of vocal tract system. Fundamental pitch and energy features are used to extract the prosodic attributes. Glottal closure instants, regions around GCIs, strength of epoch, slope of strength of epochs, and instantaneous frequencies are used as excitation source information attributes. Additionally, in order to capture spectro-temporal varia-

Table 3.7: Dialect recognition performance using SVM and ESVM on English IViE dataset

Sl.No.	Features	SVM	ESVM
1	MFCCs	84.48	86.72
2	Spectral Features	85.59	87.85
3	Prosodic Features	61.09	46.31
4	Source Features	34.89	48.85
5	Gabor Features	96.12	96.74
6	Combination of all Features	98.45	99.21

tions across dialects, 2D Gabor features are computed and used. Single classifier and multi classifier based (decision tree and SVM based) algorithms were used for developing dialect recognition systems. Performance comparison of both categories of classification algorithms have been carried out. The effectiveness of the proposed features and approach has been evaluated on KDSC and internationally known, standard, Intonation Variation in English (IViE) dataset containing nine British English dialects.

From the results, it is observed that, all four varieties of features, used in this work, carry significant information useful in classification of dialects. Prosodic and excitation source features are found to be less significant in the classification of Kannada dialects. The existence of non-overlapping dialectal traits among different features and reduction in performance with the addition of SDC features has been investigated on KDSC. Both spectral and Gabor features individually have demonstrated better performances over the prosodic and source features. Gabor features alone have successfully classified Kannada dialects with an accuracy of about 84%. This has demonstrated the usefulness of 2D Gabor features for dialect identification. These are well established features in the field of image processing. Comparatively better dialect recognition performance is observed with ensemble algorithms. Among decision tree and SVM based methods, ensemble algorithms with SVM as the base learners have shown better dialect recognition. The highest of 92.25% of recognition rate is observed with a combination of the evidence from all four features using ESVM method on Kannada dataset. Moreover, dialect recognition performance obtained from proposed features was found to be better over standard UBM based i-vector features for Kannada language. Standard IViE English corpus is used for validating the results obtained with KDSC. The proposed feature vectors evaluated on IViE have shown better performance (99.21%) with derived feature vectors using ESVM algorithm. Chapter 4 discusses the dialect identification system developed, by using non-conventional features known as chroma features (music related) and spectral shape based features.

CHAPTER 4

Classification of Dialects based on Non-Conventional (Dialect-Specific) Features

A detailed explanation of dialect identification systems developed with various conventional speech based features has been provided in the previous chapter. The present chapter includes the details of the use of non-conventional, i.e. dialect-specific features for dialect classification. The details of extraction of chroma based music related features from speech, for capturing dialectal cues are briefly discussed. In addition to chroma features, details of eight significant spectral shape-related features derived from short term spectra are provided. Details of the KDSC and IViE datasets prepared for verifying the effectiveness of the proposed features and methods are highlighted. The details of two different classification methods employed for dialect classification are also briefly covered. Analysis of dialect classification results, achieved using proposed non-conventional features with individual and in combination of features is provided.

4.1 Introduction

Dialects primarily get derivations from the phonological, lexical and grammatical variations in the usage of a language with very minor and subtle differences (Huang et al., 2007). The dialectal variations are also mainly due to specific speaking patterns followed among a group of speakers. Dialect recognition task has a lot of ambiguities since there are no linguistic rules to demarcate the dialectal boundaries. Dialectal boundaries are highly overlapped and confused since they are derived from the same language by sharing standard phoneme set and grammatical rules. Hence, it is better to develop an ADI system, by formulating the fundamental acoustic differences especially in terms of dialect-specific prosodic attributes that exist across dialects. Majority of systems reported in the literature are with the idea of correlating the prosodic evidence such as pitch, energy, and duration to the dialectal variations for developing ADI systems (Chittaragi et al., 2018b; Mehrabani and Hansen, 2015; Mary and Yegnanarayana, 2008; Ramus and Mehler, 1999).

However, ADI systems may also be developed using the non-conventional features deriving the complex prosodic features from the dialects for effective classification. In this thesis, chroma features familiar with music-related systems are used for identification of dialects. Chroma features try to aggregate spectral information and attempt to encapsulate the evidential variations, related to timbre, correlated melody, rhythmic, and intonation patterns found prominently among dialects of a few languages. As every dialect has shown variations in pitch and energy parameters, chroma features are expected to capture musical patterns in speech in terms of rhythm and intonation and in this work six chroma features are used instead of twelve chroma parameters. Unlike MFCC features, these features show robustness to variations in the timbre of the signal and noise (Müller and Ewert, 2011).

Capturing only pitch and energy features does not discriminate dialects effectively. In addition, eight significant spectral shape related features from short term spectra, are computed and combined with chroma features and named as chroma-spectral shape features. Eight spectral shape-based features are used to capture the complementary spectral information of dialectal cues. Instead of using raw features, post-processed derived features are used for the development of ADI systems. A single classifier based SVM and ensemble ESVM methods are used for classification of dialects. Performance of both methods are compared. Slightly better results are observed with proposed chroma-spectral shape features, when compared to the state-of-the-art i-vector features.

Rest of the chapter is organized in the following sections: Section 4.2 covers brief details of the dataset employed in the present study. Feature extraction procedure for proposed chroma and spectral shape based features and details of the classifiers used are provided in Section 4.3 and Section 4.4 respectively. Section 4.5 discusses the experimental setup, and results. Section 4.6 provides a summary of the present work.

4.2 Dataset Details

In this study, the performance of a dialect identification system is evaluated by using two corpora namely: Kannada dialect dataset (Chittaragi et al., 2019) and IViE British English dialect dataset (Grabe and Post, 2002) of which details are provided in section 2.7. Before feature extraction, both datasets are pre-processed, to remove silenced frames from each audio files using dynamic energy threshold based algorithm (Giannakopoulos and Pikrakis, 2014). Each audio clip is divided into smaller audio segments of around 5 seconds and is used in the experiments.

4.3 Feature Extraction

The proposed work in this chapter aims to employ music related non-conventional chroma and spectral shape features to capture dialectal cues present in small sized audio files.

4.3.1 Chroma Features

Chroma features are commonly derived by encoding the normalized short-time energy distribution of the music signal across 12 bands (Wakefield, 1999). These 12 chroma bands stand for the twelve favorite pitch classes of Western music. Generally, these features have been built to process and classify music clips into different genre, chord identification, audio matching, and so on (Müller and Ewert, 2011; Antoni, 2006; Han et al., 2010).

It is commonly observed among the people, belonging to different dialectal regions that they intuitively use different proportions of frequency bands belonging to several pitch classes in their regular speech (Mehrabani and Hansen, 2015; Mary and Yegnanarayana, 2008). Therefore, it is viable to distinguish speech signals of different dialects by observing pitch based features. In this work, an attempt has been made to use music related chroma features for identification of five Kannada dialects. Due to employment of the frequency bands and pitch profile details, the chroma based approach is quite practical for dialect classification as well. It is observed that in the case of some languages including Kannada, that energy, pitch, and loudness play a prominent role across different dialectal groups (Soorajkumar et al., 2017; Nagesha and Nagabhushana, 2007; Mehrabani and Hansen, 2015).

Chroma features are extracted cyclically to represent the frequency. According to literature, two dimensions are necessary to represent the pitch of a human voice. Helix structure is used in this work for better representation (Wakefield, 1999). Tone height and chroma are the two components used to characterize the vertical and angular dimensions, means; the perceived pitch p is factored into values of chroma c and tone height h as shown in equation 4.1.

$$p = 2^{h+c} \tag{4.1}$$

According to the equation, linear changes in c result in logarithmic changes in the associated fundamental frequency. Normal chroma values in an interval between 0 and 1 are divided into 12 equal parts, to represent 12 corresponding pitch values of the equal-tempered chromatic scale. From the equation, one can observe that it implies the distance between two pitch values depending on both c and h values. Finally, a complete chroma feature vector is represented by a 12-dimensional vector. In equation 4.1, values of chroma are, $c \in [0, 1]$ and they represent angle along the spiral. Tone hight $h \in Z$, is a set of integers; indicates the position along the common spoke (Wakefield, 1999). Calculation of a chroma features from a given frequency value is done using equation 4.2

$$c = log_2 f - \lfloor log_2 f \rfloor \tag{4.2}$$

Thus, chroma is the fractional part of the base-2 logarithm of the corresponding frequency value. In this way, chroma features model the pitch class profiles, by aggregation of several pitch values to a single class. In this work, an intuition believed where dialect-specific cues which are mainly influenced by the pitch and energy variations can be adequately captured through chroma features. However, chroma features in the case of dialectal discourse (speech) may not be as distinctive as those observed in the case of music (Giannakopoulos and Pikrakis, 2014).

During chroma information extraction, from a given dialectal speech into a sequence of chroma features are extracted, each value represents the aggregate of the energy of the signal spread over six chroma bands. Six bands or bins are chosen, unlike 12 in music, after empirical analysis through the pre-test experiments. It has been observed that only six different frequency bins are sufficient and efficient to capture perceptually the prominent harmonics that occur in natural speech. The procedure followed during chroma feature extraction is summarized in the steps given below.

1. The chroma vector is computed by grouping the FFT coefficients of a short-term window into six bins, where each bin is a representation of six equally tempered pitch classes.

2. Each of these bins brings about the mean of log-magnitude of the respective FFT coefficients.

$$v_k = \sum_{n \in S_k} \frac{X_i(n)}{N_k} \quad k \in 0, 1, ...5 \tag{4.3}$$

Where S_k is a frequency subset corresponding to the FFT coefficients. N_K is the cardinality of S_k and $X_i(n)$ is the magnitude of the FFT coefficients of the i^{th} audio frame.

3. Chroma vector v_k is computed based on short-term analysis of speech signal with a frame size of 20 ms and an overlap of 10 ms. Matrix V, with elements $V_{k,i}$, where k represents the pitch class, and i represents the frame number contain the resultant chroma features.

The obtained matrix V is also known as a chromagram and is similar to the concept of the spectrogram. Chroma features preserve the intensity associated with each of the six bins (semitones) within one octave, where all octaves are folded together (Giannakopoulos and Pikrakis, 2014; Ellis and Poliner, 2007). The

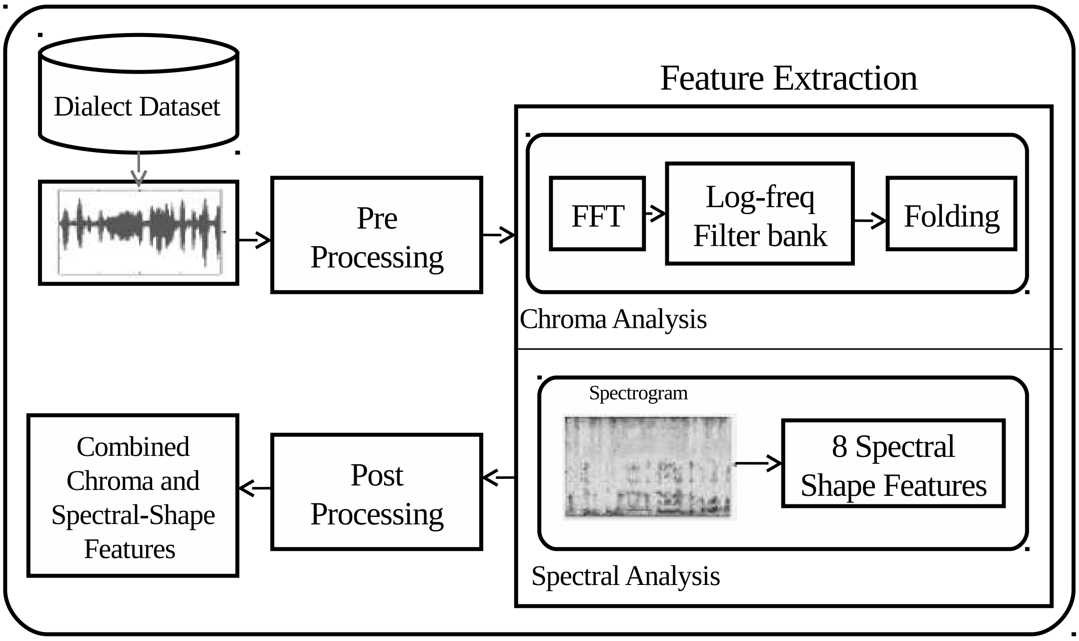


Figure 4.1: Steps in chroma, spectral shape feature extraction

fundamental frequency that is used for identifying the pitch and its corresponding bin has been given a range from 55 to 200 Hz, covering only the human range of frequencies and it is set to 120 after several pre-test experiments.

4.3.2 Spectral Shape Features

Discriminating phonetic structures of different phonemes across Kannada dialects motivate for spectral analysis of a signal. Accents or dialects of any language represent the varying pronunciation styles. These styles may follow a unique pattern of rhythm, intensity, intonation, and stress. Speaking rate is also one of the important discriminators for dialectal speaking patterns (Ma et al., 2006; Rouas, 2007). In this work, spectral shape features are extracted to capture the statistical description of spectra, through eight prominent features. Generally, spectral shape-based features are believed to represent paralinguistic information used for other speech-related tasks such as language identification, speaker and emotion recognition (Koolagudi and Rao, 2012). Eight features namely: spectral centroid, spread, flatness, flux, slope, roll-off, skewness, and kurtosis are extracted from overlapped hamming windows of size 20 ms with 10 ms steps. Overview of the feature extraction process is presented in Figure 4.1.

A Spectral Flux:

Timbre is the speaker specific feature of sound unit that helps to compare similarity between two speech utterances. Spectral flux is one of the known correlates of timbre information. Spectral flux usually corresponds to a perceptual roughness of sound. In this work, flux feature is computed and used to measure the spectral changes between two successive frames. It is computed by subtracting the power spectra of a frame against the same of the previous one (Giannakopoulos and Pikrakis, 2014). Square of the values is taken to avoid negative sign. The following equation is used,

$$Fl_{(i,i-1)} = \sum_{k=1}^{Wf_L} (EN_i(k)) - (EN_{i-1}(k))^2 \tag{4.4}$$

Where $EN_i(k) = \frac{X_i(k)}{\sum_{l=1}^{Wf_L} X_i(l)}$, here $EN_i(k)$ is the k^{th} normalized DFT coefficient at the i^{th} frame, Wf_L is the frame size.

B Spectral Entropy:

Spectral entropy of a signal measures the distribution of spectral power. It is computed using Shannon's entropy concept. Spectral entropy also captures abrupt changes occurring in the energy level of an audio signal (Giannakopoulos and Pikrakis, 2014).

While computing spectral entropy of a frame, corresponding spectrum is divided into L sub-bands (bins). The energy E_f of the f^{th} sub-band, for f = 0, . . ., L-1 is calculated using equation (4.5). Then, energies of all bins are normalized by dividing with the spectral energy of the whole frame, i.e., $ef = \frac{E_f}{\sum_{f=0}^{L-1} E_f}$, the entropy of each normalized energy value is calculated using the equation (4.6)

$$E(i) = \frac{1}{Wf_L} \sum_{k=1}^{w_L} |x_i(k)|^2 \tag{4.5}$$

$$H = -\sum_{f=0}^{L-1} ef.\log_2(ef) \tag{4.6}$$

In this work, the value of L is set to 10 indicating that each frame is divided into 10 bins.

C Spectral Centroid and Spread:

The spectral centroid and spread represent the spectral position and shape. Centroid measures the brightness of a spectrum and represents the spectral center of gravity. The spectral centroid, C_i, of the i^th audio frame is computed as follows:

$$C_i = \frac{\sum_{l=1}^{Wf_L} lX_i(l)}{\sum_{l=1}^{Wf_L} X_i(l)} \tag{4.7}$$

Where, Wf_L is the number of DFT coefficients used, X_i is the magnitude of the DFT coefficients of the i_{th} audio frame.

Similarly, spectral spread feature represents second central moment of the spectrum. It computes the fullness of the given spectrum. Spread is computed by taking the deviation of the spectrum from the spectral centroid using the following equation:

$$S_i = \sqrt{\frac{\sum_{k=1}^{Wf_L}(k - C_i)^2 X_i(k)}{\sum_{k=1}^{Wf_L} X_i(k)}} \tag{4.8}$$

On the other hand, spectral spread measures how the spectrum is distributed around its centroid. Obviously, low values of the spectral spread correspond the signal whose spectrum is tightly concentrated around the spectral centroid. Here, the centroid feature corresponds to the brighter sounds of the spectrum and spread measures how spectrum is distributed around the centroid.

D Spectral Flatness:

Spectral flatness is computed as the ratio of geometric and arithmetic mean of a power spectrum. It represents the peakiness in the vector. Instead of measuring flatness across the whole band, it is computed within each sub band.

$$Flatness = \frac{\exp\left(\frac{1}{W_{fL}}\sum_{k=1}^{W_{fL}} X_i(k)\right)}{\frac{1}{W_{fL}}\sum_{k=1}^{W_{fL}} X_i(k)} \tag{4.9}$$

E Spectral Slope:

Slope represents information on how fast energy of formants gets decreased or increased. This is computed by taking a linear regression of the amplitudes of formant values from a spectrum. It measures the voice quality aspects such as harshness, whisper, creakiness, breathiness etc. from a voice.

F Spectral Roll-off:

This feature is treated as a spectral shape descriptor of an audio signal and it can be basically used to discriminate voiced and unvoiced sounds. This feature is defined as the frequency below a certain threshold value (generally 90%-95%) of the magnitude spectrum. The m^{th} DFT coefficient corresponding to the spectral rolloff of the i^{th} frame is shown in the equation 4.10.

$$\sum_{k=1}^{m} X_i(k) = C \sum_{k=1}^{W_{fL}} X_i(k) \tag{4.10}$$

Where, C is the percentage parameter set to 95 in this paper.

G Spectral Skewness:

Skewness indicates whether the spectrum is skewed towards a particular range of values. It represents the tilt of the spectrum and spectral tilt is defined as a measure of distribution of spectral energies in a speech signal. It indicates how much the shape of the spectrum is below the center of gravity and how different it is from the shape above the mean frequency.

H Spectral Kurtosis:

This feature is the fourth central moment of the spectrum and represents the pointedness of a given spectrum. If the kurtosis is high, the spectral peaks are usually well defined. If the kurtosis is low, then the distribution of the spectrum is relatively flat (Antoni, 2006).

These features capture the dialectal information, existing in the frame level average spectrum which is computed from the given audio signal. Spectral shape features are nothing but statistically computed parameters like mean, standard deviation, etc. from the frame level features mentioned above.

From these discussions, it may be noted that chroma features have been extracted to capture features from frequency bands and pitch profiles. It is also necessary to consider the variations occurring in the vocal tract through spectral analysis. Hence, both complementary dialectal cues, are combined to form a new proposed feature vector called chroma+spectral shape features. Both features are combined to get a feature vector of size 14 (6+8), includes 6 chroma and 8 spectral shape features.

The frame-wise feature vectors are further post-processed to get derived features. The different length feature vectors obtained from varying length speech clips are converted into the fixed length ones. Mean and standard deviation (STD) is computed for each pair of consecutive frames, then mean of all these mean vectors is calculated and are named as derived features. Additionally, the performance of these features is compared with the i-vector features extracted using the GMM-UBM approach. MFCCs are very familiar spectral features, and they are said to imitate the human auditory system. Hence in this work, MFCC features are also extracted, as they convey complementary dialect-specific spectral information and are evaluated individually and in combination with chroma and spectral shape features. Thus, 13 MFCC features are converted into 26 derived features (13 mean and 13 standard deviation). Similarly, 39 MFCC-SDC set yields 78, and 14 chroma-spectral shape features give 28 features. Further details of post-processing can be found in (Chittaragi et al., 2018b).

4.4 Classification Models

In the present work, a single classifier based SVM and multi-classifier based ensemble SVM algorithms are employed for the dialect classification. A single classifier based SVM method tries to capture the discriminating parameters across the feature vectors for identification of classes. A few SVM based dialect recognition systems are reported in the literature (Chittaragi et al., 2018b; Biadsy et al., 2011; Pedersen and Diederich, 2007). Recently, the ensemble of multiple classifiers is gaining much attention due to its promised improved performance in different speech processing tasks. These are proven to be the state of the art classification approaches since they are designed to combine the prediction outcomes of several individual classifiers (Dietterich, 2000a).

In this study, bagging (bootstrap with aggregation) technique is followed with SVM classifier instead of decision trees as the base model. Optimal values are empirically chosen for essential hyper-parameters, to yield better performance. Number of estimators is set to 2048, where, 2048 subproblems are built from an available feature set, using random selection and replacement technique. Also, RBF kernel is chosen since it has shown better performance over the linear kernel, in these experimentation (Pedregosa et al., 2011). Detailed information regarding these two classification methods is given in section 3.3.

4.5 Experimental Results and Discussions

This section presents the performance evaluation of the traditional MFCC-SDC, i-vector, proposed chroma, spectral shape and chroma-spectral shape features for dialect identification. The results achieved with both Kannada and IViE dialect dataset are summarized. Performance of the proposed features is also evaluated in a noisy environment, for testing their robustness.

Several experiments are conducted to know the significances of different features for Kannada dialects recognition. The i-vectors, which are used in state-of-the-art dialectal systems are also extracted and used for comparison with the proposed ones. These i-vectors have been quite successful in speaker and language identification systems (Dehak et al., 2011b). However, few researchers have enen applied them for dialect identification task (Hansen and Liu, 2016; Zhang and Hansen, 2018).

Results obtained using GMM-UBM, and i-vector features are given in Table 3.2. It is observed from the table, that i-vector features with UBM have shown better performance over traditional GMM-UBM based systems.

