




Automatic dialect identification system for Kannada language using single and ensemble SVM algorithms

Nagaratna B. Chittaragi^{1,2}  ·
Shashidhar G. Koolagudi¹

Published online: 21 November 2019
© Springer Nature B.V. 2019

Abstract In this paper, an automatic dialect identification (ADI) system is proposed by extracting spectral and prosodic features for Kannada language. A new dialect dataset is collected from native speakers of Kannada language (A Dravidian language). This dataset includes five distinct dialects of Kannada language representing five geographical regions of Karnataka state. Investigation of the significance of spectral and prosodic variations on five Kannada dialects is carried out. Mel-frequency cepstral coefficients (MFCCs), spectral flux, and entropy are used as representatives of spectral features. Besides, pitch and energy features are extracted as representatives of prosodic parameters for identification of dialects. These raw feature vectors are further processed to get a new derived feature vectors by using statistical processing. In this paper, a single classifier based multi-class support vector machine (SVM) and multiple classifier based ensemble SVM (ESVM) techniques are employed for classification of dialects. The effectiveness and performance evaluation of the explored features are carried out on newly collected Kannada speech corpus, with five Kannada dialects and internationally known standard Intonation Variation in English (IViE) dataset with nine British English dialects. Experimental results have demonstrated that the derived feature vectors performs better when compared to raw feature vectors. However, ESVM technique has demonstrated better performance over a single SVM. Spectral and prosodic features have resulted individually with the dialect recognition performance of

✉ Nagaratna B. Chittaragi
nbchittaragi@gmail.com
Shashidhar G. Koolagudi
koolagudi@nitk.edu.in

¹ Dept. of Computer Science and Engg., National Institute of Technology Karnataka, Surathkal, India

² Dept. of Information Science and Engg., Siddaganga Institute of Technology, Tumkur, Karnataka, India

83.12% and 44.52% respectively. Further, the complementary nature of both spectral and prosodic features is evaluated by combining both feature vectors for dialect recognition. However, an increase in dialect recognition performance of about 86.25% is observed. This indicates the existence of complementary dialect specific evidence with spectral and prosodic features. The experiments conducted on standard IViE corpus have shown a higher recognition rate of 91.38% using ESVM. Proposed ADI systems with derived features have shown better performance over the state-of-the-art i-vector feature based systems on both datasets.

Keywords Dialect identification · Kannada dialect dataset · IViE dialect dataset · Spectral and prosodic features · Derived features · Single SVM · Ensemble SVM

1 Introduction

Dialect identification from speech has made great strides in the recent past with the development of speech based interactive systems. Dialect represents a unique pronunciation pattern followed among a group of native people belonging to a specific geographic region. Dialectal variations in any language exist mainly due to several surrounding factors pertaining to the speaker such as, geographical location, socioeconomic status, mother tongue (L1), the influence of neighboring state languages, cultural and education background Clopper and Smiljanic (2011), Chambers and Trudgill (1988). Next to gender information, dialectal differences are the primary factors responsible for speech variabilities that cause heavy degradation in the performance of automatic speech processing systems.

Dialect identification task primarily deals with recognition of dialect within a predetermined language from the spoken speech. Dialect identification from speech can be useful in modeling several subsystems of Automatic Speech Recognition (ASR) systems. Subsystems such as pronunciation modeling, acoustic, phonetic model training and lexicons adaptation can be aware of variations of dialectal pronunciations prior to speech recognition Ferragne and Pellegrino (2007). Dialect classification¹ is useful in indexing of historical speech corpora and spoken document retrieval systems. ADI systems are beneficial in real-life applications such as nativity identification, efficient automatic call interpretation systems, tele-medicine, e-learning, entertainment etc. ADI systems can be useful in call centers for an effective region based customer call attention. Extensive research findings are available in the literature for speaker profiling, speaker recognition and verification techniques that are applied in forensic applications. However, they are not supported by suitable dialect processing module Harris et al. (2014), Ahuja and Vyas (2018). Dialect information is extremely useful in speaker profiling task which is commonly used in linguistic forensic applications. Here, speaker profiling mainly deals with capturing linguistic and paralinguistic cues from the unknown speakers. Information such as age, gender, language, dialect, emotional state, ethnicity, geographical, and

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

socio-economic status of the speaker are drawn. Due to vast and wide variety of applications, now-a-days ADI systems are gaining more attention from the speech research community.

In general, characterization and identification of dialects of any language requires a better understanding of linguistic properties Clopper and Smiljanic (2011). For any language, dialectal differences may arise from all levels of its linguistic hierarchy (e.g., acoustic, phonetic, vocabulary, syntax, prosody) Mehrabani and Hansen (2015), Hansen and Liu (2016). Among all levels, dialect variations may be prominently observed in phonetics, acoustics, prosodic level and with a little deviation in use of standard vocabulary across dialects. Hence, for efficient discrimination of dialects, it is necessary to identify the peculiar acoustic, phonetic and prosodic variations. Among all, a few variations play a significant role in the characterization and identification of dialects namely: (1) Use of different vowels and consonant pronunciation patterns, (2) Specific phonetic rules indicating substitution, elimination, and addition, (3) Special morphological operations pertaining to different languages Hansen and Liu (2016).

Well established advanced research activities are available for several widely spoken languages across the world, such as English, Chinese, Arabic, Spanish, Hindi and French Etman and Beex (2015), Zhang and Hansen (2018). In the context of Indian languages, dialect relevant studies are found only in few languages such as Hindi, Bengali, Assami, Kannada, Telugu, Punjabi, Haryanvi and Marathi Rao and Koolagudi (2011), Sarma and Sarma (2016), Soorajkumar et al. (2017), Malmasi and Dras (2015), Mannepalli et al. (2016). However, significant and systematic work being carried out with many Indian regional languages is still lacking. Probably this may be due to the following challenges: (1) The existence of several official and unofficial regional languages across the country, (2) Difficulty in identifying clear boundaries between dialects of individual languages, (3) Every language uses a common phoneme set and follows a similar grammatical structure, and (4) Unavailability of standard speech corpora. Apart from these, research findings on many Indian, specifically Dravidian languages spoken in southern India are in its nascent stage when compared to other languages. In particular, Kannada language has got decidedly less substantial work pertaining to dialects.