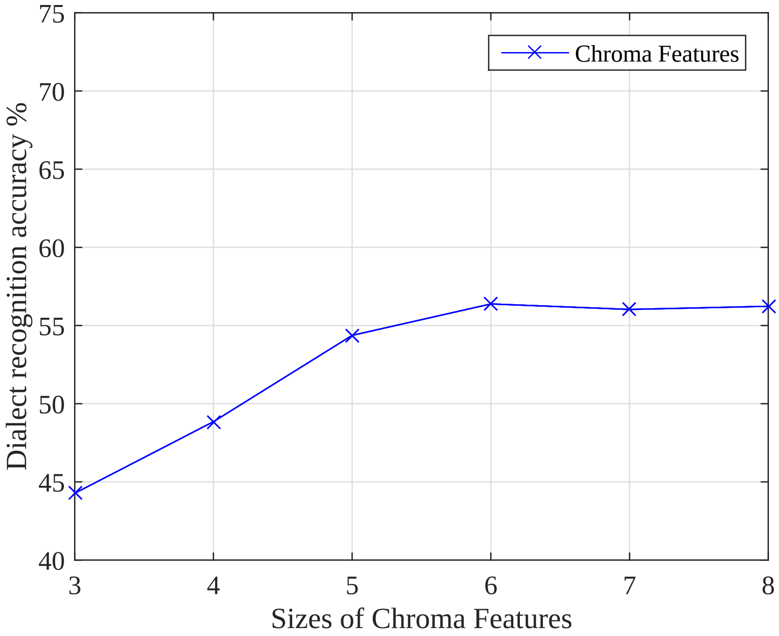


Figure 4.2: Dialect recognition accuracy with different sized chroma features

4.5.1 Proposed Chroma and Spectral shape based Features for Kannada Dialect Identification

Various experiments are conducted on dialect identification with SVM and ESVM classifiers using proposed feature sets. The chroma-spectral shape feature set is evaluated individually and in combination with MFCCs, to understand dialectal evidence from both features using a late equal fusion mechanism (combining features extracted from two different methods with equal score). Kannada dialect dataset is divided into training and testing data. For consistent, and unbiased model evaluation and to avoid over-fitting of the model, five-fold cross-validation is employed. The results obtained are summarized in Table 4.1. Proposed chroma+spectral shape features have produced slightly better results over MFCC features. However, chroma features alone have exhibited lesser recognition rate of 56.65% with SVM. Experiments are conducted repeatedly, to choose the optimum number of chroma features which needs to be selected for these experiments. Finally six chroma features are used. Later, throughout this work, six chroma features are used. Also, Figure 4.2 shows the performance accuracies obtained with varying numbers of chroma features.

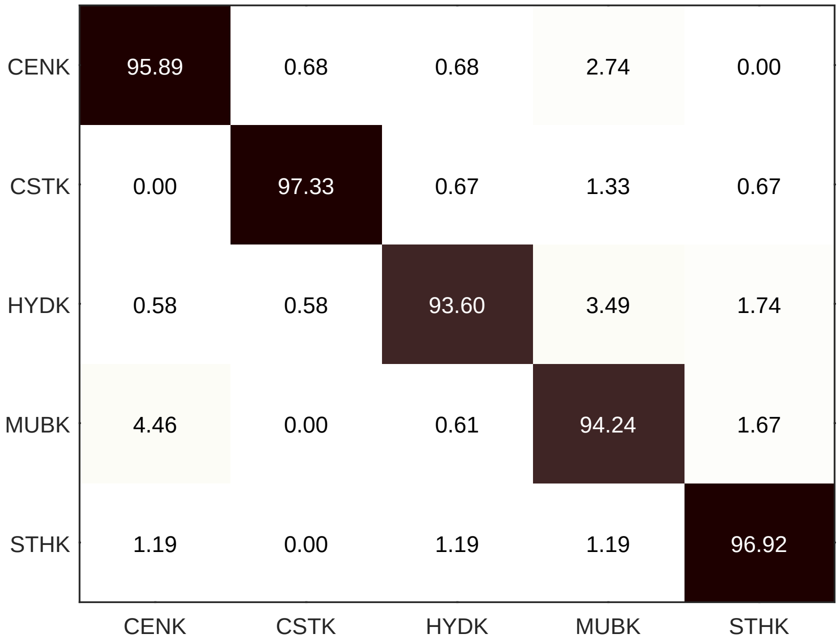
Table 4.1: Dialect recognition performance using SVM and ESVM on Kannada dialect dataset. Legend: CENK-Central Kannada, CSTK-Coastal Kannada, HYDK- Hyderabad Kannada, MUBK-Mumbai Kannada, and STHK-Southern Kannada

Features	SVM Method (Accuracy in %)						ESVM Method (Accuracy in %)					
	CENK	CSTK	HYDK	MUBK	STHK	AVG	CENK	CSTK	HYDK	MUBK	STHK	**AVG**
MFCC	91.58	86.14	85.79	84.05	90.81	**87.67**	93.05	88.14	83.24	89.41	85.60	**87.89**
Chroma Features	50.85	68.35	51.53	49.30	51.25	**56.65**	59.65	54.66	48.39	55.00	51.41	**53.82**
Spectral Shape Features	75.75	82.58	80.77	80.84	77.12	**79.00**	88.36	80.67	80.23	74.39	75.76	**79.88**
Chroma + Spectral Shape Features	92.21	91.37	84.13	93.58	88.50	**89.75**	93.75	89.69	85.45	92.40	86.78	**89.60**
MFCC + Chroma + Spectral Shape Features	93.05	92.40	96.73	95.16	94.18	**94.30**	95.89	97.33	93.60	94.24	96.92	**95.60**

Spectral shape features alone have also shown lower recognition performance compared to standard MFCC features on Kannada dialects. Lower performance with individual features may be due to chroma features; carry mainly melodic and harmonic features extracted through pitch and energy profile properties. Spectral shape features are mainly said to capture the statistical parameters from the spectra. Indeed, combining both features can exploit, both spectral and energy profile related to dialects. In this work, both features are combined and used. However, both SVM and ESVM, have exhibited more or less similar results with chroma, spectral shape based and features, MFCC features including their different combinations. To capture, complementary evidence from MFCCs, chroma, and spectral shape features, all three features are combined. This combination has resulted in considerable improvement in dialect recognition. Highest classification performance of 94.30% and 95.60% are achieved with SVM and ESVM classifiers respectively. Ensemble SVM algorithm has performed slightly better over single SVM.

In specific, it can also be noticed that classification results are comparatively better on CENT (92.21%) and MUBK (93.58%) dialects when used with proposed chroma-spectral shape features using SVM. Whereas, using proposed features, the misclassification rate is slightly higher in the cases of HYDK and STHK dialects. These two dialects are observed with a unique style and similar speaking patterns among themselves. Similar trends are noticed when the same experiments are repeated with ESVM classifiers. Improvement in performance is noticed with a combined feature set in the cases of both SVM and ESVM classifiers.

Figure 4.3 presents the colored confusion matrix for Kannada dialect identification using ESVM. The confusion matrix shows highest accuracy of 97.33% for coastal Kannada. Kannada spoken in the coastal region commonly follows a classical bookish speaking pattern. This includes the use of unique pitch and energy specific modulation that are distinguished precisely by the combined feature vector. Also, it is visible that CENK and MUBK are mutually confused. This may be due to the use of similar pitch and intonation patterns between them. How-



	CENK	CSTK	HYDK	MUBK	STHK
CENK	95.89	0.68	0.68	2.74	0.00
CSTK	0.00	97.33	0.67	1.33	0.67
HYDK	0.58	0.58	93.60	3.49	1.74
MUBK	4.46	0.00	0.61	94.24	1.67
STHK	1.19	0.00	1.19	1.19	96.92

Figure 4.3: Confusion matrix of dialect identification results obtained with ESVM using MFCC+Chroma +Spectral Shape features for Kannada dialects

ever, HYDK dialect is also misclassified with MUBK dialect. These two dialects are spoken in neighboring regions of Karnataka and are similar in terms of higher pitch range. STHK Kannada dialect is also classified with better accuracy, using the combination of complementary evidence. Comparison of dialect identification performance that is obtained using MFCCs, chroma, spectral-shape, and a combination of all these three features, is presented in the form of a bar graph for better visualization in Figure 4.4.

4.5.2 Proposed Chroma and Spectral Shape based Features for Identification of English Dialects

The set of experiments mentioned in subsection 4.5.1 are repeated on standard internationally known IViE corpus.

Table 4.2 shows the average performances achieved using MFCCs, chroma, spectral shape, and chroma-spectral shape feature vectors. It is noticed from the results that, consistent improvement in performance is observed with English dialects using proposed features. Further, ensemble-based SVM has also shown better performance over single SVM classifier with both datasets. The highest of

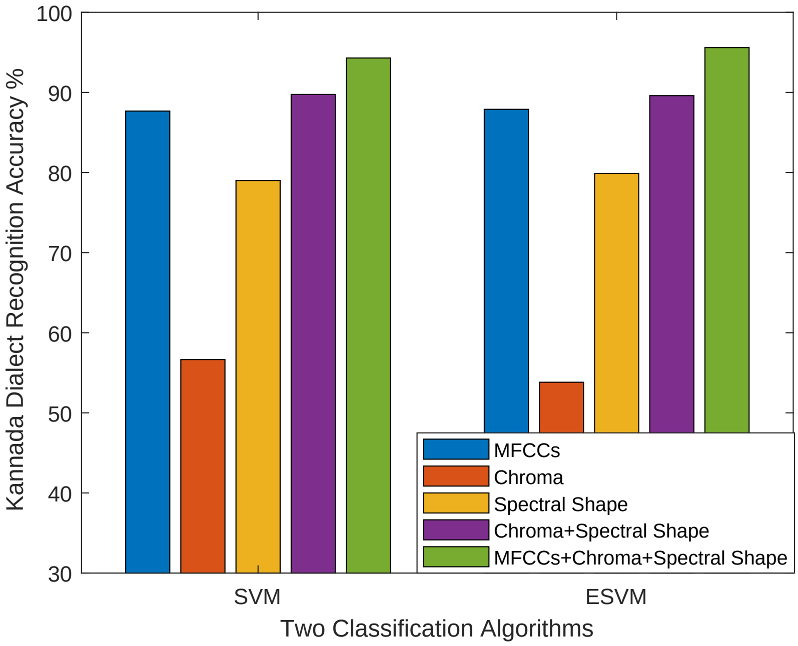


Figure 4.4: Comparison of dialect recognition performance of MFCCs, Chroma, Spectral-Shape features, and their combination using SVM and ESVM classifiers (Dataset- KDSC)

97.52 % of ADI performance is obtained with ESVM. In general, slightly better recognition performance are observed with IViE dataset over Kannada dataset. The improvement in the performance is due to the fact that IViE dataset is studio recorded read dataset. Also, it is text-dependent, with limited speaker variabilities. The Kannada dataset used in this study is text-independent speech data, with more significant speaker variabilities.

Table 4.2: Dialect recognition performance using SVM and ESVM algorithms on English IViE dataset

Features	**SVM**	**ESVM**
MFCCs	84.48	89.96
Chroma Features	82.33	83.79
Spectral Shape Features	83.57	86.30
Chroma+ Spectral Shape Features	92.95	94.78
MFCC + Chroma + Spectral Shape Features	96.66	97.52

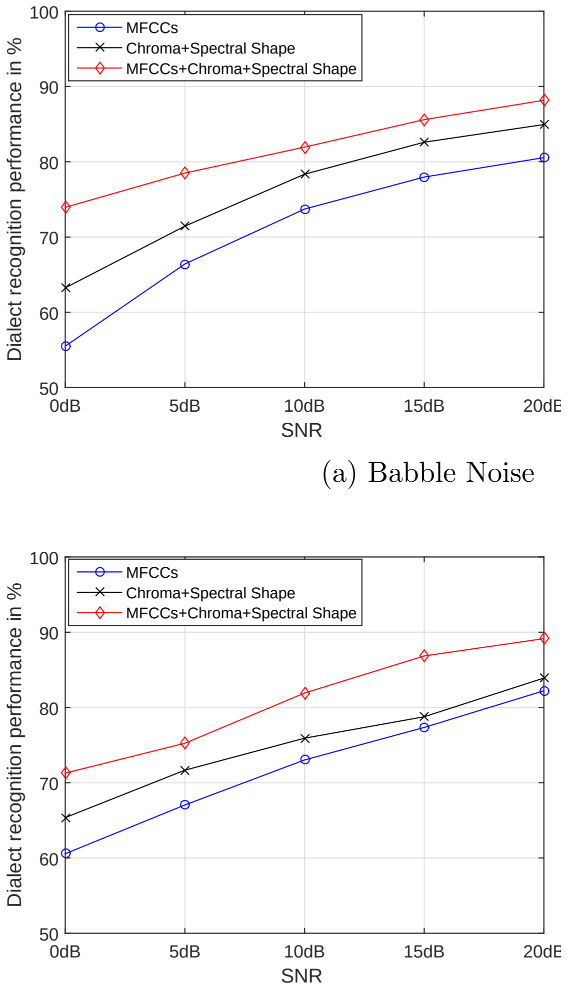
Table 4.3: Dialect recognition performance using SVM on noisy Kannada dataset with babble and spectral shape noise

Features	Babble Noise					Spectral Shape Noise				
	0 dB	5 dB	10 dB	15 dB	20 dB	0 dB	5 dB	10 dB	15 dB	20 dB
MFCCs	55.50	66.42	73.72	77.97	80.55	60.57	67.05	73.03	77.35	82.22
Chroma Features	34.02	41.67	47.00	50.32	51.35	45.85	48.32	48.77	51.97	52.80
Spectral Shape Features	45.05	52.10	59.57	63.90	67.25	60.55	61.02	62.23	63.60	64.23
Chroma + Spectral Shape Features	63.25	71.45	78.35	82.6	84.95	65.38	71.67	75.82	78.77	83.91
MFCCs + Chroma + Spectral Shape Features	73.95	78.50	81.95	85.6	88.17	71.28	75.25	81.92	86.85	89.15

4.5.3 Performance Analysis of Proposed Features in Noisy Conditions

In many realistic situations, it is required to use the ADI system under noisy conditions. However, in the literature, some systems have reported the development of ADI systems from noisy and limited speech environment (Liu and Hansen, 2011; Huang et al., 2007). The proposed chroma-spectral shape features have already demonstrated better performance in the case of clean speech. In the present study, dialect identification performance is evaluated on an audio, added with artificial additive noise.

Noisy speech is generated by adding a multi-speaker babble or Speech-Shaped Noise (SSN). Babble noise is the most challenging noise. The conversation of multiple speakers in an office environment is the kind of babble noise. This noise is uniquely challenging because of its high time evolving structure and its similarity to the desired target speech.



(a) Babble Noise

(b) Spectral Shape Noise

Figure 4.5: Kannada dialect recognition results in noisy conditions using proposed chroma and spectral shape based features

SSN noise employs long-term average spectrum and is similar to that of speech. The robustness of the ADI system with proposed features is evaluated in five distinct SNR levels such as 0, 5, 10, 15, and 20dB. For this experiment, Kannada dataset is used and results obtained are presented in Table 4.3. For easier analysis, pictorial representation of results is presented in Figure. 4.5. It shows that, the proposed features performed relatively well over MFCC features in both noisy conditions with different SNR levels. In the cases of 15dB and 20dB SNR of SSN noise very similar results are reported with MFCC and proposed features. Similarly, in the case of babble noise at 10dB, 15dB and 20dB SNR consistent improvements are observed with combined features compared to their individual use. Moreover, on an average proposed features have shown comparatively better performance during dialect recognition in severe noise conditions like 0dB and 5dB.

4.6 Summary

This chapter has demonstrated the applicability of non-conventional chroma and spectral shape features for dialect identification. In this study, a number of dialect classification experiments are conducted, using audio clips of about 5 seconds or lesser duration, for five Kannada dialects. For this purpose, chroma-based features, closely correlated with an aspect of harmony, are used to distinguish dialects. An attempt has been made, with the intuition of the existence of dialect-specific musical properties present in a speech by using chroma features. Even though these are well-established features in music data analysis and processing, no similar systems have been made available for capturing rhythmic patterns in dialects. This intuition has been found to be reasonable when evaluated on Kannada dialect identification and has resulted in better performance over standard MFCCs.

However, in this work complementary dialect specific information existing at short-term spectra are explored as spectral shape features. These spectral shape features are combined with modified chroma features to build a new feature vector of size 14 (6 chroma+8 spectral shapes). The Kannada and IViE dialect corpora are used for evaluating the performance. In the present study, three different types of classification methods namely GMM-UBM, SVM and ensemble ESVM are employed. Among all, SVM and ESVM have performed well and have given almost similar results, with an observation that the ensemble method has shown slightly better recognition performance, with an additional overhead of higher computational time, to handle larger subproblems (Marc et al., 2014). The idea of comparing the obtained results, with the other existing studies is not fair, as there are many non-matching issues to be considered such as language, type of speech corpora, type of features, changes in methodologies and many more.

Table 4.4: Dialect recognition performance on Kannada dialect dataset using various features considered in this work

Sl.No.	Features and Methods	Accuracy in %
1	MFCCs-GMM-UBM (26)	75.57
2	MFCCs-UBM-i-vector (26,300)	81.42
3	MFCCs (26,SVM)	87.67
4	MFCCs-SDCs (39,SVM)	73.10
5	MFCCs-SDCs (78,SVM)	83.38
6	MFCCs (13,SVM)	84.17
7	Chroma-Spectral Shape Features (28,SVM)	89.75
8	MFCCs + Chroma + Spectral Shape Features (54,SVM)	94.30
9	MFCCs + Chroma + Spectral Shape Features (54,ESVM)	95.60

The average results observed in these studies, with different features on Kannada dialects are summarized in Table 4.4. Proposed features have shown sustained improvement in performance over MFCC features. Addition of SDC features has demonstrated a reduction in performance. When compared with the GMM-UBM and UBM-i-vectors based systems on Kannada dialects have been classified more accurately with SVM and ESVM algorithms. Among all classifiers, an ensemble of SVMs (ESVM) has relatively shown better classification performance of 95.60%. Notably, proposed features have demonstrated improved and sustained dialect recognition performance even in a noisy environment.

The proposed features and approaches are found to be language, speaker and text-independent. For dialect recognition, a small speech segment of 5 ms and the approach is also proved to be noise robust. It is noticed that the proposed ADI systems using dialect-specific features may be better suitable for many commercial forensic application frameworks. Chapter 5 discusses about the development of ADI systems by extracting non-conventional features from the word and sentence level utterances.

CHAPTER 5

Dialect Classification from shorter speech units: Words and Sentences

Previous chapter has discussed about the development of dialect recognition systems using non-conventional, dialect-specific features representing musical aspects of speech from longer utterances. In this chapter, the details of use of non-conventional (dialect-specific) features from word and sentence level utterances for the task of dialect identification are provided. Extraction of dialect-specific dynamic and static prosodic parameters explored from pitch and energy features from the word and sentence level utterances is discussed. Inference of the dynamic and static variations in pitch and energy features across dialects is provided. Procedure followed for preparation of word and sentence level datasets and details of classification methods used are also provided. Experimental details with SVM and XGB ensemble methods under individual and combinations of features are covered. Analysis of results obtained from shorter speech utterances like word and sentences using the proposed static and dynamic prosodic features is provided at the end of the chapter.

5.1 Introduction

Over the last few years, the speech research community is trying to improve the performance of automatic speech recognition (ASR) systems by being automatic to the speech variabilities existing in natural speech. Due to this, dialect identification from speech has become increasingly interesting and essential. Generally, people belonging to a specific region follow a unique speaking style that can be easily distinguished from the utterances of word and sentences. Every dialect of a language generally shows unique distinguishing characteristics such as speech rate, stress levels, intonation, rhythmic patterns, varying patterns of usage of words at the beginning and end of sentences and so on. Indeed, all these factors assist in identifying the unique speaking patterns followed in each dialect (Rouas, 2007). Also, many times, in several instances, it may be required to recognize dialects from the shorter utterances such as words and sentences. In the literature, it is demonstrated that an accurate dialect recognition can be performed due to availability of extensive dialect specific properties at these shorter units of speech (Chittaragi and Koolagudi, 2017; Xu et al., 2016). It is also suggested in the literature that, even word and sentence level audio units are sufficient to make decisions about the languages and dialects (Xu et al., 2016). It is observed from our experiments that shorter utterances carry majority of significant dialect information. Further, word and sentence-level utterance processing for dialect identification ensures to reduction in the need of computational resources and are beneficial under limited data availability constraints (Liu and Hansen, 2011; Huang et al., 2007).

In the literature, research works carried out on words or sentence level units are significantly less exported. This may be mainly due to the fact that segmentation of the exact word or sentence level utterances by identifying the boundaries from the continuous speech is a challenging task (Biadsy, 2011). In this work, the word and sentence datasets are prepared from KDSC and IViE datasets. The sentences considered in the present context resemble the collection of words conveying complete meaning. In this chapter, we proposed an algorithm for segmentation and creations of the word and sentence datasets by using spectral features. Dialect specific non-conventional features are extracted from these words and sentences. Dynamic (local changes) and static (global changes) in pitch and energy features are extracted from the pitch and energy contours obtained from both word and sentence units. Pitch and energy contours are used to derive intonation and energy patterns respectively. Legendre polynomial fit function of an order 14 is employed to capture the proposed intonation and energy patterns. Along with these, five gross statistical features namely as mean, minimum, maximum, standard deviation and variance of pitch and energy values are also extracted. These features represent the dynamic and static variations of the pitch and energy values across words and sentences. Proposed ADI systems are built by using individual and combination of dynamic and static prosodic, and spectral features. Experiments have been carried out using single SVM and decision tree based XGB ensemble approaches. Experiments are also conducted to verify the significance of the prosodic and spectral features from words and sentences for dialect identification.

Remaining part of this chapter is organized as follows: Dataset details along with sentence and word segmentation algorithm are provided in section 5.2. Proposed dynamic prosodic feature extraction details are discussed in section 5.3 with explanation of classification models. Section 5.4 contains the discussions on the results. Section 5.5 presents an overview and summary of the present work.