First and foremost, the acoustic and phonetic analysis of dialectal variations requires an appropriate corpus. Development of dialect speech corpora for many under-resourced regional languages in a country like India is of utmost importance. This is due to the fact that India is associated with unique multicultural and multilingual diversity Chandrasekaran (2012). Specific to Dravidian languages, several dialectal variations exist in each of the four Dravidian languages (Kannada, Telugu, Malayalam, Tamil). Specifically, Kannada language spoken in Karnataka state has also demonstrated many dialects representing pronunciation variations spoken across the state. These variations occur mainly due to regional bias, unique style/pattern, and diversity associated with each geographical region. A significant performance degradation can be observed with ASR systems of Dravidian languages due to these variations. Hence, due to the lack of systematic works existence, there is a need for developing a dialect recognition system for Kannada language in order to build an efficient ASR systems.

From the literature, it is observed that, there is no publicly available standard dialect dataset for Kannada language. Therefore, in this paper, a new speech corpus is collected for Kannada language including five dialects representing distinct varieties of speaking styles. Here, in this paper, spectral and prosodic features have been studied for the development of the ADI system. Spectral analysis of speech is performed to capture dynamics of the vocal tract system regarding MFCCs, spectral flux, and entropy features. Dialect-specific prosodic features namely, pitch and energy are also extracted. Two different dialect speech datasets namely, proposed Kannada and internationally known standard IViE dialect corpora are used for evaluation of the captured features. Performances are analyzed and compared with individual and combined feature vectors since both carry non-overlapping dialect-specific features. Selection of an effective classification method is also an essential task while designing of ADI system. Hence, two different algorithms such as single classifier based SVM and ESVM classifiers are used to develop the dialect identification system. Further, performance comparison and evaluation of raw features and derived feature vectors is carried out. Significant dialectal differences existing among female and male speakers are also investigated on five Kannada dialects.

Rest of the paper is organized with the following sections: Sect. 2 gives a brief review of the existing literature on dialect identification. Detailed discussion of steps involved in collecting Kannada dialect speech corpus is presented in Sect. 3 along with a brief introduction of IViE speech corpus. Section 4 briefly discusses the proposed ADI systems along with spectral, prosodic and i-vector feature extraction process and details of two classification methods. Section 5 provides the details of experimental setup, results, and performance analysis. Section 6 gives the conclusions of the present work with a brief summary along with future research directions.

2 Literature review

This section provides a concise review of the existing literature on dialect processing. It covers details of existing works available on dialect processing with respect to acoustic-phonetic features, phonotactic models, dialect datasets, and classification techniques.

Dialect related research findings available in existing literature are concentrated towards the widely spoken languages across the world. Many studies are available in the literature with various dialects of English. American, British and Australian English are three prominent dialects of English. Also, there exist extensive research activities with many more regional dialects within these Zhang and Hansen (2018), Chittaragi et al. (2018), Hanani et al. (2013). Significant number of studies are reported on dialects of Arabic language, as it has several dialects and spoken by a large population Zhang and Hansen (2018), Biadisy (2011), Lei and Hansen (2011), Bougrine et al. (2017). Similarly, Chinese and its various dialect related studies using different tonal features, are reported in the literature Zhang and Hansen (2018), Lei and Hansen (2011), Lim et al. (2005), Ma et al. (2006). It is observed

that, some major dialect identification works are also being conducted on dialects of Spanish, French and Japanese languages Lei and Hansen (2011), Zissman et al. (1996).

Majority of existing dialect processing systems are addressed through acoustic-phonetic, and phonotactic modeling approaches Rao and Koolagudi (2011), Chen et al. (2010). Acoustic-phonetic modeling explores dialectal cues extracted from segmental, supra-segmental, and sub-segmental levels Rao and Koolagudi (2011). Spectral acoustic differences existing among dialects are studied extensively in literature by extracting MFCCs through cepstral analysis of utterances. Also, Shifted Delta Coefficients (SDC) are used along with MFCCs to capture temporal variations. Spectral features such as formants (F1–F4), spectral flux, centroid, and entropy that are extracted from the spectra obtained after spectral analysis are also applied for dialect classification Rao and Koolagudi (2011), Chittaragi et al. (2018), Chittaragi and Koolagudi (2017).

Prosodic or supra-segmental features are said to carry dialect specific cues since they impose naturalness to speech. Various studies have addressed dialect recognition with the use of pitch, intonation patterns, intensity, rhythm, and many more features. Also, these are extracted from various levels of utterances such as words, syllables, pseudo-syllables, and sentences Chittaragi and Koolagudi (2017), Rouas (2007), Biadisy and Hirschberg (2009), Chen et al. (2014), Chittaragi and Koolagudi (2018). Some works are also reported on the use of i-vector features to extract acoustic features using joint factor analysis with dimensionality reduction Hansen and Liu (2016), Dehak et al. (2011), Behravan et al. (2015). However, a slight increase in dialect recognition performance is reported with i-vectors when compared to traditional spectral and prosodic features.

Gaussian mixture models are widely used for modeling dialect identification systems using spectral and prosodic features Soorajkumar et al. (2017), Chen et al. (2001), Torres-carrasquillo et al. (2004). Existing studies have shown better dialect identification performance using SVM. However, very few have employed SVM for dialect identification Chittaragi and Koolagudi (2017), Pedersen and Diederich (2007). A familiar Universal Background Model (UBM) proposed for speaker identification is applied for dialect classification, which has shown better performance Ziedan et al. (2016). A study of modified version of UBM is reported for classification of dialects to address a smaller dataset with noisy speech and has resulted in better performance Liu and Hansen (2011).

In the literature, most of the time single classifier based classification methods are employed for dialect identification. However, recently the concept of combining multiple classifiers using ensemble algorithms, is gaining more popularity over traditional methods. However, they have also shown better predictive performance over traditional methods. Better performances noticed with these can also be due to averaging of prediction decisions obtained from multiple predictors Dietterich (2000b). There are few existing studies in the literature which address dialect identification problems using ensemble techniques Huang et al. (2007), Darwish et al. (2014). Decision tree based random forest, extreme random forest, and rotation forest algorithms have been proposed for classification of dialects Chittaragi et al. (2018), Liu and Hansen (2011), Huang et al. (2007). The AdaBoost