5.2 Word and Sentence Level Data Preparation

This section provides the details of the word and sentence datasets used in the present work. The procedures followed for creation of the word and sentence level utterances are discussed separately in the following sub-sections.

5.2.1 Sentence based Dialect Dataset

Dynamic thresholding on the energy and spectral centroid values is used to segment sentences from the spontaneous speech. Algorithm-1 shows the important steps followed for sentence segmentation based on two acoustic features namely short-term energy and spectral centroid (refer Section 3.2) combined evidence of these two features has been useful in recognition of clear sentence boundaries. These silent regions are characterized by lower energy and centroid values. Here, spectral centroid represents the center of gravity of its spectrum. Hence, lower energy and centroid features observed in between sentences were used to have identify clear boundaries of sentences. While applying this approach few of the sentences are not segmented correctly maybe due to the faster speaking style without a pause in between the sentences. Even some incomplete sentences are also observed among the segmented units. Hence, automatically segmented sentences are further processed/listened manually to retain only meaningful and complete sentence as processing units. Figure 5.1 demonstrates the use of two features for identifying segments representing the sentences in an audio clip of 15 seconds. In this audio sample, 6 sentence units out of 7 are correctly identified. About 84% of sentences are accurately segmented from spontaneous dataset used in this work.

Algorithm 1: Dynamic threshold based sentence segmentation algorithm

```
Function Sentence_Segmentation ($wavfile$)
  Input  : Wave file
  Output: Sentence segments
Step 1:  Read wave file
Step 2:  Extract spectral energy and centroid features from 50 ms non
   over lapping frames
Step 3:  Identify two thresholds dynamically from two histograms of
   energy and spectral centroid values.
Step 4:  Apply a simple thresholding criterion on the sequences and
   identify segments based on energy and centroid values
Step 5:  Name non-silent voiced regions as sentence segments
```

5.2.2 Word based Dialect Dataset

The voiced regions which resembles meaningful word utterances are segmented from spontaneous speech to prepare this dataset. Both Kannada and English datasets are segmented to have a word based dataset by using a dynamic threshold based algorithm (Giannakopoulos and Pikrakis, 2014). The steps given in

Table 5.1: Details of Word and Sentence dataset (KDSC) Legend: CENK-Central Kannada; CSTK-Coastal Kannada; HYDK-Hyderabad Kannada; MUBK-Mumbai Kannada; STHK-Southern Kannada

Sl.No.	Spoken Units	Name of the dialects					Total
		CENK	CSTK	HYDK	MUBK	STHK	
1	Words	1426	1247	1344	1240	1349	6606
2	Sentences	200	200	200	200	200	1000

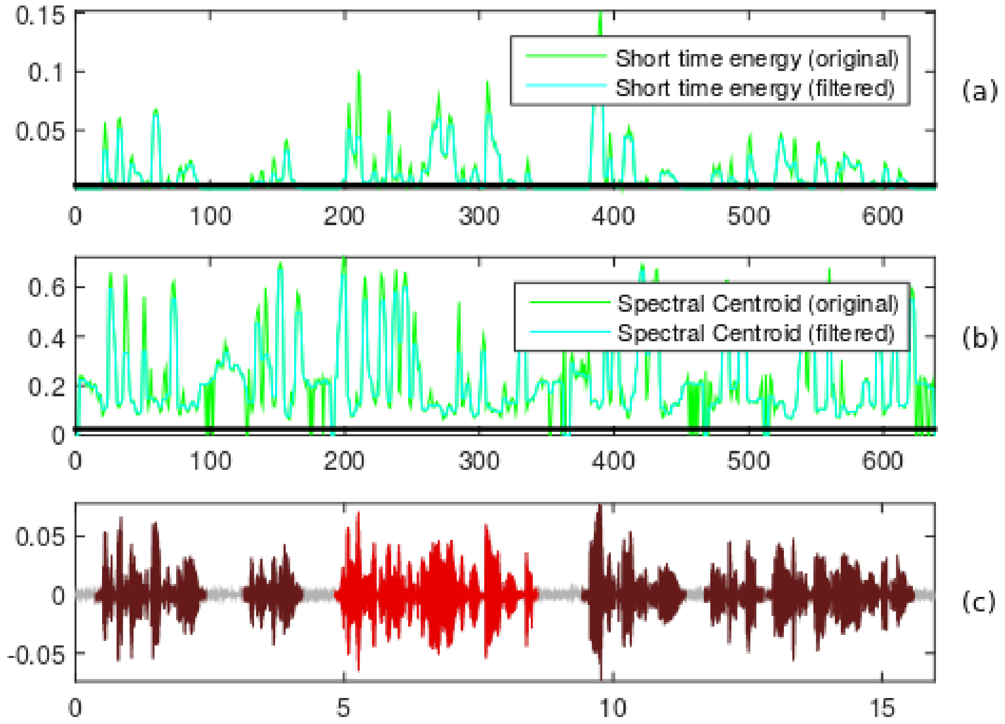


Figure 5.1: Threshold based Sentence Segmentation. (a) Energy based segmentation (b) Centroid sequence for segmentation. and (c) Segmented sentence units.

Algorithm 1 are also employed with some necessary modification and additional features for segmentation of spontaneous speech into words. Since words differ from sentences largely. Only energy and centroid features are not sufficient for extraction for segmentation. Additionally, spectral flux and ZCR features are used to capture the information of accurate word boundaries. However, among segmented word units there are several mis-identifications. Hence, joined words are further processed manually to segment and retain only correctly uttered and complete word units. Number of individual word and sentence units extracted from both Kannada and English datasets are presented in Tables 5.1 and 5.2 respectively.

5.3 Feature Extraction

This section discusses extraction of dynamic and static prosodic features used in this study.

Table 5.2: Details of Word and Sentence dataset (IViE dataset)

Sl.No.	Spoken Units	Name of the dialects									Total
		D1	D2	D3	D4	D5	D6	D7	D8	D9	
1	Words	1100	1100	1100	1100	1100	1100	1100	1100	1100	9900
2	Sentences	200	200	200	200	200	200	200	200	200	1800

5.3.1 Prosodic Features

Prosody can be viewed as one of the important speech features associated with larger units of speech such as syllables, words, phrases, and sentences. The prosody has direct correlation with the patterns of duration, intonation (F0 contour), and energy. It is known that prosodic knowledge plays a prominent role in the classification of dialects, emotions, and languages (Koolagudi and Rao, 2012). Similarly, the absence of prosodic cues can easily be perceived from the speech as it will be similar to read speech with just a concatenation of words. Prosodic features are said to add naturalness to a speech by structuring the flow of speech.

A Dynamic Prosodic Features

Majority of the times, dialects of languages vary with the use of unique prosodic attributes on the same vocabulary and script. These unique prosodic attributes are reflected through the distinct variabilities existing speaking patterns that includes intonation, stress patterns, prolongation, rhythm and so on. Indeed, the F0 change, observed due to rise and fall in the pitch while speaking, recognizes a unique phonological pattern followed in a sentence or word of a specific dialect. Prosodic features namely, duration, pitch (F0) and energy features are used in this work.

Pitch Contour: Subharmonic-to-harmonic ratio based pitch estimation algorithm with 30 ms frame and 15 ms step size is followed to extract pitch features (Sun, 2000). These features form the pitch contour for each sentence. Further, a Legendre polynomial fit function of order 14 is used to decimate these values. These 14 coefficients of Legendre polynomial have produced a good fit and these form a 14 dimensional feature vector. that represents the dynamism of feature

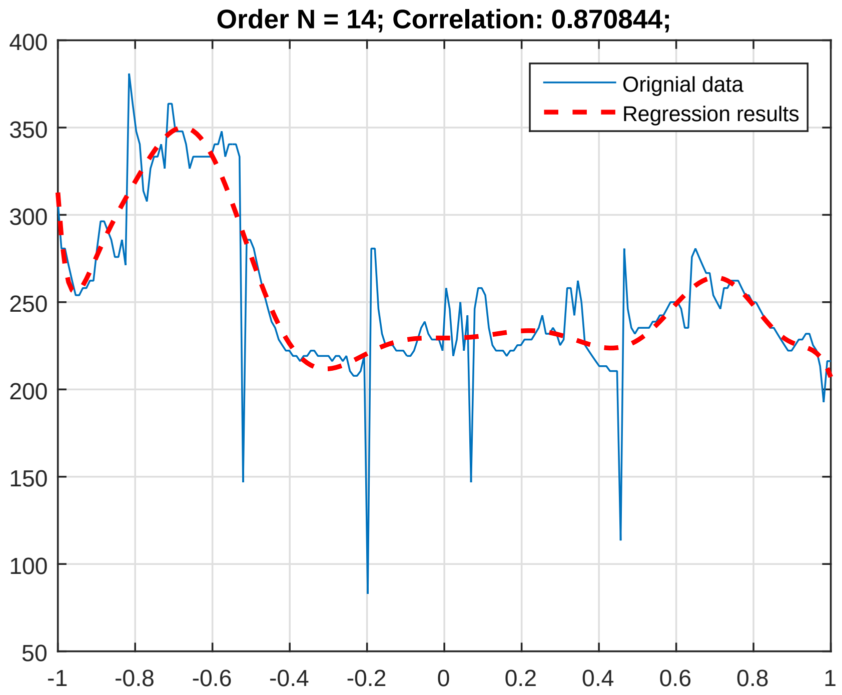


Figure 5.2: Legendre polynomial fit for a pitch contour

contours. The Legendre polynomial fit drawn for the pitch features extracted from the English sentence is shown in Figure 5.2.

Energy Contour: Similar to pitch contour, the energy contour is also derived by considering the variation in the amplitude values of the signal. Short term energies from the frames of size 30 ms and 15 ms step using the method given in Section 3.2. The number of features obtained from each speech file is different as the lengths of the files are different. Further, a Legendre polynomial function of order 14 is used to fit these values. 14 coefficients resulted in a good fit are treated in order as a 14-dimensional feature vector which also represents the dynamic stress patterns of different phonemes in a sentence. These 14 energy features are appended to the 14 pitch values creating the complete feature vector of size 28.

B Static Prosodic Features

The 28 dimensional feature vector obtained in A represents the dynamic variations in pitch and energy values in a sentence with respect to time. Apart from these, five

Table 5.3: Dialect recognition performance from **sentence** level utterances using SVM and XGB classifiers on KDSC and IViE datasets

Sl. No.	Features	Kannada Dialects		English Dialects	
		SVM	XGB	SVM	XGB
1	Dynamic Features	38.10	41.26	30.05	31.25
2	Static Features	42.51	42.79	40.12	39.67
3	Static Features	52.94	58.46	56.11	63.03
4	MFCCs	75.90	76.32	67.82	75.46
5	3+4	82.81	**85.38**	74.54	87.38

pitch and energy values namely, minimum, maximum, mean, standard deviation, and variance are derived from sentence and word units. These ten features are considered as the static prosodic evidence of the speech units used for identification of dialects. Inclusion of these 5 features forms 19-dimensional feature vector from pitch contour (14 dynamic+5 static). Similarly, the other 19-dimensional feature vector for energy leads to totally a 38-dimensional feature vector is formed when both dynamic and static parameters are combined.

5.3.2 Spectral Features

Spectral features play a decisive role in capturing the proper dialectal cues existing in languages and dialects. In this work, MFCC features (refer Section 3.2.2 for more details) are derived from the sentence and word utterances to exploit the vocal tract variabilities during the pronunciation of different dialects.

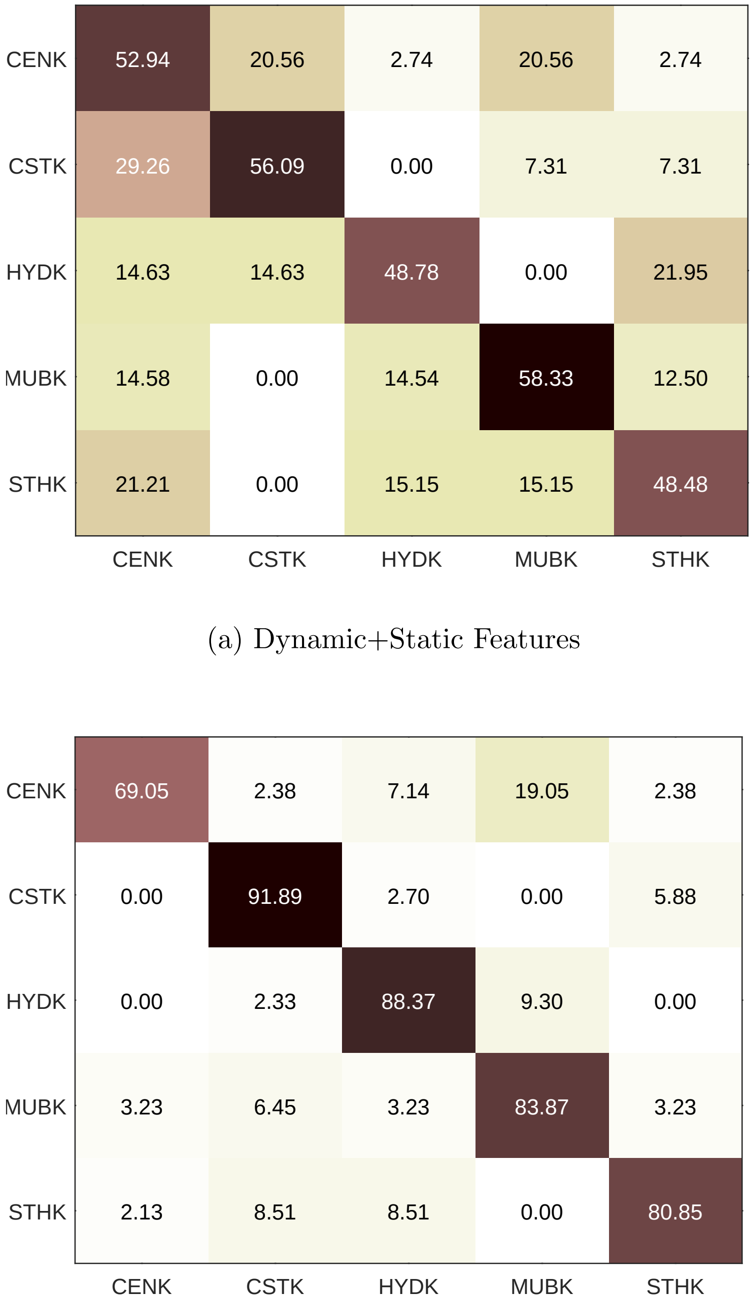
5.4 Experimental Results and Discussion

Automatic dialect identification systems are separately developed for sentence and word level units using static and dynamic prosodic features. Several experiments are conducted to evaluate the performance of the proposed approaches on Kannada and English language. Experiments were also carried out with the combination of features. Support vector machine (SVM) and XGB ensemble based algorithms are used for classification. Average performance of the classifiers is reported after employing five-fold cross-validation (CV) configuration.

5.4.1 Sentence based Dialect Identification System

The dialect recognition performance obtained by using dynamic, static, spectral features and their combination are given in Table 5.3. From the results, it is observed that the features have exhibited slightly lower dialect recognition performance when used individually. Whereas, the combination of features has demonstrated a higher accuracy. A confusion matrix obtained by combining the features is shown in Figure 5.3a. CSTK and MUBK dialect are classified with slightly better accuracies of 56.09% and 58.33% using sentences as units. In every dialect, high miss-classifications were observed. STHK dialect is found with the lowest performance, where 21.21% of sentences were misclassified with CENK and 15.15% with HYDK and MUBK dialects individually. HYDK and STHK dialects are observed to follow a similar speaking pattern that gets reflected into mutual misclassification. Where 21.95% of sentences of HYDK dialect are misclassified with STHK dialect. It is evident from the matrix that CENK & CSTK and HYDK & STHK dialects were mutually confused. This is due to the use of similar pitch and intonation patterns.

In order to investigate the system performance due to vocal tract system, 13 MFCC features are computed and appended to the set of prosodic features. Use of MFCCs has resulted in better recognition performance and demonstrated the existence of complementary information when combined with prosodic features. The average accuracy of about 82.81% is achieved with the combination of all features. The confusion matrix of this experiment is shown in Figure 5.3b. It may be observed from the figure that once again CSTK dialect is classified with the highest performance of 91.89%. Addition of MFCCs to the prosodic features has increased the classification performance of HYDK dialect to 88.37% from 48.78%. However, about 19.05% of CENK sentences were incorrectly identified with MUBK dialect and CENK sentences were remained confused even after addition of MFCCs. Results indicate that the proposed dynamic and static features have exhibited lower performances when used alone. However, their combination has improved performance indicating that both contribute complementary dialectal evidence.



(a) Dynamic+Static Features

(b) Dynamic+Static Features

Figure 5.3: Confusion matrices of dialect identification results on Kannada sentence level utterances using SVM. (a) Dynamic+static features and (b) Combination of Dynamic+static+MFCC features. Legend: CENK:Central Kannada, CSTK:Coastal (Karavali) Kannada, HYDK:Hyderabad Kannada, MUBK: Mumbai Kannada, STHK: Southern Kannada

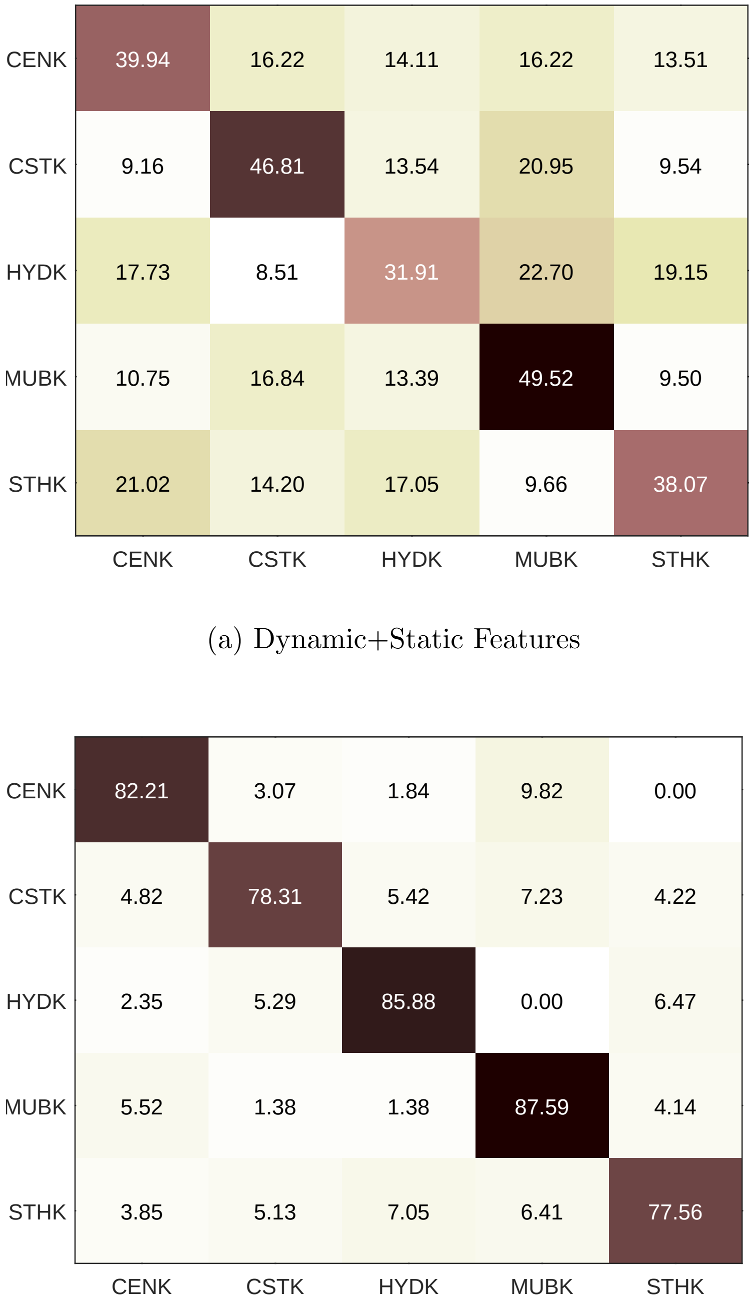
Table 5.4: Dialect recognition performance from **word** level utterances using SVM and XGB methods on KDSC and IViE datasets

Sl. No.	Features	Kannada Dialects		English Dialects	
		SVM	XGB	SVM	XGB
1	Dynamic Features	32.88	33.59	34.13	34.71
2	Static Features	35.12	44.00	38.50	44.09
3	Prosodic + Statistical Features	41.25	48.86	56.26	65.27
4	MFCCs	72.94	73.40	72.05	76.38
5	Combination of Features	82.31	**83.06**	81.50	**84.75**

Ensemble-based XGB classification algorithm has resulted in slightly better recognition performances over the single classifier based SVM method. A highest of 85.38% is achieved with a combination of all features with XGB. Higher accuracies are due to the use of several predictive models used in ensemble approaches. Besides, the proposed approach is tested on both Kannada and ADI systems are also developed on IViE dataset for comparison. It can be noted from the table that, English dialects are classified better compared to Kannada dialects using sentence level information. This may be due to proper articulation of intonation patterns compared to Kannada sentences.

5.4.2 Shorter Utterance (Word) based Dialect Identification System

Any listener can make conscious decision about the identification of dialects after listening to isolated words. Since every spoken word exhibits significant dialect discriminating cues (Purnell et al., 1999). After analyzing the role of dynamic and static prosodic features on effective sentence based dialect classification systems, experiments are also conducted over isolated word based dataset derived from KDSC and IViE speech corpora. Experiments are carried out separately using SVM and XGB classification methods on different features such as dynamic & static, MFCCs, and their combinations. The dialect recognition results are tabulated in Table 5.4.