Table 1 Details of commonly used standard dialect speech corpora

Database	Participants No. of dialects	Recording mode	Language	References
ABI (Accents of the British Isles)	300 (M + F), 15	Read	English dialects	D' Arcy et al. (2004)
CALLFRIEND	60 (M + F), 4	Telephone	English, Arabic, Spanish, Mandarin dialects	Canavan and Zipperlen (1996)
IViE (Intonational Variations in English)	112 (M + F), 9	Read and semiread	English dialects	Grabe and Post (2002)
Hindi dialect dataset	120 (M + F), 4	Read and spontaneous	Hindi dialects	Sinha et al. (2015)
Miami dataset	209 (M + F), 3	Read and spontaneous	English, Spanish	Zissman et al. (1996)
Multi-dialect multi-genre evaluation corpus (MGB-3)	-(M + F), 5	Spontaneous	Arabic dialects	Bahari et al. (2014)
NSP (Nation wide Speech Project)	60 (M + F), 6	Read and spontaneous	English dialects	Clopper and Pisoni (2006)
TIMIT (Texas Instrument and Massachusetts Institute of Technology)	630 (M + F), 8	Read and telephone	English dialects	Zue et al. (1990)
Present Kannada dialect speech corpus	156 (M + F), 5	Spontaneous	Kannada dialects	–

algorithm is being used for the classification of dialects based on word level utterances Huang et al. (2007). Similarly, extreme gradient boosting method has been used for recognition of English dialects from word and sentence utterances Chittaragi and Koolagudi (2017). These algorithms have reported a significant improvement in recognition performance over traditional single classifiers.

Apart from these, yet another most commonly applied classification method is artificial neural network (ANN). ANNs are said to be effective in capturing the complex non-linear relations present in data Rao and Koolagudi (2011). Due to advanced technological growth currently different variants of neural networks such as Deep Neural Network (DNN) and Convolutional Neural Network (CNN) models with larger number of hidden layers and activation functions are proposed in the literature for speech processing. These models are usually built by considering 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, these are found to be well suited for handling larger dataset with inherent complexities Zhang and Hansen (2018), Snyder et al. (2017). Few researchers have employed CNN based models which are familiar with image processing tasks for dialect classification Shon et al. (2018), Jiao et al. (2016). Currently, present research activities on speech processing are highly motivated to use DNNs and CNNs, since they are found to perform better over the traditional systems. However, these models are highly associated with a major issue of requiring larger speech dataset, as these models have shown poor performance with smaller datasets.

Apart from these, several researchers have applied phonotactic language-based methods for dialect identification. This is mainly due to the assumption that both language and dialects share similar linguistic and phonetic properties. Standard Phone Recognizer followed by a Language Modeling (PRLM) approach and a few of its varieties like, parallel PRLM (PPRLM) and parallel phone recognition (PPR) are proposed for classification of dialects Zissman et al. (1996), Chen et al. (2010), Biadys and Hirschberg (2009), Shen et al. (2008). However, majority of these require a transcribed speech for processing. A detailed review regarding dialect identification can be found in this paper Etman and Beex (2015).

Many times, research activities for dialect processing are limited mainly due to the unavailability of standard datasets for regional languages. Till date, identification of clear boundaries between dialects is a major challenge with many languages, except for few standard languages. A few standard speech corpora are available for language, dialect, and speech recognition tasks. Table 1 provides brief details of the most commonly used standard dialect speech datasets. Usually, recorded audio samples are in either spontaneous or read mode. Among them few datasets include telephonic conversations.

In the context of Indian scenario, limited studies have been reported on dialect processing tasks on Indian languages. A few such studies have addressed dialect processing of Hindi, Kannada, Malayalam, Punjabi, Assamese, and Telugu languages. Five prominent Hindi dialects are addressed by extracting MFCCs, pitch, energy and duration prosodic features from syllables. Auto-Associative Neural Network (AANN) and SVM classification methods are used for



Fig. 1 Five Kannada dialects spoken in different geographical regions of Karnataka: CENK: Central Kannada, CSTK: Coastal (Karavali) Kannada, HYDK: Hyderabad Karnataka Kannada, MUBK: Mumbai Karnataka Kannada, STHK: Southern Kannada

classification task Rao and Koolagudi (2011). Acoustic-phonetic feature based dialect identification system is proposed for four dialects of Hindi language using AANN method Sinha et al. (2015). An empirical analysis of differences between linguistic and paralinguistic features are investigated. This study presented the summary of dialectal differences by measuring acoustic-phonetic features such as formants, pitch, pitch slope, duration and intensity of vowel sounds Sinha et al. (2019). Gaussian mixture model (GMM) based dialect classification system is proposed by extracting MFCC features for three main dialects of Telugu language such as coastal Andhra, Telangana, and Rayalaseema Mannepalli et al. (2016). Few studies have proposed dialect recognition systems on Assamese dialects and other languages spoken in North Eastern Indian states by using the formants and prosodic features using artificial neural networks Sarma and Sarma (2016).

With respect to the Kannada language, very few systems are found addressing the task of dialect identification. An acoustic-phonetic analysis of three major dialects of Kannada spoken in the South, North and coastal areas of the state of Karnataka is proposed. In this study, preliminary analysis of variations in formant frequencies of Kannada vowels is done from dialect perspective Nagesha and Kumar (2010). Recent research has suggested the use of MFCC, prosodic features such as pitch and energy profiles for capturing dialectal traits. It has presented an improved recognition rate with GMM models along with a fusion of features Soorajkumar et al. (2017). Very few attempts are being made to address ADI problem for Kannada language due to unavailability of standard speech corpus. Until now, studies conducted on dialects of Kannada have considered only three dialects. However, there are a few more speaking styles which can be categorized as dialects of the Kannada language spoken across Karnataka. Present work focuses on five prodigious dialects representing five diversified geographical regions with unique speaking styles. A dialect dataset proposed in this paper consists of five distinct dialects named as, central Kannada (CENK), coastal Kannada (CSTK, Karavali), Hyderabad Kannada (HYDK), Mumbai Kannada (MUBK) and Southern Kannada (STHK) dialects. A map of Karnataka state showing the five dialectal regions considered in this work is given in Fig. 1, along with the neighboring states.

3 Dialect dataset details

This section initially provides the details of newly proposed Kannada dialect dataset. It also includes the details of existing English dialect dataset used for comparative analysis. A little linguistic background of Kannada language and the procedure adapted during the collection of Kannada dataset are clearly given in Subsect. 3.1, similarly details of English (IViE) dataset are given in 3.2.