(a) Dynamic+Static Features

(b) Dynamic+Static+MFCC Features

Figure 5.4: Confusion matrices of dialect identification performance using dynamic+static features and combination of other features on word dataset using SVM

From the results, it may be noticed that both dynamic and static features have reported lower dialect recognition performance when used individually. Whereas,

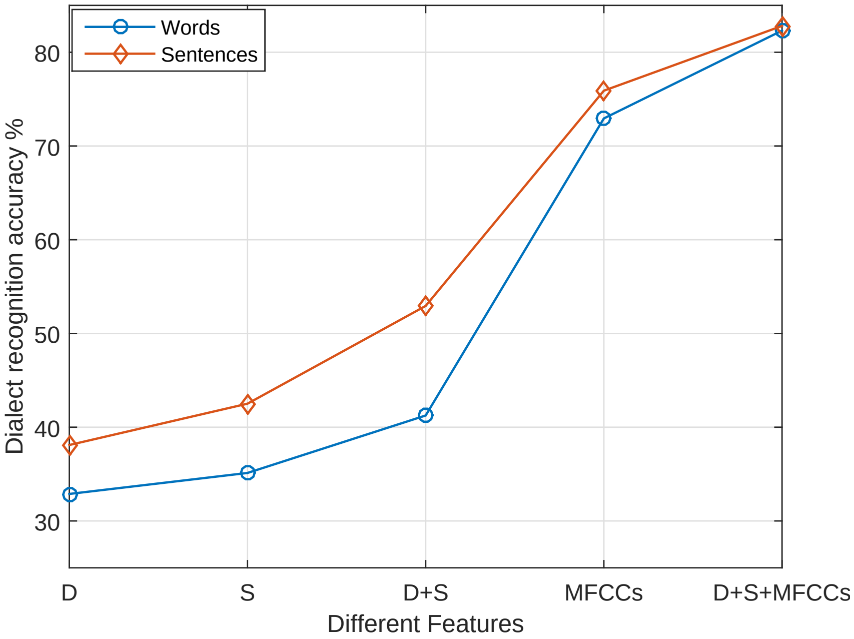


Figure 5.5: Comparison of dialect recognition performance using D-dynamic, S-static, MFCCs, and their combination using SVMs on word and sentence datasets

the dynamic and static feature set has demonstrated an average recognition rate of 41.25% using SVM on Kannada word based dialect dataset. A confusion matrix obtained with these feature sets is shown in Figure 5.4a. It can be observed from the table that, words of every dialect classes are highly confused among other dialects. The reason may be, that the word utterances are of short duration compared to the sentences; and they could not be effectively modeled with the features explored in this study. Among the dynamic and static features, static features have performed better. And the combination of both has increased the performance indicating the existence of complementary information at word level utterances too. Further, the standard MFCC features are added to the feature vector. These MFCCs alone have shown a performance of 72.94%. The combination of all features has improved the recognition rate to 82.31%. From these observations, one can conclude that from short utterances such as words, spectral features may capture dialect specific variations more effectively than the prosodic features.

In the case of word level utterances also, XGB based ensemble classifiers have resulted in somewhat better performance over individual classifier based SVMs.

One important observation is dynamic+static feature pair has resulted in improved performance of 65.27%. Combining the spectral parameters, represented by MFCCs with prosodic features has demonstrated a higher accuracy of 84.75%. Indeed, in all our experiments English words were better classified than Kannada words using proposed prosodic features and combination of features.

Generally, words and sentences represent the speech at two different dimensions, conveying dialect related information from both production and perception points of view. Dialect recognition performance achieved using proposed dynamic, static and MFCC features is presented in Figure 5.5. Dynamic and static features along with their statistical parameters are found to be successful in discriminating Kannada dialects. Pitch variations are better captured with five statistical parameters from sentences over the word units. Words are found to be shorter in length to capture intonation features. However, spectral (MFCC) features could identify the dialect related features from both sentence and words with almost similar performance.

5.5 Summary

In this chapter, an attempt has been made to recognize Kannada and English dialects from shorter (word) and (sentence) utterances using SVM and XGB ensemble algorithms. The dynamic threshold-based segmentation algorithm is used for creation of the word and sentence datasets from KDSC and IViE datasets. Dynamic and static prosodic features are used for assessment of dialectal differences along with MFCC features. Highest recognition rates of 82.81% and 85.38% are achieved with a combination of different features using SVM and XGB classification methods from sentence level utterances. Similarly, 82.31% and 83.06% of recognition results are observed from word level utterances. Further, the proposed prosodic features have shown better results when sentence level utterances are used for dialect classification. Comparatively, English dialects are classified efficiently with the use of both word and sentence level utterances. However, ADI performance obtained from these shorter word and sentence level utterances are comparatively lower than the little bigger utterances. In the next chapter we address the development of language dependent dialect identification systems for Kannada language. Kannada phonemes namely, vowels and consonants are considered for development of these ADI systems.

CHAPTER 6

Characterization and Identification of Kannada Dialects

Chapters 3, 4 and 5 have contained the discussion about the development of language independent ADI systems by using conventional and non-conventional (dialect-specific) speech features. This chapter provides details of development of language dependent ADI systems for Kannada language. The basic phonetic units namely, vowels and consonants are used for this purpose. Apart from these, details of Kannada *case* based ADI systems are also included in this chapter. The procedure followed for preparation of Kannada vowel, consonant and *Case* based dataset is provided. Details of acoustic-phonetic speech features extracted from speech to capture the dialectal cues are briefly discussed. Extraction and use of dynamic and static behaviors of speech through temporal dynamic properties and frame level global statistics features from vowel sounds are provided. Similarly, details of extraction of dialect-specific spectral variations from consonants and *cases* is discussed along with details of classification methods employed. Statistical analysis of vowels using Single Factor-ANOVA (Analysis of Variances) tests is provided. Experimental details with three decision tree based ensemble and SVM methods are covered. Analysis of results obtained from vowels, consonants and *Cases* using acoustic-phonetic spectral and prosodic features is provided at the end of the chapter.

6.1 Introduction

Dialects of a language are identified by the unique speaking patterns followed by the community of speakers belonging to a particular geographical region. The current chapter focuses on the classification of five dialects of Kannada language based on vowel and consonant sounds. The Kannada language is chosen, as it is one among the twenty-two official languages recognized by the constitution of India and it is the mother tongue for people of Karnataka. However, research activities are still in its nascent stage as far as dialectal studies are considered in Kannada language.

Kannada language vocabulary is equipped with forty-nine phonemic letters, which are divided mainly into three groups namely; *Swaragalu* (vowels), *Vyanjanagalu* (consonants) and *Yogavahakagalu* (Shridhara et al., 2013). Words in Kannada are a meaningful combination of *Aksharas*. Further, words are mainly bi and tri-syllabic sometimes even larger consonant clusters (Rajapurohit, 1982). In Kannada language, the speakers should be careful, while uttering any phoneme sounds, such as short vowels (/i/, /e/, /a/, /u/, /o/), long vowels (/i:/, /e:/ /a:/, /u:/, /o:/), unaspirated plosive (/p/, /b/, /t/, /d/, /k/, /g/, /T/, /D/) sounds, aspirated plosive (/ph/, /bh/, /th/, /dh/, /kh/, /gh/, /Th/, /Dh/), unaspirated affricates (/c/, /j/), aspirated affricates (/ch/, /jh/), fricatives (/h/, /s/, /sh/, /Sh/) and Nasal (/m/, /n/, /N/, /nY/, /nG/) sounds. Kannada script is characterized by an inherent vowel around which a syllable is built. For example, /ka/, in Kannada is a syllable of two phonemes as opposed to one phoneme in the English language. In Kannada, vowels may appear individually. Whereas consonants appear along with the support of vowels sometimes they appear individually at the end of the word called as dead consonants. Vowels are often observed as carriers of dialectal variations in Kannada than consonants (Zhenhao, 2015; Arslan and Hansen, 1996; Nagesha and Nagabhushana, 2007).

Stop consonants or plosive sounds represent one of the broad categories of phones in Kannada language. The production of a stop involves a complete closure of the oral cavity followed by the release of air in the form of noise burst. The stop consonants are differentiated from each other in terms of the manner of articulation (whether voiced and aspirated) and the place of articulation. Past research has suggested that acoustic attributes extracted from the burst spectrum can be useful in the classification of unvoiced stops Karjigi and Rao (2013). Also, it is reported in the literature that, dialectal variations can also be extracted from these shorter duration stop consonants.

Apart from vowels and consonants, the *cases* in Kannada language exhibit morphological peculiarities across the dialects. The grammatical function of a noun or pronoun is called as a *case* (Vibhakthi). Vibhakthi's expressed as through appending the suffixes (pratyayas) to the words specifically noun or pronouns. These convey the relationship between the words used in the sentences. These *cases* do not have any independent or unique meaning when used alone; however they, convey a different meaning when combined with the nouns. This distinctive usage of nouns and pronouns has demonstrated variations among five dialectal speakers of Kannada language. Based on this observation, in this thesis, ADI systems are developed by considering *case* level utterances.

In this chapter, three different ADI systems are discussed by using vowels, consonants, and *case* level information of Kannada language. First, the vowel-based ADI system is developed by analyzing dialectal variations in ten monophthongal vowels manually segmented from continuous speech. Three formant frequencies (F1-F3) followed by prosodic parameters such as F0 or pitch, energy or loudness and duration are extracted from the steady-state vowel region. These features are measured at different places of a vowel region to represent the fluctuations due to dialectal differences. Features like mean represent the global (static) phenomenon and features like fluctuations in vowel region depict the local (dynamic) behavior of in the corresponding vowels. To capture dynamic behavior, a Legendre polynomial fit of degree five is employed, as each coefficient of the polynomial relates to the formant frequency characteristics within the contour. Similarly, pitch and energy contours are also extracted for modeling dialectal variations in Kannada. This method of features extraction is also employed for extracting the features from consonants and *case* based utterances. Further, three decision tree based ensemble classification methods based on bagging and boosting are employed for evaluating the contribution of both local and global features towards recognizing the dialects.

This chapter in the subsequent sections discusses the procedure followed in the preparation of vowel, consonant, and *case* based dialect datasets in section 6.2. Feature extraction details, classification algorithms employed, statistical analysis of vowels along with result analysis are discussed in section 6.3. Features explored, result analysis and discussions on consonant-based ADI system are provided in section 6.4. Section 6.5 briefs about the various *cases* of Kannada language, features explored and analysis of results. Section 6.6 provides the summary of the present work.

6.2 Kannada Dialect Dataset Preparation for Characterization

Vowels highly influence the speaking pattern at phonetic and information levels (Clopper et al., 2005). Consonant pronunciation though lasts for shorter duration does exhibit unique dialectal cues (Themistocleous, 2016). *Case's* in Kannada have their own structures and patterns of pronunciation across different dialects. The procedure for suitable data preparation is discussed in subsequent subsections.

6.2.1 Vowel Dataset

In order to prepare the vowel dataset, few speakers whose speech is clear and legible are selected cautiously from the newly proposed KDSC (Refer section 2.7). In a continuous speech, vowel onset points are detected using spectral peak and excitation source features for automatic segmentation of vowels (Prasanna et al., 2009). However, these VOTs are found to be not accurate so that they cannot be used to segment complete vowel utterances. Since the present study demands demarcation of an accurate vowel steady state region, the manual segmentation process is applied after use of automatic one.

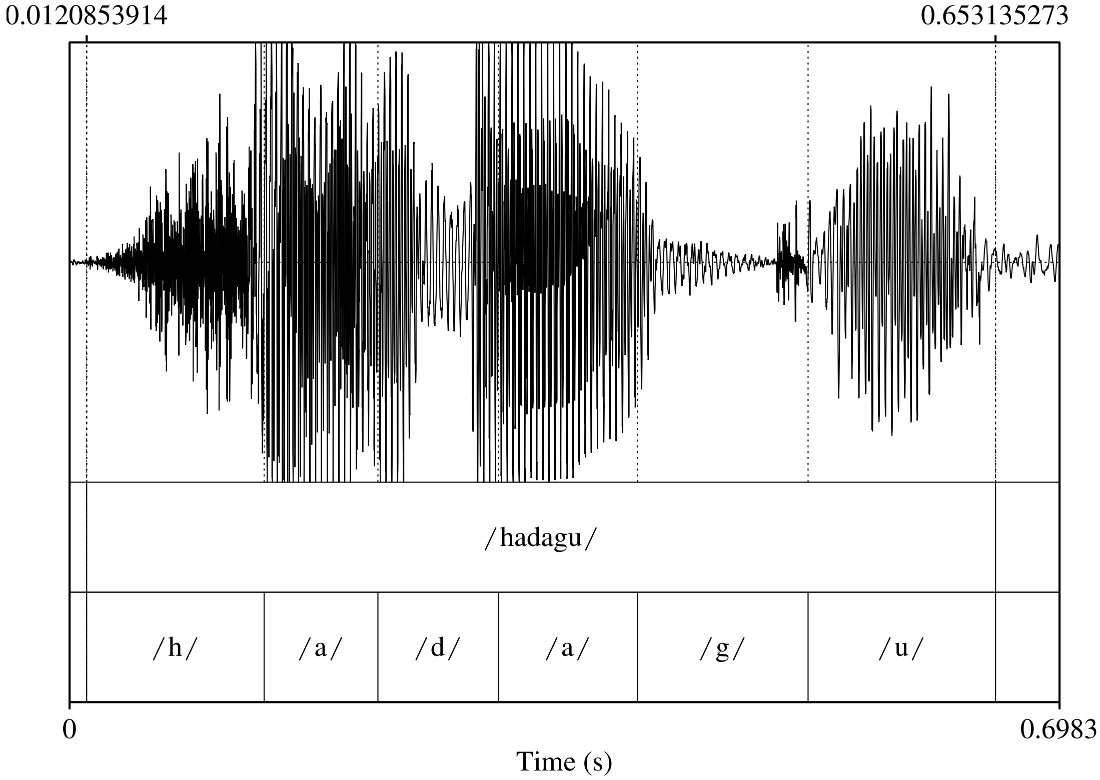


Figure 6.1: Manual segmentation of the word "/hadagu/" (Kannada word for the English word "Ship") in /hvd/ format using Praat

Manual segmentation process to segment out vowels includes the following steps. Annotated vowel regions using Praat are shown in Figure 6.1 (Boersma et al., 2002).

(i) Continuous speech utterances are segmented into words,

(ii) Specific words that contain interested vowels which pronounced intelligibly are chosen. Monosyllabic words are rare in standard Kannada; however, they are observed in few dialects. Majority of words are either bi or tri-syllabic or sometimes even more (Rajapurohit, 1982). While extracting vowels from spontaneous speech containing syllables of the form /VcV/ and /cVcV/ (V-vowels, c-consonants /p/,/k/,/d/,/b/,/h/,/s/) are chosen. The co-articulation effect of these consonants on the steady-state of the preceding and succeeding vowel is comparatively minimal. In specific, words with syllables of the form /hVdV/ are preferred over others since co-articulation is still minimal with /h/ and /d/ vowels (Hillenbrand et al., 2001).

(iii) From these syllables, vowels are located by manual inspection of the wave-

Table 6.1: Kannada Vowel dataset details, Legend:CENK-Central Kannada; CSTK-Coastal Kannada; HYDK-Hyderabad Kannada; MUBK-Mumbai Kannada; STHK-Southern Kannada

Sl. No	Dialect Name	Total No. of vowels	No. of Speakers (Male+Female)
1	CENK	520	5+5
2	CSTK	520	5+7
3	HYDK	520	5+4
4	MUBK	520	6+4
5	STHK	520	6+5

form, formants and intensity features in Praat tool. Vowel onset and offsets are located based on F1 and F2 values, by identifying the beginning of the gradual rise in the intensity and beginning of the gradual fall in the intensity respectively. The signal between vowel onset and offset is considered as the steady state region for both short and long vowels. Majority of the vowels are extracted from the initial and final positions of the words since there is less impact of co-articulation in this region. Also, vowels are said to be more distorted and dynamic, if exist in the middle of the words (Fogerty and Humes, 2012).

6.2.2 Consonant Dataset

Consonants dataset used in this work is derived from newly proposed KDSC. In this study, eight un-aspirated unvoiced and voiced consonants namely, /p/, /b/, /T/, /t/ /D/, /d/, /k/, and /g/ are used to study Kannada dialects. These eight consonants are commonly known as plosive sounds, as they are produced due to the constriction occurred at different regions in the mouth. Among these, /k/ and /g/ are produced with the constriction at the back of the tongue against the back of the roof of the mouth, the soft palate (velar); /k/ is voiceless and /g/ is voiced. /p/ and /b/ are produced because of constriction at the lips (bilabial), /p/ is unvoiced due to the absence of vocal fold vibration and /b/ is voiced. Similarly, /t/ and /d/, the dental consonants are produced with the constriction of the blade of the tongue against the ridge behind the upper teeth (dental); /t/ is voiceless. Retroflex consonants are also called cerebral consonants are produced with the tongue tip turned back to the hard palate behind the alveolar ridge. The distinctive features of some retroflex consonants is that the tongue curls up slightly on itself when produced. There are no English equivalents for these. Very rarely vowels occur in the initial position of the word in the Kannada language. It is observed that the duration of plosive sounds is concise; hence segmentation of these consonants and extraction of significant dialect-specific features from them is sensitive and a tedious task. Hence, in this work, all monosyllabic consonants (CV units) where, plosive sounds are combined with vowels /a/, /u/, /i/ or /o/ are used for experimentation, since co-articulation between them is comparatively lesser (Kalaiah and Bhat, 2017). The /CV/ tokens are identified in spoken utterances manually using Praat tool. The onset of the stop burst and offset of the vowels are identified through simultaneous inspections of both waveforms and the spectrograms (Boersma et al., 2002). /CV/ syllables are segmented by detecting burst onset, /CV/ transition and complete vowel utterances. Majority of the /CV/ units are chosen from the word-initial positions. A consonant dataset considered in this study consists of 2417 consonants extracted from 16 (9 Female + 7 Male) speakers from each dialect. The total number of consonant clips available for each dialect are as follows, CENK-455, CSTK-478, HYDK-484, MUBK-501, STHK-499.

6.2.3 Kannada *Case* Dataset

The grammatical function of a noun or pronoun is called *Case* (Vibhakthi). These are suffixes that get added to the words specifically to noun or pronouns. These convey the relationship between the words used in the sentences. These cases do not have any independent or unique meaning when used alone; however, convey and alter the meaning when combined with the nouns in different contexts. In this work, an attempt has been made to utilize the dialectal cues present in the usage of this Kannada *cases*. For this purpose, a *case* dataset is prepared from KDSC. Details of the dataset with five out of seven *cases* used in the Kannada language is given in Table 6.2. Every speech sample was listened and the words with *cases* have been segmented manually using Praat tool.

Table 6.2: Kannada *Case* dialect dataset

Sl. No	**Kannada Dialects**	**Case Dataset**		**Total**	**Kannada Suffixes**
		Male	Female		
1	CENK	266	138	404	/u/, /annu/, /inda/, /ge/, /ke/, /ige/, /alli/
2	CSTK	164	138	302	/u/, /annu/, /inda/, /ge/, /ke/, /ige/, /alli/
3.	HYDK	326	62	388	/u/, /annu/, /inda/, /ge/, /ke/, /ige/, /alli/
4.	MUBK	54	140	194	/u/, /annu/, /inda/, /ge/, /ke/, /ige/, /alli/
5.	STHK	278	116	394	/u/, /annu/, /inda/, /ge/, /ke/, /ige/, /alli/

Table 6.3: Kannada vowel organization

Kannada Vowels				
		Front	Central	Back
High	Short	/i/		/u/
	Long	/i:/		/u:/
Mid	Short	/e/		/o/
	Long	/e:/		/o:/
Low	Short		/a/	
	Long		/a:/	

6.3 Vowel based Dialect Identification System

Standard Kannada language has vowels classified into three groups based on place and manner of articulation namely, front, central and back vowels. Similarly, based on the height of tongue hump vowels can be classified into high, mid and low vowels. Vowel height such as low or high refers to the vertical position of the tongue relative to the roof of the mouth (Reetz and Jongman, 2009). Table 6.3 provides the organization chart for ten Kannada vowels. This study includes five long and five short vowels for classification of dialects.

6.3.1 Feature Extraction

Acoustic characteristics such as formant frequencies extracted from steady vowel regions play a significant role in discrimination of different vowels. The articulatory configuration corresponding to each sound, co-articulation effect, the context of the sound units, gender variability, and emotional status are also important reasons for acoustic variability (Reetz and Jongman, 2009). Average LP spectra are drawn for the vowel /e/ uttered by a randomly selected speaker in five dialects are shown in Figure 1.4. From spectra, vowel inherent spectral properties indicating the existence of systematic and significant differences among five dialects can be observed. These variations may be observed through the gaps in energy levels, spectral peaks, spectral sharpness and positions of formant frequency values (F1-F4) among five dialects of the Kannada language. In this study, values of F1-F3 (formant frequencies), pitch (F0), the energy of speech frames and vowel duration are used for characterization of five dialects.

Formants of a phoneme correspond to resonant frequencies of the oral cavity structure during the pronunciation. First three formants play a prominent role in identifying vowels. Among all formants, F1 the lowest formant demonstrates vowel heights. High front vowels namely, /i/, /i:/, /u/, /u:/ (refer Table 6.3) are associated with low F1 values and low mid vowels such as /a/, /a:/ are with higher F1 values. Similarly, F2 is associated with tongue advancement. Fronting and backing are observed with high and low F2 values respectively (Reetz and Jongman, 2009). F3 is commonly used for discrimination of rounded and unrounded vowels. Both F2 and F3 vary proportionally to the degree of constriction and amount of lip rounding.

LPC based McCandless formant tracking algorithm is employed to extract the three formant frequencies from the steady-state region of the vowel with a 10 ms overlapped 20 ms frame (McCandless, 1974). LPC is a widely used method for formant extraction due to its compact and accurate computation. 14 order LP spectrum is computed on 8kHz down sampled speech signal by using continuity constraints from voiced regions (Rabiner and Juang, 1993). As per this algorithm, the first three peaks of a spectrum are considered as the three formants of interest obtained from the steady-state region of the vowel. It is also noticed that due to physiological differences in vocal tracts, formants obtained from female and male speakers may differ slightly. In this work, gender differences w.r.t. formants are observed to be relatively less (Johnson, 2008).

Use of formant values alone for recognizing dialects may not be a practical approach due to the following reasons: (i) Inappropriate modeling of physiological aspects of speakers, a perceptual approximation of vocal tract length for measuring formants. (ii) Presence of noise in the recorded speech. (iii) Use of formant frequencies computes the spectral variations in the pronunciation. (iv) Dialectal differences of language mainly span over more extended duration units, such as suprasegmental or prosodic features (Harris et al., 2014; Rouas, 2007). The features representing the perceptual properties of speech such as intonations, loudness, speaking rate, and so on also exhibit significant dialectal cues. The imposition of intonation, different durations, and intensity patterns in the spoken speech make it more natural. Majority of dialects of the Indian languages are said to have prosodic variations (Mehrabani and Hansen, 2015; Rao and Koolagudi, 2011; Reddy et al., 2013). It is also observed in the case of Kannada dialects (Soorajkumar et al., 2017).

With this understanding to verify the existence of prosody cues; pitch, energy and duration features are extracted from shorter vowel segments. Pitch (F0) features are extracted using a subharmonic-to-harmonic ratio based pitch estimation method (Sun, 2000). Rise and fall of pitch over a time constitute the intonation patterns, and these variations represent unique patterns in dialects. Energy features associated with the speech signal is time varying in nature. It directly corresponds to the loudness. Short time energy is computed from vowel sounds for describing the loudness. It plays a prominent role in human aural perception. Energy variations with time are measured from samples' amplitudes within a frame.

Instead of using spectral and prosodic features directly use of dynamic and static behaviors of these features would be beneficial for characterizing the dialects. As steady-state region of a vowel, generally spreads across many frames, it would be useful if vowel behaviors are analyzed across the frames. The procedure followed is as follows. Local variations during the pronunciation of vowels across dialects are extracted from three formants (F1, F2, F3), pitch and energy contours. A

Legendre polynomial fit function of order five is employed to capture the dynamics These represent the temporal dynamic characteristics of phonation, intonation and loudness respectively. These features are combined in the order F1, F2, F3, pitch, and energy to form a feature vector of size 30 (5*6). Later, the duration of a single vowel in ms is added to this feature vector leading to a feature vector of dimension 31, to capture local variations among vowels. In addition to local features, global features are also extracted from vowels. For each vowel, five statistical values namely, mean, min, max, standard deviation, and variance of F1, F2, F3, pitch, and energy are extracted to form 25 dimensional feature vector. These are generally known as the global features representing static behavior of the vowel.

6.3.2 Statistical Analysis of Features

Every dialect follows a unique set of articulatory patterns of tongue, lips, jaw, palate, and teeth while producing speech. Due to this, differences do exist between dialects in terms of formant and pronunciation duration values of vowels (Adank et al., 2004; Zheng et al., 2012). The significance of these features is analyzed in this subsection with respect to ten vowels of five Kannada dialects. The present study is conducted on the mean values of chosen features of ten identified vowels.

A smoothed average LP spectra drawn for vowel /e/ are shown in Figure 1.4. The figure shows the existence of variations in spectral shape across dialects. A traditional F2-F1 plot shown in Figure 6.2 is drawn using mean values of F2 and F1 for five vowels chosen from each of the dialects. Plot shows the vowels in the order, /i/, /e/, /a/, /o/, and /u/ based on the place of articulation from fronting to backing. Front vowels /i/ and /e/ of STHK and HYDK dialects are noticed with high F1 and F2 values than remaining dialects. Speaking styles followed in these dialectal regions are found to be distinct with higher energy, pitch, and speaking rate. Due to which, higher F2 values are observed for front vowels. Central-low vowel /a/ is found with almost similar F1 and F2 values for all dialects.

CSTK dialect spoken in the coastal region of Karnataka has exhibited lesser F1 and F2 values for the vowels /e/, /a/ and /u/. Back-mid vowel /o/ pronounced

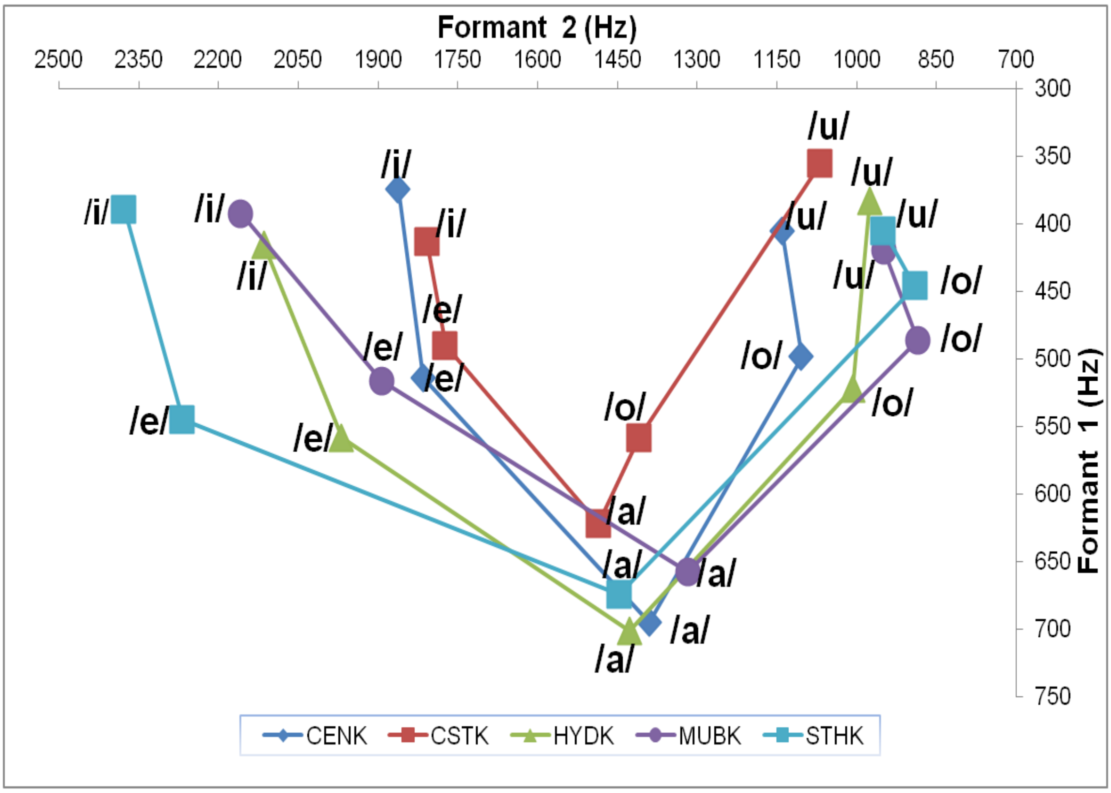


Figure 6.2: Articulatory F2-F1 plot of vowels of five Kannada dialects

in this dialect has shown higher F1 and F2 values. Speakers from this region speak Kannada occasionally (being Tulu as majority language) with consciousness resulting in fewer variations. Also, Kannada is an L2 language, the majority of the people in CSTK dialectal region speak Tulu language (L1) and are noticed with lower energy and pitch features. Furthermore, the CENK vowels are also found with small values of F1 and F2 similar to CSTK dialect. Speaking patterns of CENK and CSTK are closer to the written/standard form of Kannada, and every utterance is phonetically clearly pronounced. The plot has clearly shown the existence of varying patterns followed in vowel pronunciation among different dialects.

Statistical analysis of the features is carried out using one-way ANOVA. Mean values of F1, F2, F3 and duration features are computed for all ten vowels. F-test results, mean and standard deviation of ten vowels of 5 different dialects are given in Tables 6.4 and 6.5. *F-stat or F-ratios* measure whether the means of different samples significantly vary or not. Many a times, considering only

F-stat is not sufficient and hence even *p-value* is considered. *p-value* represents the probability of finding the observed results. Here, the hypothesis considered is formants obtained from vowels have significant differences across five dialects. Significance level α is set to 0.05.

From Table 6.4, higher F1 values are observed with /i/, /a/, /o/ and /o:/ vowels of HYDK dialect. F2 values are observed to be high with front vowels (/i/, /e/, /e:/, /a/) in STHK region. Among all dialects F2 values of MUBK are found to be low for all vowels. Similarly, F3 values are also low for Dialect CENK, CSTK and HYDK utterances. Vowels /e:/, /a/ and /u/ of MUBK and /i/, /i:/, /e/ and /u:/ of STHK have shown higher values of F3. From these tables, with /e/ and /u:/ vowels it is clearly noticed that *p-value* is less than 0.05 and also *F-stat* is greater than *F-crit* indicating that there are significant differences across dialects in pronunciation. ANOVA results for vowel /e/ are, [F(4,454)=4.61, p=0.00032] for F1, [F(4,454)=2.81, p=0.036] for F2 and [F(4,454)=3.67, p=0.018] for F3. Where, F(4,454) indicates that F statistic of ($df_{between}$ (number of class-1, 5-1), df_{within} (N-k)). Also, /u:/ vowel is seen to be vary significantly with [F(4,250)=10.38, p=0.0001] for F1 and [F(4,250)=10.09, p=0.0001] for F2 values across dialects. Whereas, F3 is not contributing in dialect discrimination with vowel /u:/.

Table 6.4: Mean and standard deviation of the Three formants (F1, F2, F3) and F -test of ten long and short vowels, SD-Standard Deviation, D1-CENK, D2-CSTK, D3-HYDK, D4-MUBK,D5-STHK

Kannada Vowels	ANOVA F- Test for three Formant frequencies								
	F1 Formant			F2 Formant			F3 Formant		
	F-stat	p-value	F-crit.	F-stat	p-value	F-crit	F-stat	p-value	F-crit
/i/	0.55	0.70	2.61	1.27	0.29	2.59	1.79	0.147	2.59
/i:/	2.09	0.12	2.70	1.35	0.27	2.67	4.94	**<0.003**	2.67
/e/	4.61	**<0.0032**	2.56	2.81	**<0.036**	2.57	3.67	**<0.018**	2.56
/e:/	0.96	0.44	2.62	0.99	0.42	2.61	1.05	0.390	2.63
/a/	1.87	0.11	2.43	2.00	0.10	2.43	1.06	0.376	2.43
/a:/	1.43	0.24	2.64	0.35	0.83	2.64	1.49	0.224	2.64
/o/	2.32	0.11	3.11	9.21	**<0.0007**	3.11	1.26	0.328	3.12
/o:/	0.88	0.49	3.06	1.13	0.39	3.11	1.44	0.268	3.06
/u/	1.22	0.33	2.84	2.16	0.11	2.92	0.62	0.658	3.01
/u:/	10.38	**<0.0001**	2.89	10.09	**<0.0001**	2.89	0.74	0.578	2.90

Table 6.5: Mean and standard deviation of the duration feature measured in milliseconds and F-test of ten long and short vowels, SD-Standard Deviation, D1-CENK, D2-CSTK, D3-HYDK, D4-MUBK, D5-STHK

Vowels	**Duration (MS)- Mean**					**Duration (MS)- SD**					**F- Test**		
	D1	**D2**	**D3**	**D4**	**D5**	**D1**	**D2**	**D3**	**D4**	**D5**	**F-stat**	**p-value**	**f-crit**
/i/	61	60	53	76	56	13	12	4	15	14	2.69	<0.043	2.58
/i:/	107	98	89	102	107	11	20	20	22	22	0.99	0.420	2.66
/e/	58	70	56	64	59	12	11	11	15	16	1.98	0.113	2.56
/e:/	108	126	118	113	108	26	28	10	19	14	1.08	0.376	2.63
/a/	57	63	54	64	71	13	16	10	16	16	6.02	<0.0002	2.43
/a:/	113	107	89	105	118	22	15	8	9	11	4.35	<0.0054	2.62
/o/	80	66	51	75	91	13	15	2	2	3	9.5	<0.0006	3.11
/o:/	116	106	92	117	106	6	6	13	11	19	2.61	0.081	3.11
/u/	58	95	59	65	79	10	28	7	12	13	4.95	<0.006	2.84
/u:/	125	130	85	119	79	26	14	7	9	52	7.43	<0.0006	2.81

From Table 6.4 representing F2 statistics, it is noticed with vowel /o/ with [F(4,250)=9.21, p=0.0007] indicating that F2 values have statistically significant differences among dialects. Existence of differences in F2 values are also noticed. Further, the vowels /i:/ [F(4,260)=4.94, p=0.003] and /e/ have indicated statistical variation across five Kannada dialects. From these tables, it is noticed that few of the vowels show varying values of F1, F2, and F3. Intern showing their significant evidential variations across five Kannada dialects and hence these (F1, F2, F3) features can be used for automatic classification.

Further, the statistical analysis on the duration of vowels is carried out with ANOVA tests. A hypothesis was made that the phonetic duration of the vowels plays a major role in distinguishing five dialects. The results revealed a considerable difference for vowels /i/ [F(4,260)=2.69, p=0.043], /a/ [F(4,350)=6.02, p=0.0002], /a:/ [F(4,210)=4.35, p=0.0054], /o/ [F(4,250)=9.5, p=0.0006], /u/ [F(4,290)=4.95, p=0.006], and /u:/ [F(4,250)=7.43, p=0.0006]. Overall, HYDK has shorter average vowel duration. Most of the vowels except /e/ have shown the distinction between dialects. Further, in the next section contribution of every feature considered in this work is assessed w.r.t. classification of dialects.

The energy values extracted from the speech samples of five dialects are analyzed, and the box plot is drawn to represent the first order statistics of the five Kannada dialects in Figure 3.6.Even more interesting observations are made from Figure 6.2 and 3.6, where higher values of F1 and F2 are observed HYDK and STHK dialects and reflected with larger interquartile range indicating higher energy range of values. Similarly, lesser F1 and F2 values are seen for CSTK region with a shorter interquartile range. It can also be noted that F1 and F2 values are correlated with energy features.

6.3.3 Experimental Results and Discussion

Dialect classification experiments are performed using acoustic-phonetic features considering their global and local behavior. Separate systems are developed with dynamic and static feature vectors using three decision tree based ensemble algorithms. The dialect classification performance is evaluated through simple vali-

Table 6.6: Dialect recognition performances using dynamic (local) features, Legend: E-Energy, Dur-Duration, RF-Random forest, ERF-Extreme random forest, XGB-Extreme gradient boosting

Feature Set	Accuracy in (%) Contour features					
	RF		ERF		XGB	
	SV	CV	SV	CV	SV	CV
Formants	60.83	56.45	62.5	60.2	63.88	55.97
F0+E+Dur.	59.72	54.44	60.55	57.01	55	54.86
Formants +F0+E+Dur.	69.16	64.37	72.78	68.41	71.94	63.95

dation (SV) and cross-fold validation (CV). Combination of formant and prosody features is evaluated towards dialect identification.

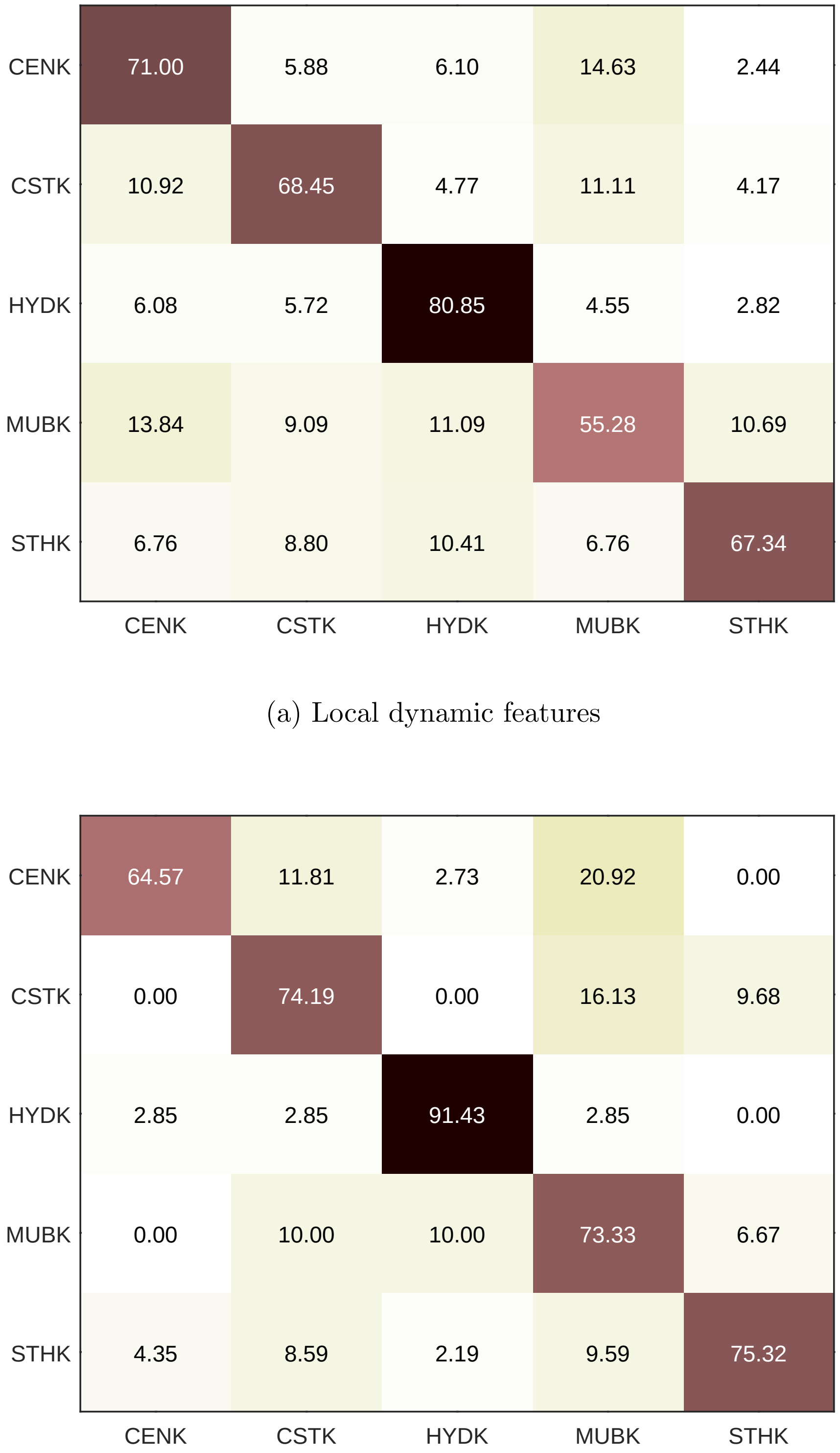
The overall performance of vowel-based ADI system using dynamic behavior is given in Table 6.6. Only CV results are considered for further analysis as their assessment is more practical. The individual contribution of formant frequencies and prosodic features of vowels is assessed from the results. Dynamic formant frequencies have resulted in the classification of Kannada dialects with the highest performance of 60.20%. Similarly, F0, energy, and duration of vowel regions yielded around 57% dialect identification. In order to use the complementary information about dialect recognition, both features are combined to prepare the new feature vector. With this combined feature vector a better performance of 68.41% is noticed.

ADI system is also developed using global features, which include statistical mean values of the features. The overall performance of vowel-based ADI system using global behaviors is given in Table 6.7. Highest dialect recognition performance of 62.33% and 64.96% are obtained with ERF model using formants and prosodic features respectively. Slightly increased classification rate is observed with prosody features of vowels. Similar to dynamic features, the combined feature vector produced an overall improved dialect recognition of 75.74% . Comparison of classification results obtained with three classification methods is given in Figure 6.4.

For clear understanding, confusion matrices obtained with both local and global features are given in Figure 6.3. Confusion matrices are also shown for a fusion of formants and prosodic features. Figure 6.3 (a) is the confusion matrix showing an average accuracy of about 68.41% with dynamic local features using the ERF model. Figure 6.3 (b) is the confusion matrix of the results of global features. The average dialect recognition accuracy of about 75.74% is obtained with global features.

Statistical analysis performed with ANOVA has shown a similar contribution of F1, F2, and F3 in the classification of Kannada dialects. Phonetic durations of Kannada vowels varies extensively, indicating the use of different articulation rates for identifying different Kannada dialects. Duration is expected to contribute positively to dialect discrimination. Acoustic features are extracted from vowels to evaluate dynamic and static behavior of signal across the dialects. Static features derived from statistical parameters have yielded in higher performance. It is found to be better over the local features derived from contour trends.

From the confusion matrices, it is noticed that HYDK dialect is classified with high performance among all other dialects. Speakers of HYDK dialect use a unique pronunciation style with high pitch, loudness, and faster speaking rate. These patterns are captured clearly through the global features over local features (Chittaragi et al., 2019). Vowels of MUBK dialect are classified with only 55.28% accuracy. 13.84% of MUBK vowels are misclassified as CENK, and 14.63% of CENK vowels are misclassified as MUBK dialect due to the similar energy characteristics during pronunciation. Vowels of CENK, CSTK, and STHK dialects are classified with the better performance when dynamic features are used. Compared to local features, better performance is observed with the use of global vowel features. It is noticed that CENK and MUBK pair are mutually confused when local features are used for classification. Whereas, with global features, CENK clips are misclassified with MUBK, but vice versa is not observed. This is due to the reason that global statistics which successfully characterize the MUBK dialect. Among these, 16.13% of CSTK vowels are misclassified with MUBK ones.