3.1 Kannada dialect speech corpus (KDSC)

India is a country with 23 official languages including English and consists of more than 2000 varieties of speaking styles within the languages spoken across the country including many regional languages Chandrasekaran (2012), Jain and Cardona (2007). A majority of them belong to the group Indo-Aryan (Hindi, Bengali, and Marathi, etc.) spoken by 76.86% of the population of India. Dravidian languages (Kannada, Malayalam, Tamil, and Telugu) are spoken by 20.82% of the Indian populace Vanishree (2011). Very few are tonal (Punjabi, Haryanvi, Assamese, etc.) languages with the tone (pitch) distinguishing the grammatical meaning of words pronounced Sarma and Sarma (2016). Dravidian languages with their several dialects are mostly spoken in the southern region of India. Karnataka is a state in southern India. Kannada is the official language and spoken by more than seventy million people in Karnataka, a few other regions of India and the world. There are approximately more than 22 linguistically identified varying speaking styles of Kannada based on the 2011 census Jain and Cardona (2007), Rajapurohit (1982).

3.1.1 Kannada language background

Kannada is a highly agglutinative (concatenative) and morphologically rich language with the influence of Sanskrit in it. Similar to the other Dravidian languages, the agglutinative property includes the creation of new words with suffixes to the root word. Hence, complex words are formed by adding morphemes (meaningful word elements) together without changing them in spelling or phonetics. Morpho-syntax is determined by the order in which suffixes get attached to the root word. Kannada is an example of a verb-final inflectional language including relatively free word order Rajapurohit (1982). There are more than 10,000 basic stems (root words) in Kannada language. Also, there are more than a million morphed variants owing to more than 5000 distinct character variants Soman et al. (2011). 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 appeared in combination with vowels. Vowels are often observed more as carriers of dialectal variations in Kannada than consonants Nagesha and Kumar (2010), 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). It is observed that, spoken Kannada varies from region to region whereas its written counterpart remains unchanged.

In this study, five significant dialects representing five distinct broad geographical regions of Karnataka are considered. Each region covers a community with its own specific cultural identity, which speaks a particular dialect predominantly. Kannada speaking styles are primarily vary in each of the dialectal regions. The observed dialectal variations are found with speaking rate, melody, intonation patterns, unique rhythmic patterns, distinction in vowels, consonants, syllables, and word pronunciations Rajapurohit (1982). These differences can be en-cashed for dialect discrimination tasks. From a computational linguistic perspective, for dialect processing significant work is not reported in the literature. Kannada language is not much explored, for dialect processing, as even current research activities are in the nascent stage when compared to other Indian languages. Lack of systematic works on Kannada dialects is the has motivation for this work to be carried out. Due to unavailability of standard dialect dataset, a new dataset KDSC has been collected and the procedure followed while recording is briefly discussed in the following subsection.

3.1.2 Considerations in designing new Kannada dialect corpora

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. are needed to be controlled Clopper and Pisoni

Table 2 Kannada dialect speech corpus (spontaneous Speech)

Sl.No.	Name of dialect	Age (years)	Number of participants				Total speakers	Duration (minutes)
			Male		Female			
			No.	Dur. (in min)	No.	Dur. (in min)		
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

(2006). In this study, the demographics of the speakers are ensured by considering the following aspects during recording: (1) age above 20 years, (2) a balanced gender ratio, (3) minimum education, (4) resident in the same place for more than 5 years, (5) native speakers of the dialect, with parental history. Along with these aspects, the recording equipment that is used and the environment in which the recording is done, does influence the quality of the recordings and speaking styles. Usually, in the literature two types of datasets are found, one is studio recorded speech or recorded in a clean, controlled atmosphere. The other is recording done on-site, infield or marketplace or home which is realistic and available as the natural conversation. This type of recording involving natural interviews generally captures background noise as well. Poor quality of recording devices can also negatively influence the quality of the speech Clopper and Pisoni (2006). The speech collected may be text dependent or read in nature where pre-typed text is recorded. In the other case speech recorded may be a spontaneous conversation which is generally text independent.

The text independent Kannada dialect speech corpus, projected in this study is recorded from various parts of the state of Karnataka, mostly from rural and interior places. It has been found in the literature that, dialectal cues are naturally prominent in spontaneous speech rather than in read speech Rouas (2007). Spontaneous speech has obvious prosodic cues such as different speaking rate, filled pauses, intensity, intonation, hesitations, repetitions, and partially spoken words, etc. Liu et al. (2010). These features are said to convey dialectal variations to a greater extent. It is also true that even the native dialectal cues vary due to changes 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 formal/school education and are from rural areas. This is done to reduce the influence of pseudo accents of city-bred, well-educated speakers and to keep the recordings as legible (audible) and clear as possible. Details of the dataset recorded are presented in Table 2. The recording procedure followed is as follows:

- Speakers are made to participate in an informal discussion with an interviewer.
- The conversation begins as the interviewer asks participants to share their personal details such as name, age, qualification, place etc. Then, they are made to talk about their work experiences, businesses, tourist places visited, their hobbies, festivals, rituals, recent movies watched, agricultural activities conducted and so on; based on the development of discourse. This elicited speech covers topics on different work situations, travel, cultural aspects as well as entertainment. This is done to prompt the speakers to speak on diversified aspects of life's experiences, by using maximum vocabulary possible.
- 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.

Table 3 IViE dialect speech corpus (Semi-spontaneous speech)

Sl. No.	Region	Dialect names	Number of participants (Male + Female)	Duration (in mins.)
1	Belfast	ID1	12 (6 + 6)	32
2	Bradford	ID2	12 (6 + 6)	31
3	Cardiff	ID3	12 (6 + 6)	35
4	Cambridge	ID4	12 (6 + 6)	37
5	Dublin	ID5	12 (6 + 6)	33
6	Leeds	ID6	12 (6 + 6)	31
7	Liverpool	ID7	12 (6 + 6)	26
8	London	ID8	12 (6 + 6)	38
9	Newcastle	ID9	12 (6 + 6)	31
Total duration:			~ 5 h	

The proposed speech dataset is assumed to be the first standard corpus recorded in a systematic way, for analysis of the five Kannada dialects of a regional language. This dataset is sufficiently large to analyze the five distinct dialectal regions of Karnataka sufficiently covering the speakers' age, gender, and text variabilities. The works done thus far by other researchers in this domain have been reported for dialect classification considering a maximum of two or three dialects and also with limited speakers over a shorter duration of time Soorajkumar et al. (2017), Nagesha and Kumar (2010).