(a) Local dynamic features

(b) Global static features

Figure 6.3: Confusion matrices of dialect identification results using dynamic and static features with ERF classifier on Kannada dataset

Table 6.7: Dialect recognition performances using static (global) features, F0-Fundamental Frequency, E-Energy, Dur-Duration, SV-Simple Validation, CV-Cross-fold Validation, RF-Random forest, ERF-Extreme random forest, XGB-Extreme gradient boosting

Feature Set	Accuracy in (%) For Frame wise statistically derived features					
	RF		ERF		XGB	
	SV	CV	SV	CV	SV	CV
Formants	68.89	61.23	68.41	62.33	68.89	57.22
F0+E+Dur.	70.56	65.43	71.11	64.96	65.55	58.70
Formants+F0+E+Dur.	77.22	74.13	78.89	**75.74**	75.56	73.21

Table 6.8: Performance contribution of different prosodic features in dialect identification, E-Energy, Dur-Duration, F0-Pitch, Rank of the features is shown in brackets

Features	Formants (62.33%)			Prosodic (64.96%)			Combined (75.74%)					
	F3(1)	F2(2)	F1(3)	Dur. (1)	E(2)	F0(3)	Dur.(1)	E(2)	F0(3)	F3(4)	F2(5)	F1(6)
Contribution in (%)	33.97	33.17	33.10	36.42	33.97	29.59	23.13	21.64	15.04	14.91	12.74	12.6

Further, it may be interesting if the contribution of each of the features is analyzed to understand their significance in vowel-based dialect classification. In this regard, the analysis of the contribution of each feature individually and in combined form is carried out. The results obtained with the ERF model using cross-validation are tabulated in Table 6.8. ERF algorithm is used as it has performed better (refer Figure 6.4). A few interesting facts observed are as follows: The role of three formants is found to be similar. Prosodic features have contributed in an order duration, energy and F0 to the classification of dialects indicating the importance of duration (speaking rate) among the dialects. F0 has shown comparatively less contribution. When features are combined, a sequence of duration, energy, F0, F3, F2, F1, has shown the significant contribution for dialect classification based on Kannada vowels.

Though vowels exhibit dialect wise distinction of formant frequencies, these are not sufficient to discriminate the dialects of Kannada language. Prosodic features have shown a little higher contribution toward dialect classification, especially prosodic global features have performed better in Kannada dialect classification over formants features. Among acoustic features considered, the contribution of the vowel duration, indicating the speaking rate of the vowel is identified to be higher (Table 6.8). Interestingly, HYDK dialect is classified in a better way with both dynamic and static features. It follows where they follow different and faster speaking rate among five dialects indicating that the duration playing a prominent role in classification. Among intensity and F0, intensity contributed better over F0 towards Kannada dialect identification. Combination of both formants and prosodic features has resulted in better dialect recognition performance from vowels sounds.

6.4 Consonant based Dialect Identification System

Consonants have a significant role in the speech however, these consonants have no meaning if used alone in the Kannada language. These consonants are com-

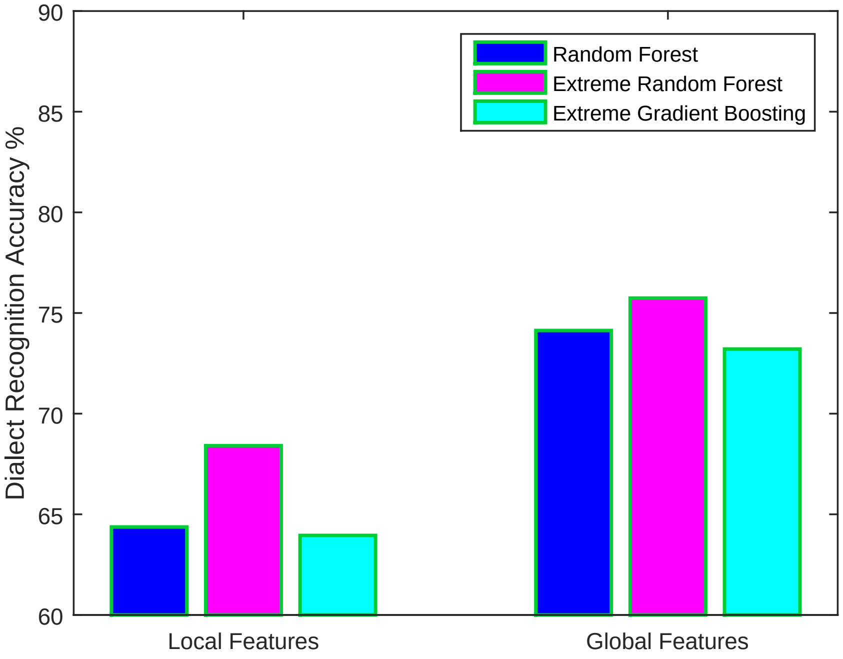


Figure 6.4: Comparison of Kannada dialect recognition performance using local and global features

plete and carry meaning when used along with the vowels. Studies conducted on consonant processing are comparatively lesser over vowel studies due to their short duration and abrupt behavior. However, in the literature, several systems are proposed primarily for classification of consonants, as every consonant is uttered with unique articulatory activity (Karjigi and Rao, 2013; Kodukula, 2009). Literature has reported substantially fewer systems that have addressed dialect classification using consonant properties.

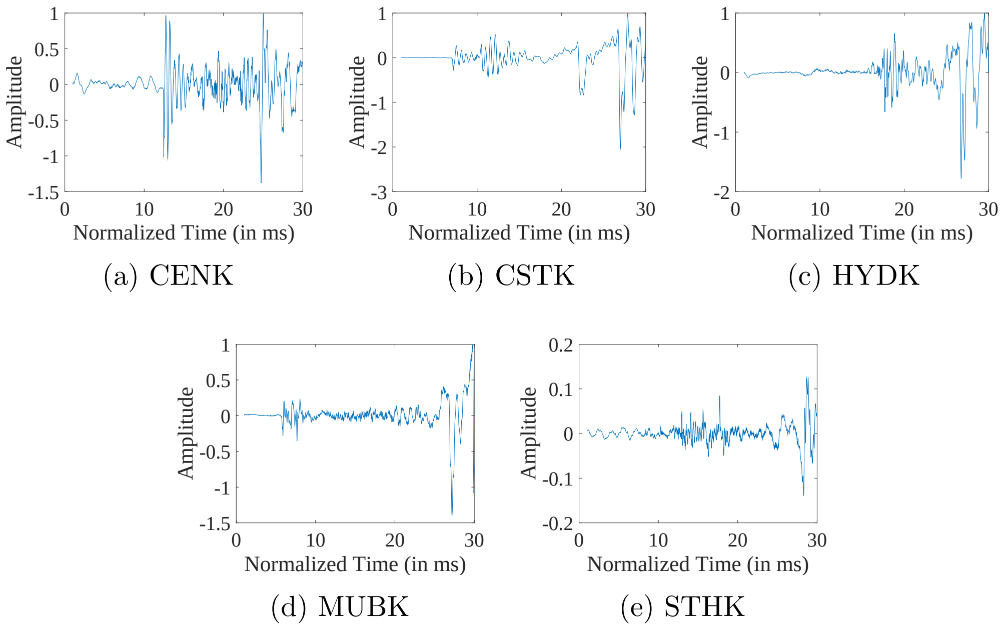
This study focuses on the investigation of the significance of the stop consonants on five dialects of Kannada language. Eight un-aspirated unvoiced and voiced consonants namely, /p/, /b/, /T/, /t/ /D/, /d/, /k/, and /g/ are used for classification of Kannada dialects. These eight consonants are commonly known as plosive sounds, as they are produced due to the constriction occurred at different regions in the mouth. Plosives of Kannada have been typically observed with silent period in the closure phase, as closure duration of /k/, /t/, /p/, and /T/ are longer when compared to /g/, /d/, /b/, and /D/. Also, voiced plosives show voicing bar at the low frequencies. Among these, /k/ and /g/ are produced with the constriction of the back of the tongue against the end of the roof of the mouth, the soft palate (velar); /k/ is voiceless. /p/ and /b/ are produced because of constriction at the lips (bilabial), where, /p/ is unvoiced due to the absence of vocal fold vibration and /b/ is voiced. Similarly, /t/ and /d/ dental consonants are produced with the constriction of the blade of the tongue against the ridge behind the upper teeth (dental); /t/ is voiceless. Retroflex consonants are also called as cerebral consonants produced with the tongue tip turned back to the hard palate behind the alveolar ridge (Rajapurohit, 1982).

The research findings in the existing literature show that besides use of linguistic information, such as place of articulation, and production mechanism, the spectral analysis of speech signals can be performed for the identification of language and/or dialect information. Hence, in this study, dialect specific cues are explored from spectral statistical features. Classification models such as SVM and XGB are used for developing dialect identification systems.

6.4.1 Feature Extraction

Features those significantly differentiate the dialects of Kannada are extracted from the spectral representation of the speech signal.

Formants: As oral cavity is closed during the constriction at the specific place in vocal tract, formants may not be available. However, closure is followed by the release of noise burst, due to which front cavity is excited by a sudden reduction in downstream. This, in turn, leads in shifting of formants either upwards or downward along with amplitude. This is depending on the place of constriction of the consonant and the following vowel. Three formant frequencies play a significant role in identifying different vowels. Similarly, vowels, along with stops, can also encapsulate variations that occur during the pronunciation variations (Reetz and Jongman, 2009). LPC based McCandless formant tracking algorithm is employed to extract the three formant frequencies from plosives with a 10 ms overlapped 20 ms frame (McCandless, 1974). LPC is a widely used method for formant extraction due to its compact and accurate computation. Fig. 6.5 presents the utterance of /k/ consonant manually segmented from the word ”Kannada” from five male speakers of five Kannada dialects.



(a) CENK (b) CSTK (c) HYDK

(d) MUBK (e) STHK

Figure 6.5: Utterance of /k/ manually segmented from word "Kannada" five male speakers of five Kannada dialects, (a) CENK,(b) CSTK,(c) HYDK,(d) MUBK,(e) STHK

Spectral Features: In this study, spectral variations of consonant utterances are measured using standard MFCCs, spectral flux, centroid, rolloff, and entropy. These features try to capture vocal tract variations of different consonants from Kannada dialects. The details of feature extraction are discussed in subsection 3.2.

6.4.2 Experimental Results and Discussion

This section provides few details about the spectral analysis of consonants through features extracted from five Kannada dialects.

From Figure 6.5 it can be observed that there are variations in pronunciation of consonant /k/ across five dialects. The variation in length and in energy of the burst regions can be seen across dialects. Spectral rolloff is generally treated as a spectral shape descriptor of an audio signal and it is usually used for discrimination of voiced and unvoiced sounds. In this study, the histogram has been plotted for the utterance of /k/ from the speakers of five dialects and this is presented in Figure 6.6. From these histograms, it may be noticed that the spectral rolloff dis-

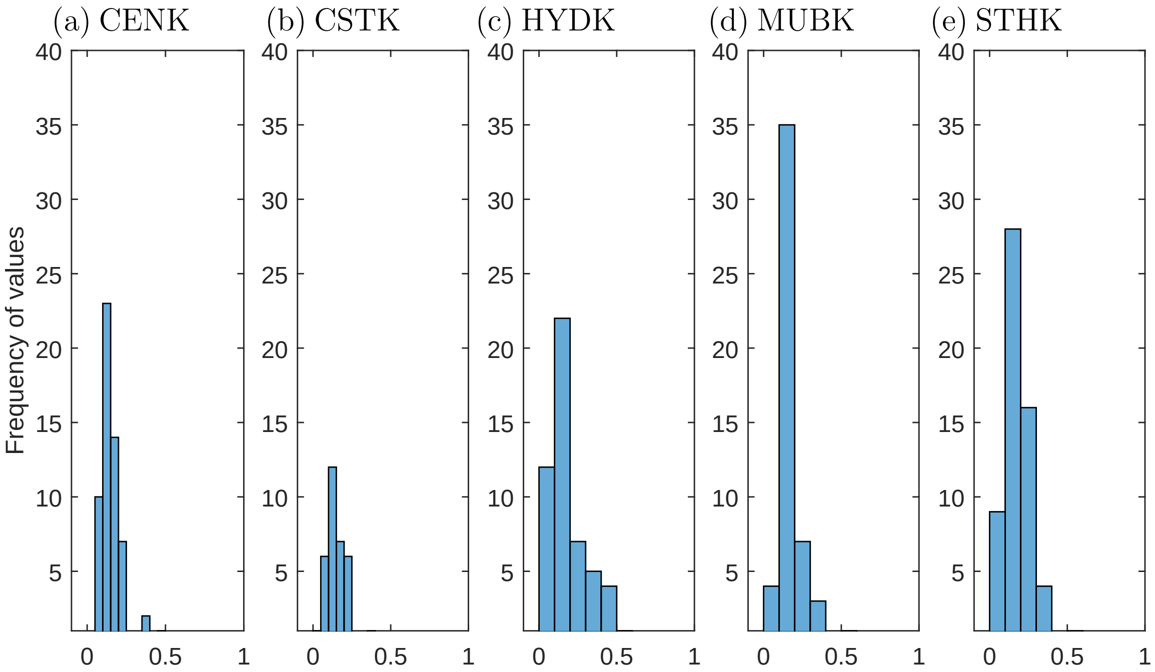


Figure 6.6: Spectral rolloff variations among five Kannada dialects

tributions are comparatively narrow spread in case of CSTK and CENK dialects. Distributions are slightly more with HYDK and STHK dialects.

It may be easily observed that MUBK and STHK dialects have shorter histograms using h higher values of the spectral rolloff values. Besides, the variation is more intense for these two dialects. Whereas, CSTK and CENK dialects are with lower values. The histograms of the mean value of the spectral rolloff sequences of consonants are drawn and are shown in Figure 6.7. It can be seen that the values of spectral flux are higher for the CENK dialect class and are lower for HYDK class. These variations are because of local spectral changes are more frequent in speech signals due to the rapid alternation among phonemes usage and pronunciation. However, local spectral changes are comparatively lower with HYDK dialect. In order to show the differences across dialects with consonant /k/ the features namely, centroid, flux, and two formants (F1 and F2) are considered. The same above mentioned histogram is used. Figure 6.8 is drawn to show the spectral variations across dialects. Formants F1 and F2 are seen to be the distinguishing features among dialects based on /k/. However, the spectral centroid is seen with lesser variations. Even spectral flux feature considered is

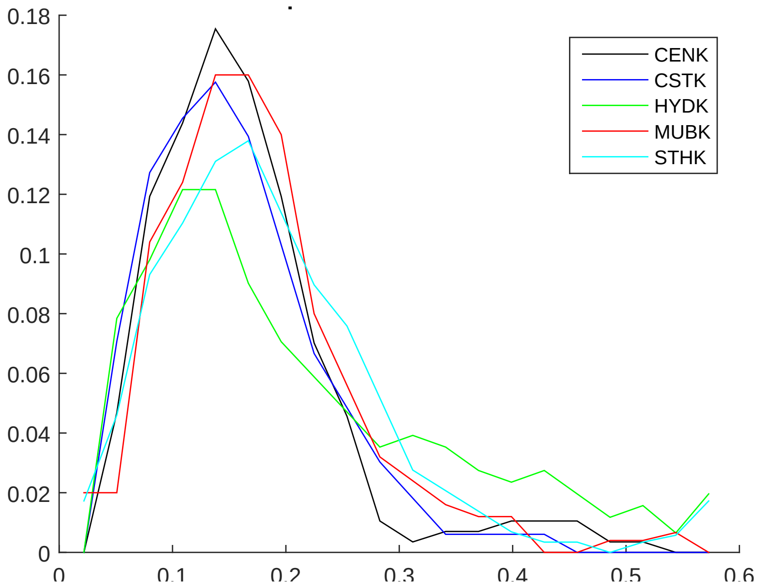


Figure 6.7: Histograms of mean values of sequences of spectral rolloff from audio segments of five Kannada dialects

Table 6.9: Dialect recognition performance from consonants using SVM and XGB classifiers (Accuracies in %)

Sl. No.	Features	Kannada Dialects	
		SVM	XGB
1	MFCCs	73.33	78.00
2	Spectral flux, rolloff, centroid, F1 and F2	59.37	68.76
3	Spectral flux, rolloff, centroid, F1 and F2 + MFCCs	77.65	**78.33**

also contributing with several differences among dialects. Based on analysis of the features as mentioned above in the characterization of dialects, several experiments are conducted by using these features for classification of dialects. Single and ensemble SVM classification methods are employed for the development of ADI systems. Average dialect recognition performance obtained from consonant utterances are presented in Table 6.9.

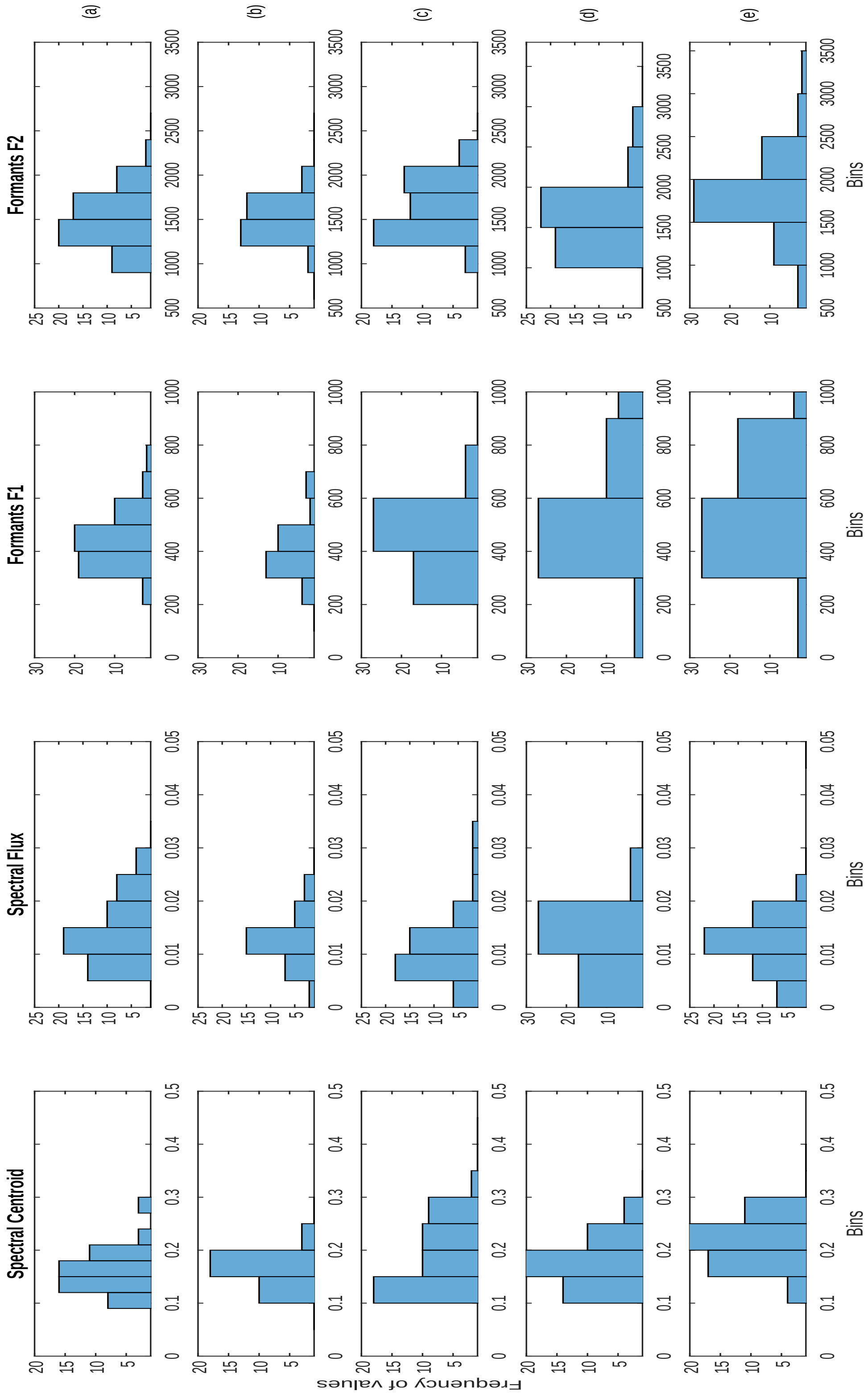


Figure 6.8: Histograms of four spectral features, drawn for the segment /k/ taken from the speakers of five Kannada dialects (a) CENK, (b) CSTK, (c) HYDK, (d) MUBK, (e) STHK

From the results obtained from the Table 6.9, it is observed that MFCCs are seen to be powerful features in classification of dialects from very shorter utterances such as stop consonants. Even spectral features are also captured the dialectal cues from stops. The combination of features has demonstrated an accuracy of 78.33%, which is slightly higher than MFCC features alone. However, from these analyses, it is noticed that dialect specific evidence are present even at consonant level utterances. Spectral attributes can effectively model the dialectal variations of stop consonants.

6.5 *Case* based Dialect Identification System

The Kannada language has complex grammar, and various morphological changes are possible for nouns and verbs when compared to English which has only two variations. For example ”Boy” may change only to ”Boys” in the plural. However, in Kannada, the scenario is significantly different as nouns, and verbs take multiple variations based on *cases* in which they are used. For example, the noun word ”Mara” in Kannada which means a tree in English can be expressed in seven different forms known as *cases* in Kannada. These seven *case* variations are as follows, ”Maravu” (”Prathama”), ”Maravannu” (”Dviteeya”), ”Maradinda” (”Triteeya”), ”Marakke” (”Chaturthi”), ”Marada-deseyinda” (”Panchami”), ”Marada” (”Shashti”) and ”Maradalli” (”Saptami”). After the careful observation and listening to speaking patterns followed in Kannada language, it can be clearly noticed that variations exist in usage patterns of these seven *cases* across dialects. In some dialects, it is noticed that the standard form of usage of *case* form is completely different and replaced with other forms only. Detailed information regarding the seven *cases* used in the Kannada language and their English equivalents and functions are presented in Table 6.10. With this motivation, dialect identification experiments are conducted by using these *case* level utterances. The dataset consists of words carefully chosen with this *case* information embedded in it. These *case* suffixes do not have any independent or unique meaning when they are used alone. Words extracted from the sentences along with case shown to have high-

Table 6.10: Different Kannada *cases* and their English Equivalents used in this study for dialect characterization

Sl. No.	**Kannada**	**Function**	**Prepositions**	**Kannada Suffixes**
1	Prathama Nominative	Subject	-	/u/
2	Dvithiya Accusative	Object	To,	/annu/
3	Trithiya Instrumental	Instrument	By/with/through	/inda/
4	Chaturthi Dative	Receiver	To/for	/ge/, /ke/, /goskar/
5	Panchami Ablative	Point of separation	From/than	/deseyinda/
6	Shashti Genitive	Posession/ Relation	Of/'s	/a/
7	Sapthami Locative	Location	In/on/at/among	/alli/

level dialectal cues; hence it is useful to investigate dialect specific characteristics from these words.

6.5.1 Feature Extraction

In this thesis, chapter 5 had discussed how individual word utterances had shown the existence of dialect-specific features and their comparative performance analysis. *Case*, the spoken units considered in this work also resemble the word groups specifically with Kannada *case* variations. Hence, the same prosodic and spectral features are extracted for analyzing the dialectal differences from *case* information since each word do carry dialect dependent cues.

6.5.2 Experimental Results and Discussion

After analyzing the influence of dynamic and static prosodic features on effective word based dialect classification systems, the same features are used for carrying out experiments on *case* based dataset derived from Kannada dataset. Each such experiments are carried out separately using SVM and XGB method by using

Table 6.11: 'Case'-based dialect recognition performance (Accuracies in %)

Sl. No.	Features	Kannada Word based Dialects		Kannada *Case* based Dialects	
		SVM	XGB	SVM	XGB
1	Prosodic + Statistical Features	41.25	48.86	56.31	64.96
2	MFCCs	72.94	73.40	75.77	80.83
3	Combination of Features	82.31	**83.06**	84.69	**86.73**

dynamic& static, MFCCs, and their combinations. The dialect recognition performances obtained by using these features on *case* dataset are given in Table 6.11. From the results, it is noticed that both dynamic and static combination has shown slightly better dialect recognition performance when compared to word-based ADI systems. Whereas, the dynamic and static feature set has demonstrated an average recognition rate of 56.31% and 64.96% using SVM and XGB methods respectively. Kannada case level utterances used in this work have demonstrated the significant contributions for classification of Kannada dialects. These phonological variants are fairly salient markers in the Kannada language, typically for classification of dialects. Which can be observed with the highest accuracy of about 86.73% has resulted in a combination of features. Even MFCCs features also have performed better on *cases*.

It can be observed from the table that, words of every dialect classes are highly confused with each other classes. However, the words especially with *case* different information have shown the contribution of these on dialect classification. The reason may be, word utterances with *case* information makes each word unique and even they are of short duration, dialect recognition performance is comparatively better over just use of every word for classification of dialects.

6.6 Summary

In this study, characterization, and classification experiments are carried out exclusively for Kannada dialects. For this purpose, Kannada vowels and consonants are considered. Also, a unique grammatical, morphological concept known as *cases* is used for discrimination of Kannada dialects. These vowels, consonants, and *case* level utterances have demonstrated distinct dialect characteristics. Hence, independent studies are carried out concerning the development of ADI systems. Various experiments are conducted using vowel data from five Kannada dialects. This chapter has provided the analysis and contributions of acoustic features in characterizing five Kannada dialects using ten vowels. More importantly, formant frequencies F1, F2, F3, along with F0, energy/loudness, and duration are found to produce better results. In this study, three different decision tree based ensemble algorithms RF, ERF and XGB are used for Kannada dialect classification on Kannada vowel dataset. Through these results, it is observed that ERF has produced better and stable results over RF and XGB algorithms. The similar study is performed with consonants by extracting spectral features namely, MFCCs, flux, centroid, rolloff, and first two formant frequencies. Prosodic features not considered while classifying dialect from stops. SVM and XGB algorithm have performed well and resulted with better performances among other algorithms. At last, a study is carried out to develop ADI systems by using *cases*. These are known to be the suffixes those get added to nouns in Kannada language. However, studies conducted in this work, have examined the apparent differences across dialects in use of the words with this *cases*. This thesis has examined the possible phonetic differences in the way of how vowels and consonants are different and convey distinct dialectal cues with them. In general, vowels are the primary factor producing distinct regional dialects. However, studies carried out have revealed even consonants too responsible for dialects of a language.

CHAPTER 7

Summary and Conclusions

This chapter summarizes the overall research work presented in this thesis. Significant contributions and conclusions drawn out of research work have been highlighted. This chapter also discusses the scope for future improvements.

This thesis has been organized into seven chapters. Chapter 1 provides the introductions to automatic dialect identification system from linguistic and speech perspective. Chapter 2 enlightens critically reviewed existing literature and research activities carried out with respect to dialect identification. In the end, this chapter provides the motivation, scope of the present work derived after the literature review along with the problem formulated for the current thesis work. Chapter 3 discusses the various conventional speech features extracted from speech for identification of dialects. Musical aspects of speech signal extracted in terms of non-conventional chroma-spectral shape features for dialect recognition are discussed in chapter 4. Also, non-conventional dialect-specific features extracted from shorter word and sentence level utterances are presented in chapter 5. Kannada language specific studies, carried for classification of dialects by extracting dialectal cues from vowels, consonants, and *cases* are discussed in chapter 6. Chapter 7 concludes the present thesis work and elucidates on the directions for further research.

7.1 Summary of the Work

Development of dialect identification systems is more useful in enhancing the performance of the speech recognition systems. However, unavailability of the adequately designed dialect datasets for regional languages is the prime reason for lack of dialect recognition systems in a multi-lingual country like India. In this work, a new text-independent spontaneous KDSC is collected from native Kannada speakers. Five dialects namely; CENK, CSTK, HYDK, MUBK, and STHK represent the distinct speaking patterns spoken across Karnataka state. Further, standard IViE speech corpus is used for comparative studies. This thesis work presents the development of ADI systems using different speech features namely, spectral, prosodic, excitation, and spectro-temporal features. It also includes the investigations carried out for the existence of dialectal cues at various levels of spoken units.

Research findings from the literature, have suggested that excitation source and temporal time-varying properties of dialects are not addressed. This work includes extraction of MFCCs, SDCs, spectral flux and entropy features to model vocal tract system. Pitch and energy features are used to capture prosodic attributes. Also, GCIs, regions around GCIs, SOE, SSOE, and instantaneous frequencies are used to capture source level dialectal information. Additionally, spectro-temporal evidences are derived through 2D Gabor features. SVM, and decision tree & SVM based ensemble classification methods are employed for developing dialect recognition systems. Performance comparison, of both categories of classification algorithms, has shown that ensemble algorithms perform better over single classifier based SVM. Investigation of the obtained results, has revealed that four varieties of conventional speech features carry significant and complementary dialect specific cues. The combination has resulted with higher accuracy of about 92.50% on KDSC. It is also observed that, spectral and Gabor features alone have demonstrated better performances over the considered prosodic and source features.

Majority of the dialect recognition systems proposed in the literature have employed conventional speech based spectral and prosodic features. However, several dialects have shown the existence of dialect-specific pitch and energy profile features similar to musical characteristics. Hence, a study is conducted by extracting the music related aspects, in terms of non-conventional features, for classification of dialects. Chroma features are well-established features in the field of music processing are used to capture pitch and energy variations present in a speech along with spectral shape features. Experimental results have shown that the intuition of applying chroma features combination has performed well and has resulted in better performance over traditional MFCC features. The combined feature set has shown sustained improvement with ESVM method on both Kannada and IViE English dialect dataset. However, the proposed feature set has also demonstrated the noise robustness over the MFCCs.

Most of the studies in the literature have only focused on addressing dialect recognition from long utterances, where, length of audio clip may be from duration 5 sec to several minutes. However, many-a-times it is required to recognize dialect from shorter word and sentence units. This study has focused on considering word and sentence level utterances for recognition of dialects. Word and sentence datasets are created by using a dynamic threshold-based segmentation algorithm from KDSC and IViE datasets. To capture dynamic and static prosodic variations across dialects, this study has proposed extraction of intonation and intensity variations from pitch and energy contours. These features are used for assessment of dialectal differences along with MFCC features. A maximum of 85.38% is achieved, with a combination of feature set, using the XGB method from sentence utterances.

Exclusive studies of characterization and classification are carried out on Kannada dialects. Kannada vowels, consonants, and morphological variations known as *cases* are used for discrimination of Kannada dialects. Individually, these units have posed distinct characteristics w.r.t. dialects. Hence, independent studies are conducted on these. Acoustic-phonetic features such as formant frequencies F1, F2, F3, pitch, energy/loudness, and duration have been extracted for vowel specific dialect studies. Explicitly, stop consonant-based dialectal study is carried out by capturing of spectral features through MFCCs, spectral flux, centroid, rolloff, and formant frequencies. Stop consonants are considered to have a higher influence of spectral variations over prosodic features. Apart from this, *case* level utterances resemble word units; hence word based studies are repeated for classification of Kannada dialects. This thesis has examined the possible phonetic differences in the way of how vowels and consonants are different and convey distinct dialectal cues with them. In general, vowels are the primary factor producing distinct regional dialects. However, studies carried out have revealed that variations in consonants too are responsible for dialects of a language. *Cases* are the suffixes added to nouns in the Kannada language. Studies conducted on these have examined the apparent differences across dialects with the use of *cases* information.

7.2 Novel Contributions

A few major contributions made in this thesis are summarized as follows:

- Design and development of new text-independent spontaneous dialect speech dataset with five dialects of Kannada language. This dataset includes more significant speaker variabilities; i.e. sufficient to develop speech recognition systems.
- ADI systems are proposed by extracting conventional spectral, prosodic, and excitation source features to capture dialect-specific information.
- Spectral-temporal features are extracted to capture temporal variations among Kannada dialects through 2D Gabor features and are found to be efficient for Kannada dialect identification.
- Melodic, rhythmic and intonation variations among dialects are captured through chroma features, combined with spectral shape features. ADI systems developed using these features have successfully encapsulated musical aspects and shown better recognition performance. Robustness of the proposed features is demonstrated.
- Statistical post-processing of the explored raw features is performed to get

derived features, where, the mean and standard deviation of the features have resulted in better performance.

- Dialectal cues existing at shorter word and sentence level spoken units, are studied through dynamic and static prosodic features. Word and sentence based ADI systems are developed using proposed features.
- Single classifier based support vector machines and multiple classifiers based ensemble techniques are used for designing the dialect identification system. Comparative performance analysis is carried out.
- This work highlights the studies performed for analysis of the contributions of vowel and consonant characteristics, in classification of dialects. Also, statistical analysis of spectral attributes of vowels is performed using Single Factor-ANOVA (Analysis of Variances) tests.
- The concept of *cases* (Vibhakthi) known as morphological variations of a noun or pronoun are explored for classification of Kannada dialects.

7.3 Conclusions

- Spectral analysis of speech signal through standard MFCC features, has proved to capture useful dialect specific cues. However, the addition of SDC features have shown a reduction in performance indicating less significant with Kannada dialects.
- In this study, Kannada vowels and stop consonants are used for differentiating five dialects. Where, vowels are supposed to carry more dialectal evidence than the stop consonants.
- Chroma and spectral shape feature combination have comparatively performed better and they also depict robustness against the two different noises in variant SNRs.
- Dialect specific evidences in terms of spectro-temporal variations are successfully captured using 2D Gabor features. Improved recognition performance is noticed over MFCC features.
- Integrated ADI systems developed by combining four different features, ex-

tracted from four various aspects of speech signal, have demonstrated higher recognition performance.

- Overall experimental results have shown higher dialect recognition performance on English dialects of IViE speech corpus, than the Kannada dialects of KDSC.
- Post-processed derived features are found to be better over the use of raw features in discrimination of dialects.
- Investigation of the recognition performances obtained throughout the thesis, show that, majority of times ensemble classification algorithms have performed considerably better on KDSC and IViE speech datasets.

7.4 Future Research Directions

- In this work, language independent spectral, prosodic, and excitation source features are extracted to capture dialectal information. Majority of the features proposed in this work are explored from the speech signal. In future, language dependent phonotactic approaches can be examined for enhancing performance.
- Present work has considered ZFF signal for extraction of epoch locations along with other source information. LP residual is used to extract RMFCC features. These features together are used as excitation source information. In future, implicit source features namely; the glottal volume velocity (GVV), raw LP residual samples, its magnitude and phase components at various levels, glottal flow derivative (GFD) parameters for modeling the glottal pulse characteristics, and so on can be used.
- Many times, prosodic features represent the relation between syllables and words, indicating that they are linked to each other. In natural conversation, the duration parameter may convey dialect specific evidence. This usefulness of duration feature of syllable's pauses among words may be explored along with other prosodic features.
- This study has included only stop consonants (plosive sounds) for classifi-

cation of dialects. Apart from these, even fricatives, affricative and nasal sounds can be used to explore dialectal cues.

- Consonant based studies have included only spectral analysis. In future, instead of spectral features, one may consider vocal onset time, burst onset, burst features (duration and amplitude), the complete transition of the consonant to vowel and the entire duration of consonants, etc. In order to perform analysis with respect to these features requires a larger dataset. Hence, in the future, an automatic segmentation algorithm for separating CV units from the continuous speech needs to be designed for the Kannada language.
- Present work has applied the block processing approach through frame level analysis for extraction of speech features. However, spectrogram based analysis can be employed to extract the visual audio features for classification of dialects.
- Integrated dialect recognition systems are proposed by combining spectral, prosodic, excitation source and Gabor features. Further, ADI systems may also be developed by combining dialect-specific features namely, chroma-spectral shape and dynamic and static prosodic features extracted in this work.
- Present work has used the i-vector framework as the state-of-the-art model for comparison of the proposed models. However, there is a scope for using the improved bottleneck features extracted from the i-vector framework for performance comparison.
- In this work, dialect models are developed by a single classifier based nonlinear SVM classification method. Four different multi-classifier based ensemble algorithms are employed those uses decision tree and SVM as the base learners. Proposed classification models performance may be fine-tuned through hyper-parameter tuning. Appropriate selection of optimized hyper-parameters for classifiers ensemble methods can also enhance the dialect recognition performance.

- In future, proper investigations can be carried out by combining dialect-specific evidences explored from vowel, consonants and case for enhancing the Kannada dialect recognition performance.
- Recently, artificial neural network (ANN) and advanced deep learning algorithms are gaining attention. Hence, ADI systems can be developed using, DNN, CNN, RNN, and LSTM frameworks.

Bibliography

Adank, P., Van Hout, R., and Smits, R. (2004). An acoustic description of the vowels of Northern and Southern Standard Dutch. *The Journal of the Acoustical society of America*, 116(3):1729–1738.

Agrawal, S. S., Jain, A., and Sinha, S. (2016). Analysis and modeling of acoustic information for automatic dialect classification. *International Journal of Speech Technology*, 19(3):593–609.

Ali, A., Dehak, N., Cardinal, P., Khurana, S., Yella, S. H., Glass, J., Bell, P., and Renals, S. (2015). Automatic dialect detection in Arabic broadcast speech. *arXiv preprint arXiv:1509.06928*.

Alorifi, F. S. (2008). *Automatic identification of Arabic dialects using hidden Markov models*. PhD thesis, University of Pittsburgh.

Antoni, J. (2006). The spectral kurtosis: a useful tool for characterising non-stationary signals. *Mechanical systems and signal processing*, 20(2):282–307.

Arslan, L. M. and Hansen, J. H. L. (1996). Language accent classification in American English. *Speech Communication*, 18(4):353–367.

Bahari, M. H., Dehak, N., Hamme, H. V., Burget, L., Ali, A. M., and Glass, J. (2014). Non-Negative Factor Analysis of Gaussian Mixture Model Weight Adaptation for Language and Dialect Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(7):1117–1129.

Barkat, M., Ohala, J. J., and Pellegrino, F. (1999). Prosody as a distinctive feature for the discrimination of Arabic dialects. In *EUROSPEECH*, volume 99, pages 395–398.

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

Belinkov, Y. and Glass, J. (2016). A character-level convolutional neural network for distinguishing similar languages and dialects. *arXiv preprint arXiv:1609.07568*.

Benesty, J., Sondhi, M. M., and Huang, Y. (2007). *Springer handbook of speech processing*. Springer.

Biadsy, F. (2011). Automatic Dialect and Accent Recognition and its Application to Speech Recognition. *PhD Thesis, Columbia University.*

Biadsy, F. and Hirschberg, J. (2009). Using prosody and phonotactics in Arabic dialect identification. In *INTERSPEECH*, pages 208–211.

Biadsy, F., Hirschberg, J., and Ellis, D. P. (2011). Dialect and Accent Recognition Using Phonetic-Segmentation Supervectors. In *INTERSPEECH*, pages 745–748.

Biadsy, F., Hirschberg, J., and Habash, N. (2009). Spoken Arabic dialect identification using phonotactic modeling. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL), workshop on computational approaches to semitic languages*, pages 53–61. Association for Computational Linguistics.

Boersma, P., Weenink, D., and Petrus, G. (2002). Praat, a system for doing phonetics by computer. *Glot international*, 5.

Bořil, H., Sangwan, A., and Hansen, J. H. (2012). Arabic Dialect Identification-" Is the Secret in the Silence?" and Other Observations. In *Thirteenth Annual Conference of the International Speech Communication Association.*

Bougrine, S., Cherroun, H., and Ziadi, D. (2017). Hierarchical Classification for Spoken Arabic Dialect Identification using Prosody: Case of Algerian Dialects. *arXiv preprint arXiv:1703.10065.*

Bougrine, S., Cherroun, H., and Ziadi, D. (2018). Prosody-based spoken Algerian Arabic dialect identification. *Procedia Computer Science*, 128:9–17.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brown, G. (2015). Moving towards automatic accent recognition for forensic applications. *Interspeech Doctoral Consortium.*

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

Canavan, A. and Zipperlen, G. (1996). Callfriend American English-non-southern dialect. *Linguistic Data Consortium, Philadelphia*, 10:1.

Chambers, J. K. and Trudgill, P. (1998). *Dialectology.* Cambridge University Press, Second edition.

Chan, M. V., Feng, X., Heinen, J. A., and Niederjohn, R. J. (1994). Classification of speech accents with neural networks. In *Proceedings of International Conference on Neural Networks (ICNN)*, volume 7, pages 4483–4486. IEEE.

Chang, C. C. and Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27.

Chen, N. F., Shen, W., and Campbell, J. P. (2010). A linguistically-informative approach to dialect recognition using dialect-discriminating context-dependent phonetic models. In *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 5014–5017. IEEE.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Twenty second International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

Chen, T., Huang, C., Chang, E., and Wang, J. (2001). Automatic accent identification using Gaussian mixture models. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 343–346.

Chittaragi, N. B. and Koolagudi, S. G. (2017). Acoustic features based word level dialect classification using SVM and ensemble methods. In *Tenth International Conference on Contemporary Computing (IC3)*, pages 1–6. IEEE.

Chittaragi, N. B. and Koolagudi, S. G. (2018). Sentence based Dialect Identification System using Extreme Gradient Boosting Algorithm. In *Sixth International Conference on Advanced Computing, Networking, and Informatics [ICACNI-2018]*, pages 131–138. Springer.

Chittaragi, N. B., Limaye, A., Chandana, N. T., Annappa, B., and Koolagudi, S. G. (2019). Automatic Text-Independent Kannada Dialect Identification System. In *Information Systems Design and Intelligent Applications*, pages 79–87. Springer.

Chittaragi, N. B., Mothukuri, S. K. P., Hegde, P., and Koolagudi, S. G. (2018a). Robust Dialect Identification System using Spectro-Temporal Gabor Features. In *IEEE Region 10 Tencon Conference*, pages 1589–1594. IEEE.

Chittaragi, N. B., Prakash, A., and Koolagudi, S. G. (2018b). Dialect identification using spectral and prosodic features on single and ensemble classifiers. *Arabian Journal for Science and Engineering*, 43(8):4289–4302.

Chitturi, R. and Hansen, J. H. L. (2007). Multi-stream dialect classification using SVM-GMM hybrid classifiers. In *IEEE Workshop on Automatic Speech Recognition Understanding (ASRU)*, pages 431–436.

Choudhury, A. R., Chittaragi, N. B., and Koolagudi, S. G. (2018). Dialect Recognition System Using Excitation Source Features. In *Fifteenth IEEE India Council International Conference (INDICON) (in-press)*.

Clopper, C. G. and Pisoni, D. B. (2006). The nationwide speech project: A new corpus of American English dialects. *Speech Communication*, 48(6):633–644.

Clopper, C. G., Pisoni, D. B., and De Jong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *The Journal of the Acoustical Society of America*, 118(3):1661–1676.

Clopper, C. G. and Smiljanic, R. (2011). Effects of gender and regional dialect on prosodic patterns in American English. *Journal of phonetics*, 39(2):237–245.

Curzan, A. (2013). *How English works: A linguistic introduction*. Pearson Education.

D'Arcy, S. M., Russell, M. J., Browning, S. R., and Tomlinson, M. J. (2004). The accents of the British Isles (ABI) corpus. *Proceedings Modélisations pour l'Identification des Langues*, pages 115–119.

Darwish, K., Sajjad, H., and Mubarak, H. (2014). Verifiably Effective Arabic dialect identification. In *Empirical Methods in Natural Language Processing(EMNLP)*, pages 1465–1468.