3.2 Intonational variation in English (IViE) speech corpus

The standard IViE speech corpus is used to evaluate the ADI system proposed in this work. Nine dialectal variants of British English spoken in nine different regions in British Isles are covered in this dataset. The corpus has been recorded in order to investigate the cross-varietal, stylistic variations, and intonation patterns across nine dialects of British English. The nine dialectal regions included are: Belfast (ID1), Bradford (ID2), Cardiff (ID3), Cambridge (ID4), Dublin (ID5), Leeds (ID6), Liverpool (ID7), London (ID8), and Newcastle (ID9). Actual corpus provides the recordings in text-read and semi-spontaneous mode. The read speech includes the speech recorded when the 'Cinderella' story is read from a printed script by female and male speakers. Whereas, in semi-spontaneous mode recordings, the speakers are made to rephrase or narrate the story of 'Cinderella' in their own words after having read the story. Therefore, the recording is termed as semi-spontaneous mode. The recording has been carried out using 12 subjects (6 Female + 6 Male adolescents) in each dialect, in a studio environment Grabe and Post (2002). Table 3 presents brief information about IViE corpus.

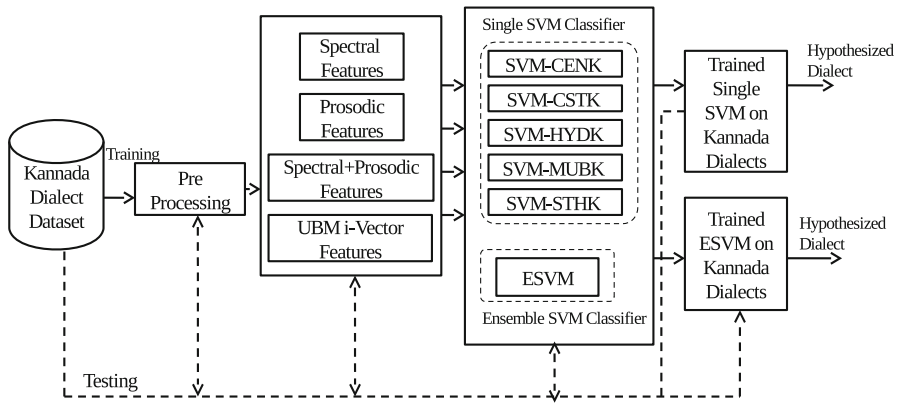


Fig. 2 Flow diagram of dialect identification system (SVM-Support Vector Machine, ESVM-Ensemble SVM)

4 Kannada dialect identification system

The proposed ADI system is implemented by using single and ensemble SVM classification methods by extracting spectral and prosodic features. Different sub-systems are developed to capture the prominent dialect features through spectral MFCCs alone, MFCCs combined with spectral flux, entropy features and the other using prosodic features (pitch and energy). Individual sub-systems are developed by using respective sets of feature vector. Further, dialect identification system is designed by combining all features extracted through equal weight fusion method.

Single classifier based SVM methods are found to be a very powerful classification and prediction algorithms. These have been designed with the intention to handle high dimensional input feature vectors. However, very few studies are found in the literature that addresses dialect classification problems using SVM Campbell et al. (2006). Nowadays, ensemble algorithms are gaining more attention as they have demonstrated better performance. Because they combine the predictions of several classifiers built on smaller instances of data Dietterich (2000a). Hence, in this work, an attempt has been made to compare the performances of these two algorithms.

Subsequently, the systems developed are used for analysis of the gender-specific influences on dialect identification systems. Further, gender dependent (individual system for male and female speakers) and independent (combined with male and female speakers) systems are developed. The flow diagram of the proposed ADI system is presented in Fig. 2. Details of various features extracted and classification methods that have been employed in the present work are discussed in the following subsections.

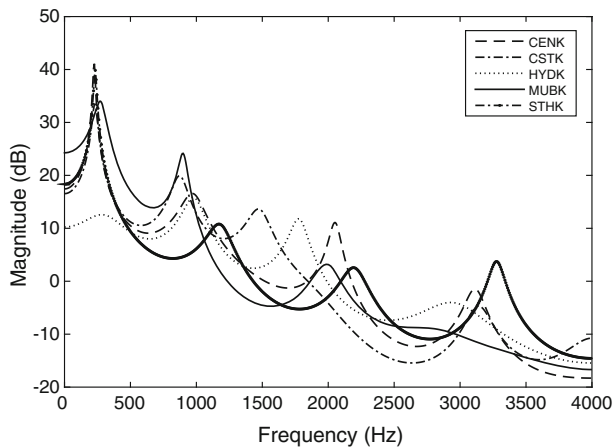


Fig. 3 Average LP spectra of vowel /e/ for five dialects

4.1 Feature extraction

In this paper, spectral features MFCCs and SDCs are extracted along with spectral flux and entropy features to model the vocal tract variations effectively through short term processing. Since cepstral coefficients resemble the human perception system, past research studies have explored MFCCs for solving a wide array of problems for language, and dialect-related tasks. Besides, SDC features are also considered to incorporate temporal cues into the feature vector Liu and Hansen (2011), Huang and Hansen (2007). Besides these, prosody is one of the essential features to be considered for dialect identification Rouas (2007). Indeed prosodic features such as different speaking rate, pitch variations, and energy capture unique pronunciation patterns across dialects. The obtained features are processed statistically to get new derived features. Further, the details of the extraction of i-vector features are discussed.

4.1.1 Spectral features

A specific sound unit of a particular language represents a unique pronunciation pattern due to the unique articulatory configurations followed in the vocal tract. Every spoken word or phoneme in any language follows a unique sequence of vocal tract shapes Reddy et al. (2013). Similarly, this assumption holds good for dialects also, since dialects are varying pronunciation patterns of a language. Due to which, every uttered unit has its own certain uniqueness concerning each dialect. These dialectal variations can be modeled through short-term spectral analysis of speech. Similarly, spectral variations are noticed in the pronunciation of a Kannada vowel /e/ uttered by a randomly selected speakers from five dialects. Average LP spectra drawn for the vowel is shown in Fig. 3. From spectra, the existence of systematic and significant differences among five dialects may be observed through vowel inherent spectral properties. These variations may be noticed through the

differences in energy levels, spectral peaks, spectral sharpness and positions of formant frequency values (F1–F4) among five dialects of Kannada language.

MFCCs Traditional RASTA (Relative Spectra) processed 13 MFCC features along with shifted delta coefficients (SDC, 13 deltas and 13 delta-delta) are derived for effective speech analysis. A total of 39 MFCC+SDC features are extracted from 20 ms frame with 10 ms frame shift from 40 filter banks Hermansky and Morgan (1994). MFCCs is the representation of the overall characteristics of the vocal tract system. 13 MFCCs gives the power spectral envelope of a frame, whereas, trajectories of the MFCC coefficients over time are measured using SDCs representing dynamic cues in speech Liu and Hansen (2011). RASTA technique based processing assists in the suppression of noise and channel distortion by using filter in the log domain of the power spectrum.