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

Dehak, N., Torres Carrasquillo, P. A., Reynolds, D. A., and Dehak, R. (2011b). Language recognition via i-vectors and dimensionality reduction. In *INTERSPEECH*, pages 857–860.

Dietterich, T. G. (2000a). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.

Dietterich, T. G. (2000b). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40(2):139–157.

Du Bois, J. W., Chafe, W. L., Meyer, C., Thompson, S. A., and Martey, N. (2000-2005). Santa Barbara corpus of spoken American English, Parts 1-4.

Ellis, D. P. and Poliner, G. E. (2007). Identifyingcover songs' with chroma features and dynamic programming beat tracking. In *International Conference on Acoustics, Speech and Signal Processing-ICASSP*, pages 1429–1432. IEEE.

Etman, A. and Beex, A. L. (2015). Language and dialect identification: A survey. In *SAI Intelligent Systems Conference (IntelliSys)*, pages 220–231. IEEE.

Etman, A. and Louis, A. A. (2015). American dialect identification using phonotactic and prosodic features. In *SAI Intelligent Systems Conference (IntelliSys)*, pages 963–970. IEEE.

Ferragne, E. and Pellegrino, F. (2007). Automatic dialect identification: A study of British English. In *Speaker classification II*, pages 243–257.

Flanagan, J. L. (2013). *Speech analysis synthesis and perception*, volume 3. Springer Science & Business Media.

Fogerty, D. and Humes, L. E. (2012). The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences. *The Journal of the Acoustical Society of America*, 131(2):1490–1501.

Freund, Y. and Schapire, R. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(1612):771–780.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., and Dahlgren, N. (1993). {DARPA TIMIT} Acoustic-phonetic continuous speech corpus.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63(1):3–42.

Giannakopoulos, T. and Pikrakis, A. (2014). *Introduction to Audio Analysis: A MATLAB Approach*. Academic Press.

Grabe, E. and Post, B. (2002). Intonational variation in the British Isles. In *Speech Prosody*, pages 127–132.

Gray, S. and Hansen, J. H. L. (2005). An integrated approach to the detection and classification of accents/dialects for a spoken document retrieval system. In *Workshop on Automatic Speech Recognition and Understanding*, pages 35–40. IEEE.

Han, B. J., Rho, S., Jun, S., and Hwang, E. (2010). Music emotion classification and context-based music recommendation. *Multimedia Tools and Applications*, 47(3):433–460.

Hansen, J. H. L. and Liu, G. (2016). Unsupervised accent classification for deep data fusion of accent and language information. *Speech Communication*, 78:19–33.

Harris, M. J., Gries, S. T., and Miglio, V. G. (2014). Prosody and its application to forensic linguistics. *LESLI: Linguistic Evidence in Security Law and Intelligence*, 2(2):11–29.

Hassani, H. and Hamid, O. H. (2017). Using Artificial Neural Networks in Dialect Identification in Less-resourced Languages-The Case of Kurdish Dialects Identification. In *IJCCI*, pages 443–448.

Hermansky, H. and Morgan, N. (1994). RASTA processing of speech. *IEEE transactions on speech and audio processing*, 2(4):578–589.

Hillenbrand, J. M., Clark, M. J., and Nearey, T. M. (2001). Effects of consonant environment on vowel formant patterns. *The Journal of the Acoustical Society of America*, 109(2):748–763.

Huang, R. and Hansen, J. H. L. (2007). Unsupervised discriminative training with application to dialect classification. *IEEE transactions on Audio, Speech, and Language processing*, 15(8):2444–2453.

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

Jankowski, C., Kalyanswamy, A., Basson, S., and Spitz, J. (1990). NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database. In *Acoustics, Speech, and Signal Processing. ICASSP*, pages 109–112.

Jebara, T. (2012). *Machine learning: discriminative and generative*, volume 755. Springer Science & Business Media.

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

John Godfrey and Edward Holliman (1993). Switchboard-1 Release 2 LDC97S62.

Johnson, K. (2008). 15 speaker normalization in speech perception. *The handbook of speech perception*, page 363.

Kalaiah, M. K. and Bhat, J. S. (2017). Effect of Vowel Context on the Recognition of Initial Consonants in Kannada. *Journal of audiology & otology*, 21(3):146.

Karjigi, V. and Rao, P. (2013). Knowledge-based features for place classification of unvoiced stops. *Journal of Intelligent Systems*, 22(3):215–228.

Khurana, S., Najafian, M., Ali, A., Al Hanai, T., Belinkov, Y., and Glass, J. (2017). QMDIS: QCRI-MIT advanced dialect identification system. In *INTERSPEECH*, pages 2591–2595.

Kim, H. C., Pang, S., Je, H. M., Kim, D., and Bang, S. Y. (2002). Support vector machine ensemble with bagging. In *First International Workshop on Pattern Recognition with Support Vector Machines*, pages 397–408.

Kodukula, S. (2009). Significance of excitation source information for speech analysis. *PhD Thesis, Dept. of Computer Science, IIT, Madras.*

Konnerth, L., Morey, S., Sarmah, P., and Teo, A., editors (2015). *North East Indian Linguistics (NEIL)*, volume 7. Asia-Pacific Linguistics.

Koolagudi, S. G. and Rao, K. S. (2012). Emotion recognition from speech: a review. *International journal of speech technology*, 15(2):99–117.

Kotnik, B., Vlaj, D., Kacic, Z., and Horvat, B. (2002). Robust MFCC feature extraction algorithm using efficient additive and convolutional noise reduction procedures. In *International Conference on Speech and Language Processing (ICSLP)*, volume 2, pages 445–448.

Kotsiantis, S. B., Zaharakis, I., and Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24.

Kulshreshtha, M., Singh, C., and Sharma, R. (2012). Speaker profiling: The study of acoustic characteristics based on phonetic features of Hindi dialects for forensic speaker identification. In *Forensic Speaker Recognition*, pages 71–100. Springer.

Lachachi, N. E. and Adla, A. (2016). Two approaches-based L2-SVMs reduced to MEB problems for dialect identification. *International Journal of Computational Vision and Robotics*, 6(1-2):1–18.

Lei, H., Meyer, B. T., and Mirghafori, N. (2012). Spectro-temporal gabor features for speaker recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4241–4244. IEEE.

Lei, Y. and Hansen, J. H. L. (2011). Dialect classification via text-independent training and testing for Arabic, Spanish, and Chinese. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):85–96.

Lessmann, S., Baesens, B., Seow, H. V., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136.

Li, H., Ma, B., and Lee, K. A. (2013). Spoken language recognition: from fundamentals to practice. *Proceedings of the IEEE*, 101(5):1136–1159.

Lim, B. P., Li, H., and Ma, B. (2005). Using local & global phonotactic features in Chinese dialect identification. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 577–580. IEEE.

Liu, G., Lei, Y., and Hansen, J. H. L. (2010). Dialect identification: Impact of differences between read versus spontaneous speech. In *Eighteenth European Signal Processing Conference*, pages 2003–2006. IEEE.

Liu, G. A. and Hansen, J. H. L. (2011). A systematic strategy for robust automatic dialect identification. In *Nineteenth European Signal Processing Conference*, pages 2138–2141.

Ma, B., Zhu, D., and Tong, R. (2006). Chinese dialect identification using tone features based on pitch flux. In *International Conference on Acoustics Speech and Signal Processing Proceedings (ICASSP)*, volume 1, pages 1029–1032.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580.

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

Marc, C., Frank, D. S., Johan, S., and Bart, D. M. (2014). EnsembleSVM: A library for ensemble learning using support vector machines. *Journal of Machine Learning Research*, 15:141–145.

Mary, L. and Yegnanarayana, B. (2008). Extraction and representation of prosodic features for language and speaker recognition. *Speech Communication*, 50(10):782–796.

McCandless, S. (1974). An algorithm for automatic formant extraction using linear prediction spectra. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 22(2):135–141.

Mehrabani, M. and Hansen, J. H. L. (2015). Automatic analysis of dialect/language sets. *International Journal of Speech Technology*, 18(3):277–286.

Meyer, B. T., Ravuri, S. V., Schädler, M. R., and Morgan, N. (2011). Comparing Different Flavors of Spectro-Temporal Features for ASR. In *Twelfth Annual Conference of the International Speech Communication Association*, pages 1269–1272.

Meyer, Bernd T and Kollmeier, Birger (2011). Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition. *Speech Communication*, 53(5):753–767.

Müller, M. and Ewert, S. (2011). Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In *Proceedings of the Twelfth International Conference on Music Information Retrieval (ISMIR).*

Murty, K Sri Rama and Yegnanarayana, B (2008). Epoch extraction from speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8):1602–1613.

Nagesha, K. S. and Nagabhushana, B. (2007). Acoustic-phonetic analysis of Kannada accents. In *Proceedings of Frontiers of Research on Speech and Music Signal Processing, AIISH*, pages 222–225.

Najafian, M., DeMarco, A., Cox, S., and Russell, M. (2014). Unsupervised model selection for recognition of regional accented speech. In *Fifteenth Annual Conference of the International Speech Communication Association.*

Nakamura, M., Iwano, K., and Furui, S. (2008). Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech & Language*, 22(2):171–184.

Nandi, D., Pati, D., and Rao, K. S. (2015). Implicit excitation source features for robust language identification. *International Journal of Speech Technology*, 18(3):459–477.

Nandi, D., Pati, D., and Rao, K. S. (2017). Parametric representation of excitation source information for language identification. *Computer Speech & Language*, 41:88–115.

P. Price, W. Fisher, Jared Bernstein, and D. Pallett (1993). Resource Management RM1 2.0.

Pedersen, C. and Diederich, J. (2007). Accent classification using support vector machines. In *Sixth IEEE/ACIS on Computer and Information Science*, pages 444–449.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vandeeplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Edouard, D. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct):2825–2830.

Prahallad, K., Kumar, E. N., Keri, V., Rajendran, S., and Black, A. W. (2012). The IIIT-H Indic Speech Databases. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Prasanna, S. M., Gupta, C. S., and Yegnanarayana, B. (2006). Extraction of speaker-specific excitation information from linear prediction residual of speech. *Speech Communication*, 48(10):1243–1261.

Prasanna, S. M., Reddy, B. S., and Krishnamoorthy, P. (2009). Vowel onset point detection using source, spectral peaks, and modulation spectrum energies. *IEEE Transactions on audio, speech, and language processing*, 17(4):556–565.

Purnell, J. T. and Magdon-Ismail, M. (2009). Learning American English accents using ensemble learning with GMMs. In *International Conference on Machine Learning and Applications (ICMLA)*, pages 47–52. IEEE.

Purnell, T., Idsardi, W., and Baugh, J. (1999). Perceptual and phonetic experiments on American English dialect identification. *Journal of language and social psychology*, 18(1):10–30.

Rabiner, L. R. and Juang, B. H. (1993). Fundamentals of Speech Recognition.

Rajapurohit, B. B. (1982). *Acoustic characteristics of Kannada*. Central Institute of Indian Languages.

Ramus, F. and Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. *The Journal of the Acoustical Society of America*, 105(1):512–521.

Ranjan, R. and Dubey, R. K. (2016). Isolated Word Recognition using HMM for Maithili dialect. In *International Conference on Signal Processing and Communication (ICSC)*, pages 323–327.

Rao, K. S. and Koolagudi, S. G. (2011). Identification of Hindi dialects and emotions using spectral and prosodic features of speech. *International Journal of Systemics, Cybernetics and Informatics*, 9(4):24–33.

Rao, K. S. and Nandi, D. (2015). *Language Identification Using Excitation Source Features*. Springer.

Reddy, V. R., Maity, S., and Rao, K. S. (2013). Identification of Indian languages using multi-level spectral and prosodic features. *International Journal of Speech Technology*, 16(4):489–511.

Reetz, H. and Jongman, A. (2009). *Phonetics Transcription, Production, Aoustics and Perception*. Wiley Blackwell.

Rizwan, M. and Anderson, D. V. (2018). A weighted accent classification using multiple words. *Neurocomputing*, 277:120–128.

Rouas, J. L. (2007). Automatic prosodic variations modeling for language and dialect discrimination. *IEEE Transactions on Audio, Speech and Language Processing*, 15(6):1904–1911.

Sadjadi, S. O., Slaney, M., and Heck, L. (2013). Msr identity toolbox v1. 0: A matlab toolbox for speaker-recognition research. *Speech and Language Processing Technical Committee Newsletter*, 1(4):1–32.

Sarma, M. and Sarma, K. K. (2016). Dialect Identification from Assamese speech using prosodic features and a neuro fuzzy classifier. In *Third International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 127–132. IEEE.

Schädler, M. R., Meyer, B. T., and Kollmeier, B. (2012). Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition. *The Journal of the Acoustical Society of America*, 131(5):4134–4151.

Schröder, J., Goetze, S., and Anemüller, J. (2015). Spectro-temporal Gabor filterbank features for acoustic event detection. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(12):2198–2208.

Shon, S., Ali, A., and Glass, J. (2018). Convolutional neural networks and language embeddings for end-to-end dialect recognition. *arXiv preprint arXiv:1803.04567*.

Shridhara, M., Banahatti, B. K., Narthan, L., Karjigi, V., and Kumaraswamy, R. (2013). Development of Kannada speech corpus for prosodically guided phonetic search engine. In *International Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 1–6. IEEE.

Siddhant, A., Jyothi, P., and Ganapathy, S. (2017). Leveraging native language speech for accent identification using deep Siamese networks. In *Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 621–628. IEEE.

Sinha, S., Agrawal, S. S., and Jain, A. (2015a). Influence of regional dialects on acoustic characteristics of Hindi vowels. In *International Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 166–171.

Sinha, S., Jain, A., and Agrawal, S. S. (2014). Speech Processing for Hindi Dialect Recognition. In *Advances in Signal Processing and Intelligent Recognition Systems*, pages 161–169.

Sinha, S., Jain, A., and Agrawal, S. S. (2015b). Acoustic-phonetic feature based dialect identification in Hindi Speech. *International Journal on Smart Sensing & Intelligent Systems*, 8(1).

Sinha, S., Jain, A., and Agrawal, S. S. (2017). Empirical analysis of linguistic and paralinguistic information for automatic dialect classification. *Artificial Intelligence Review*, pages 1–26.

Snyder, D., Garcia-Romero, D., Povey, D., and Khudanpur, S. (2017). Deep neural network embeddings for text-independent speaker verification. In *Proc. INTERSPEECH*, pages 999–1003.

Soorajkumar, R., Girish, G., Ramteke, P. B., Joshi, S. S., and Koolagudi, S. G. (2017). Text-Independent Automatic Accent Identification System for Kannada Language. In *Proceedings of the International Conference on Data Engineering and Communication Technology*, pages 411–418. Springer.

Sun, X. (2000). A pitch determination algorithm based on subharmonic-to-harmonic ratio. In *The Sixth International Conference of Spoken Language Processing*, pages 676–679.

Themistocleous, C. (2016). The bursts of stops can convey dialectal information. *The Journal of the Acoustical Society of America*, 140(4):EL334–EL339.

Toohill, B. J., Mcleod, S., and Mccormack, J. (2012). Effect of dialect on identification and severity of speech impairment in indigenous australian children. *Clinical Linguistics & Phonetics*, 26(2):101–119.

Torres-Carrasquillo, P. A., Gleason, T. P., and Reynolds, D. A. (2004). Dialect identification using gaussian mixture models. In *ODYSSEY 2004-The Speaker and Language Recognition Workshop*, pages 297–300.

Trousdale, G. (2010). *Introduction to English Sociolinguistics*. Edinburgh University Press.

Tsai, W. H. and Chang, W. W. (1999). Chinese dialect identification using an acoustic-phonotactic model. In *Sixth European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 367–370.

Tsai, Wuei He and Chang, Wen Whei (2002). Discriminative training of Gaussian mixture bigram models with application to Chinese dialect identification. *Speech Communication*, 36(3):317–326.

Tzudir, M., Sarmah, P., and Prasanna, S. R. M. (2018). Dialect Identification Using Tonal and Spectral Features in Two Dialects of Ao.

Utami, Iut Tri, Sartono, Bagus, and Sadik, Kusman (2014). Comparison of single and ensemble classifiers of support vector machine and classification tree. *Journal of Mathematical Sciences and Applications*, 2(2):17–20.

Wakefield, G. H. (1999). Mathematical representation of joint time-chroma distributions. In *Advanced Signal Processing Algorithms, Architectures, and Implementations IX*, volume 3807, pages 637–646. International Society for Optics and Photonics.

Wightman, C. W. (1992). Automatic detection of prosodic constituents for parsing. *Doctoral dissertation.*

Wray, S. and Ali, A. (2015). Crowdsource a little to label a lot: Labeling a speech corpus of dialectal Arabic. In *Sixteenth Annual Conference of the International Speech Communication Association*, pages 2824–2828.

Xu, F., Wang, M., and Li, M. (2016). Sentence-level dialects identification in the Greater China region. *International Journal on Natural Language Computing (IJNLC)*, 5(6):9–20.

Yang, X., Audhkhasi, K., Rosenberg, A., Thomas, S., Ramabhadran, B., and Hasegawa-Johnson, M. (2018). Joint modeling of accents and acoustics for multi-accent speech recognition. *arXiv preprint arXiv:1802.02656.*

Yanguas, L. R. and Quatieri, T. F. (1999). Implications of glottal source for speaker and dialect identification. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 813–816. IEEE.

Yegnanarayana, B. and Murty, K. S. R. (2009). Event-based instantaneous fundamental frequency estimation from speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4):614–624.

Zaidan, O. F. and Callison-Burch, C. (2014). Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

Zhang, Q. and Hansen, J. H. L. (2018). Language/dialect recognition based on unsupervised deep learning. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 26(5):873–882.

Zheng, D. C., Dyke, D., Berryman, F., and Morgan, C. (2012). A new approach to acoustic analysis of two British regional accents—Birmingham and Liverpool accents. *International Journal of Speech Technology*, 15(2):77–85.

Zhenhao, G. (2015). Improved accent classification combining phonetic vowels with acoustic features. In *Eighth International Congress on Image and Signal Processing (CISP)*, pages 1204–1209.

Zissman, M. A. and Berkling, K. M. (2001). Automatic language identification. *Speech Communication*, 35(1-2):115–124.

Zissman, M. A., Gleason, T. P., Rekart, D., and Losiewicz, B. L. (1996). Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech. In *Acoustics, Speech, and Signal Processing, ICASSP*, volume 2, pages 777–780.

List of Publications

Journal Publications

1. Chittaragi, N. B., Prakash, A., and Koolagudi, S. G. (2018). ”Dialect Identification Using Spectral and Prosodic Features on Single and Ensemble Classifiers”, Arabian Journal for Science and Engineering, 43(8): 4289-4302. (SCI Indexed)

2. Chittaragi, N. B. and Koolagudi, S. G. (2020) Automatic dialect identification system for Kannada language using single and ensemble SVM algorithms. Language Resources & Evaluation 54, 553–585 (SCI Indexed)

3. Chittaragi, N. B. and Koolagudi, S. G. (2019) Acoustic-phonetic feature based Kannada dialect identification from vowel sounds. International Journal of Speech Technology 22, 1099–1113 (2019) (Scopus Indexed)

4. Chittaragi, N. B. and Koolagudi, S. G. Text Independent Dialect Identification using Chroma-Spectral Shape Features with Ensemble Technique. Computer Speech and Language Communication (Under Review). (SCI Indexed)

Conference Publications

1. Chittaragi, N. B. and Koolagudi, S. G. (2017). Acoustic features based word level dialect classification using SVM and ensemble methods. Tenth International Conference on Contemporary Computing (IC3), pages 1-6 . IEEE.

2. Chittaragi, N. B. and Koolagudi, S. G. (2018). Sentence based Dialect Identification System using Extreme Gradient Boosting Algorithm. In Sixth International Conference on Advanced Computing, Networking, and Informatics [ICACNI-2018], pages 131-138, Springer.

3. Chittaragi, N. B., Limaye, A., Chandana, N. T., Annappa, B., and Koolagudi, S. G. (2019). Automatic text-independent Kannada dialect identification system. In Information Systems Design and Intelligent Applications, pages 79-87.

4. Chittaragi, N. B., Mothukuri, S. K. P., Hegde, P., and Koolagudi, S. G. (2018). Robust Dialect Identification System using Spectro-Temporal Gabor Features. In IEEE Region 10 Tencon Conference, pages 1589-1594. IEEE.

5.Choudhury, A. R., Chittaragi, N. B., and Koolagudi, S. G. (2018). Dialect Recognition System Using Excitation Source Features, Fifteenth IEEE India Council International Conference (INDICON), pages. 1-6. IEEE.

6. Chittaragi, N. B. and Koolagudi, S. G. (2019). Spectral Feature based Kannada Dialect Classification from Stop Consonants, Eighth International Conference on Pattern Recognition and Machine Intelligence (PReMI 2019), pages 82-90.

BRIEF BIO-DATA

Personal Details

Name - Nagaratna B. Chittaragi
Date of Birth - 20/07/1982

Work Address

Nagaratna B. Chittaragi
Asst. Professor, Dept. of ISE.
Siddaganga Institute of Technology
B. H. Road,Tumkur 572103
Email: nbchittaragi@gmail.com, chittaragi@sit.ac.in

Permanent Address

Nagaratna B. Chittaragi
Shree Dhaneshwari, 1st Cross,
Shivaganga Nagar, Batawadi
Tumkur, 572103 Karnataka

Qualification

M. Tech. Computer Science & Engineering, SIT, Tumkur, 2010.
B. E. Information Science & Engineering, VTU, Belagaum Karnataka, 2005.

Current Employment

Lecturer, Department of ISE. SIT, Tumkur, Karnataka (2005 July - 31/12/2010)
Assistant Professor, Department of ISE. SIT, Tumkur, Karnataka (01/01/2011 - Till date)