Spectral flux Timbre is an auditory sensation and speaker specific feature used to compare the similarity of speech utterances. It is commonly measured using the spectral flux feature. Slight differences appear in pronunciation among dialects of Kannada can be measured using flux feature. These can be distinguished by identifying the quick variations in the power spectrum of a signal. The spectral flux usually corresponds to a perceptual roughness of the sound. Flux feature is calculated by subtracting the power spectra of two consecutive frames Giannakopoulos and Pikrakis (2014).

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

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 Spectral entropy of a signal measures the distribution of spectral power. Spectral entropy feature captures the abrupt changes occurring in the energy level of an audio signal. Computation of the spectral entropy feature is done by dividing the spectrum of a frame into L sub-bands (bins). The energy E_f of the f th sub-band, for $f = 0, \dots, L-1$ is calculated and energies of all bins are normalized by taking the total spectral energy, where, $nf = \frac{E_f}{\sum_{f=0}^{L-1} E_f}$. The entropy of each normalized energy value is computed using the following Eq. (2)

$$H = - \sum_{f=0}^{L-1} nf \cdot \log_2(nf) \quad (2)$$

In this work, L value is set to 10 indicating that every frame is divided into small sized 10 bins.

4.1.2 Prosodic features

Prosodic features contribute a 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, reduction or prolongation of few speech units, imposing stress on few sound units while pronouncing, melody, and so on increases the intelligibility of the spoken words Ramus and Mehler (1999). Rhythm, stress and intonation features are complex perceptual entities usually expressed primarily through energy, pitch, and duration respectively. However, prosody features play a prominent role in the perception of different dialects, since most of the dialects primarily demonstrate variations in both pitch and energy features. In this work, pitch values for utterance are computed by using the autocorrelation method Giannakopoulos and Pikrakis (2014). Fundamental frequency (F0) which is a physical correlate of the pitch represents the vocal fold vibration rate. Due to this, a different F0 can be realized among dialects.

Energy feature associated with the speech signal is time varying in nature. It corresponds to the loudness attribute. 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 with amplitudes of concerning samples within a frame. Short-term energy feature is calculated as per the formulation as given in Eq. (3).

$$E(i) = \sum_{n=1}^{w_L} |x_i(n)|^2 \quad (3)$$

here $x_i(n)$, $n = 1, \dots, W_L$ be the audio samples in the i th frame, where W_L is the length of the frame.

$$E(i) = \frac{1}{W_L} \sum_{n=1}^{w_L} |x_i(n)|^2 \quad (4)$$

Energy is normalized over a frame by dividing it with W_L to remove the dependency on the frame length Giannakopoulos and Pikrakis (2014).

4.1.3 Feature extraction and post processing

The speech signal is segmented into M frames of length 20 ms with a 10 ms overlap. The spectral features discussed earlier are extracted from each frame resulting $M \times N$ dimensional matrix with N feature vector size. Similarly, prosodic features are also extracted to form $M \times N$ matrix. In this study, these features are named as raw features (since extracted from speech). Further, statistical processing of these matrices is done to obtain a new feature set called derived feature set using mean and standard deviation statistical parameters. These are computed by taking the statistical mean between F_i and F_{i+1} two consecutive frames to obtain first N features. Later, features from $N+1$ to $2N$ represents the standard deviation of the same frames. The dimension of the feature vector of each frame is doubled to $2N$ from N Giannakopoulos and Pikrakis (2014), Chittaragi et al. (2018). The steps involved in feature extraction and post-processing of features is shown in Fig. 4. These two statistics extracted from two consecutive frames exhibit similar behavior, and a few temporal variations are discarded due to averaging of two consecutive

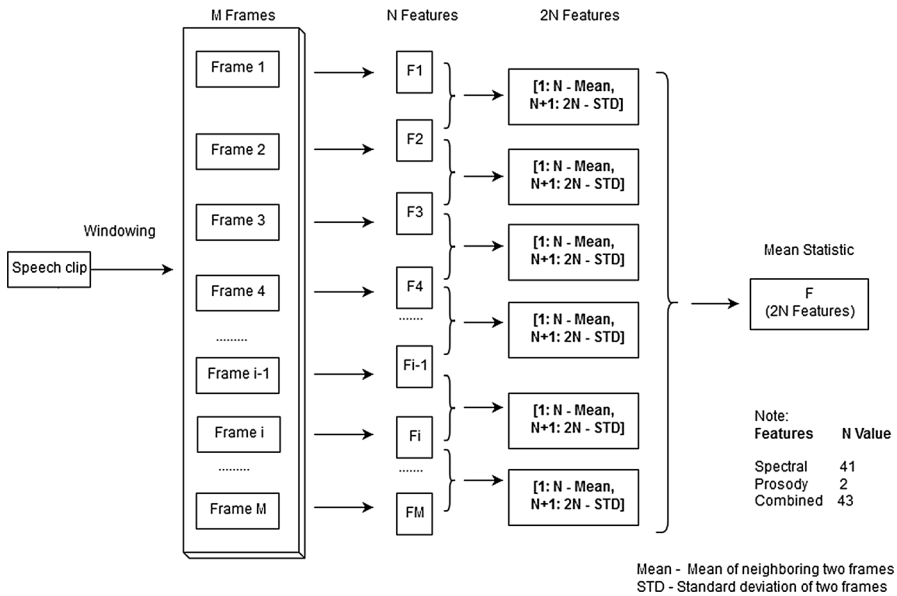


Fig. 4 Flow chart for deriving statistical features

frames. A matrix with size $M \times 2N$ is obtained. Now, at this stage averaging of all frames is done to get the final derived feature vector of size $2N$. However, the N raw features computed from taking the mean of all frame at the earlier stage are noticed to be different from $2N$ derived features. In this work, derived feature vector size is 82 $(39+1+1) \times 2$ for spectral features, 4 (2×2) and 86 (43×2) for prosodic and combined features respectively. With these features, a hypothesis made is that, derived features additionally contribute to dialect recognition performance over raw features.

4.1.4 UBM based i-vector features

Nowadays, state-of-the-art i-vector features are found to be quite successful features in speaker and language recognition systems Dehak et al. (2011). Very few attempts are made in the literature using them for dialect recognition tasks Hansen and Liu (2016), Zhang and Hansen (2018). RASTA processed 13 MFCC, and 39 MFCC-SDC raw features are used to obtain i-vectors using the method proposed in Dehak et al. (2011), Hansen and Liu (2016). The i-vectors are represented in a low ranked matrix T that captures relevant variabilities concerning total variability matrix. Conceptually, the 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. The basic idea is adapting the Universal Background Model (UBM)² to a set of given speech frames. Estimation of utterance

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

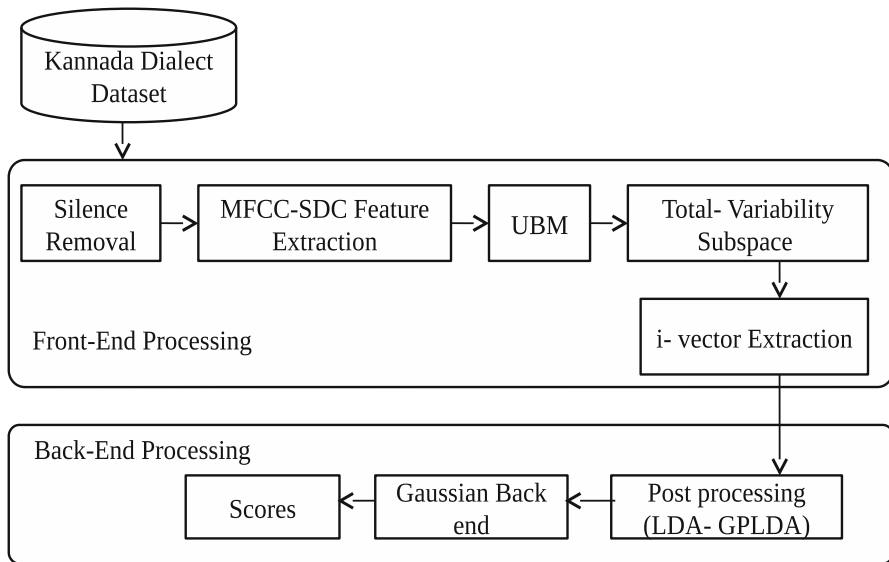


Fig. 5 Steps involved in UBM-i vector feature extraction

dependent GMM is done through the Eigenvoice adaptation technique Dehak et al. (2011). The GMM super-vector is obtained by stacking all mean vectors from the GMM for a given utterance. The i-vector is modeled as follows Sadjadi et al. (2013):

$$M = m + T\omega + \epsilon \quad (5)$$

where M is the speaker and session dependent, UBM super-vector derived from either MFCCs or MFCC-SDC features, and m is speaker and session independent super-vector taken from UBM. T is a rectangular low rank matrix called a total variability matrix, and ω is referred to as total vectors or i-vector features. The residual noise term is shown as ϵ . Various stages, followed during the extraction of UBM based i-vectors, are presented in Fig. 5.

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 ten iterations for every mixture are selected heuristically. From the trained UBM, a total variability (TV) matrix is also trained with the same data. From this, i-vectors of 300 dimension are computed so that the fixed length feature vectors are obtained from the varied length audio files. Expectation-Maximization algorithm is employed to train both UBM and total variability matrix using complete training. Later, the i-Vectors for both train and test sets are extracted, and Gaussian Back-end algorithm is used for five class dialect modeling.

4.2 Classification algorithms

In the present work, a single classifier based SVM and multi-classifier based ensemble based SVM algorithms are employed for the dialect classification. A single classifier based SVM method is used since it has shown better generalization performance across several applications of speech. 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 Chittaragi et al. (2018), Pedersen and Diederich (2007), Biadys et al. (2011). In order to handle high-dimensional feature space, radial basis function (RBF) kernel is used for implementation of SVM along with the one-versus-rest approach for handling five class problem.

Single classifier based SVM algorithm In this work, a dialect identification system has been developed by using MFCC, SDC, pitch and energy features using single classifier based SVM. SVM method is modeled to capture dialect specific cues. Five different individual SVMs are trained on five dialects of Kannada using one-versus-rest technique to solve five dialects classification problem. Radial basis function (RBF) kernel is used for separating hyperplane from the maximal margin in a high-dimensional feature space Chang and Lin (2011).

Both spectral and prosodic features are extracted individually from each of the five separate SVM models and are trained with individual and combination of features. Training inputs from all five classes are of the form $\left\{ \left\{ (x_i, k) \right\}_{i=1}^{N_k} \right\}_{k=1}^n$, where N_k is the total speech inputs belonging to k th dialect class, k takes five labels $k = 1, 2, 3, 4, 5$. All these are used to train the SVM model individually on five classes. SVM for each class k is constructed by using the set of training inputs and the desired outputs, $\left\{ \left\{ (x_i, y_i) \right\}_{i=1}^{N_k} \right\}_{k=1}^n$, the desired output y_i for the training example x_i , takes value $+1$ if $x_i \in k$ th class representing positive example, else -1 represents the negative example. After the training, evaluation of the system is performed by deriving feature vectors from test speech clips. These are given as inputs to all trained SVM models. For instance, a test pattern x , the evidence $TS_k(x)$ is obtained from all five SVM models. The class label k associated with SVM, which gives maximum evidence is hypothesized as the dialect class C of the test pattern, i.e. $C(x) = \operatorname{argmax}_k (TS_k(x))$.

Ensemble SVM (ESVM) algorithm Ensemble algorithms have shown better performances over single classifier based systems. In these algorithms, predictive performance rely on decisions made by multiple classifiers employed on the smaller sub-problems (a), Utami et al. (2014). In this work, an ensemble algorithm is employed for implementation of the dialect identification system where SVM is used as a base learner (classifier). Bagging and boosting are two techniques used commonly while implementation of ensemble algorithms Chittaragi et al. (2018). In this work, bagging (bootstrap aggregating) technique is adapted to construct ESVM. Bagging technique combines predictions from independent base models that are derived from bootstrap samples by sub-sampling through replacement of the original data Breiman (2001).

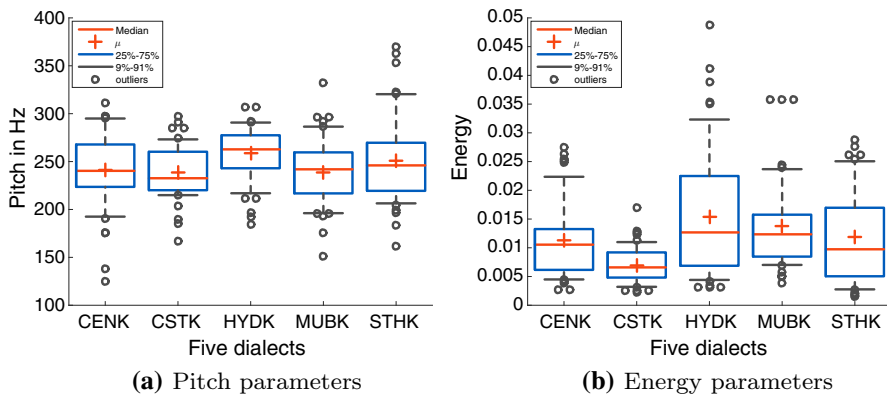


Fig. 6 Statistical range for pitch and energy values of five Kannada dialects, CENK:Central Kannada, CSTK:Coastal (Karavali) Kannada, HYDK:Hyderabad Kannada, MUBK: Mumbai Kannada, STHK: Southern Kannada

Using this, training set $TR = (x_i, y_i) | i = 1, 2, 3, 4, 5$ is divided into K training sample set to construct the K independent SVMs. Here K is set to 2048 empirically. Hence, 2048 training sets are created from TR through random sampling with replacement. Each of these is trained with SVM classifier individually. Independently trained SVMs are later aggregated either using majority voting or the probability sum. After training, since it is a classification problem, the majority voting method is followed to combine SVMs. Binary classifier based SVM is extended by using the one-vs-rest method to handle the multi-class problem (here five dialect classes here) using libsvm implementation Pedregosa et al. (2011). RBF kernel is chosen among available kernel functions empirically that have resulted in better performance. Using ESVM classification accuracy is expected to improve since combinations of several SVMs tries to expand the correctly classified area incrementally.

5 Experiments, results and discussions

In this work, experiments are carried out to evaluate the performances of ADI systems that have been implemented using spectral, prosodic features and the combination of both. In addition to this, the performance comparison of both single SVM and ESVM classification methods is performed with features extracted. Initial subsection covers details of the statistical analysis of the prosodic features namely, pitch and energy. Later, it includes analysis of results obtained with two classification methods individually followed by a comparative study of developed systems with the existing state-of-the-art system. Finally, performances achieved on Kannada datasets are compared with results that have been obtained on standard IviE dataset.

Table 4 Mean values of the prosodic features for each 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

5.1 Statistical analysis of prosodic features on Kannada dialects

Prosody cues namely, stress, rhythm, and intonation are complex perceptual entities expressed primarily through acoustic features: pitch, energy, and duration. These parameters existence in speech conveys important dialect specific information. Also, few studies on language identification have specifically confirmed that, pitch and energy prosodic information can be used where acoustics and phonotactics information is degraded Mary and Yegnanarayana (2008).

In order to understand the influences of prosodic pitch and energy features on 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 distribution of data based on five statistical parameters namely: median, minimum, maximum, two intervals: one between 25 to 75%, and other between 9 to 91%. Statistical parameters obtained with Kannada dialects are presented in terms of box plot as shown in Fig. 6. In addition to this, the mean and standard deviation of pitch and energy features from female and male speakers of five dialects are presented in Table 4.

From Fig. 6 and Table 4 following observations are made. It has been noticed that both pitch and energy parameters are comparatively low with CSTK dialect. The span of both energy and pitch feature is found to be smaller in CSTK dialect which is spoken in the coastal region. CSTK dialect speakers are found to follow lower energy and pitch value range and use very similar speaking style across the region. This is mainly because Kannada is the second language for coastal region speakers (Tulu is the first language for majority of the speakers). Kannada is an acquired language for them and Kannada is spoken mostly on formal occasions. They are likely to pronounce Kannada words more consistently and consciously. This results in minor variation in acoustic properties of vowels and consonants across phonetic contexts in both read and spontaneous speech Nagesha and Kumar (2010). They are also said to follow a unique intonation with rhythmic styles.

Speakers of HYDK dialect, that is spoken in Hyderabad region have demonstrated higher values of pitch and energy when compared to all other Kannada dialects. The span of energy feature range is observed to be broader in this region. Also, while speaking the rate followed in this dialect is found to be higher and in turn results in reduction of syllables in the final words. Speakers of this region are found using highly stressed words and syllables. Even though high pitch values are noticed; the span of pitch value range is found to be smaller with this dialect speakers Soorajkumar et al. (2017). It is also observed from these speaking styles that they follow unique pronunciation patterns very different from other Kannada dialects. These variations are mainly due to the influence of two languages of the neighboring states. The HYDK dialect is influenced by Marathi and Telugu languages, due to its proximity to Maharashtra and Andhra Pradesh states respectively. Apart from these, Kannada spoken in HYDK dialect is slightly different due to the influence of the former Hyderabad Nizam Kingdom and the use of a fair chunk of Urdu vocabulary Rajapurohit (1982).

The speakers of the Mumbai region (MUBK) also have the influence of Marathi language on Kannada and also there is fair use of the Urdu vocabulary while

Table 5 Average dialect recognition performance using single SVM classifier method on KDSC; SF-Spectral flux, SE-Spectral entropy, size of feature vector is given in brackets()

Accuracies in %					
Sl. No.	Feature set	MFCC (13)		MFCC + SDC (39)	
		Raw features		Raw features	
		Derived 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)

speaking. Although MUBK region is close to the Hyderabad region, the speaking rate noticed here is comparatively slow, the pitch values are lesser and slightly lesser in energy values Soorajkumar et al. (2017). Interesting observations are made with CENK and STHK dialect speakers as they follow very little similarity in the speaking pattern. For CENK and CSTK dialects, Kannada language is a primary or mother tongue and these two speaking styles are known to be closer to the written form of Kannada language. In specific, CENK dialect is very closer to the written/standard form of Kannada where, every utterance is phonetically clearly pronounced. However, STHK dialect speakers have shown a little faster speaking styles similar to that of HYDK dialect with higher energy values. Whereas, a difference exists with pitch variations between them; both feature intervals are found to be with wider span. Outliers are represented with “o” and are found in all dialects with both pitch and energy features. Outliers are slightly more among CENK and STHK dialects. Apart from these, outliers are detected to be more with STHK dialect with pitch feature. Similarly, with the HYDK dialect there are outliers with energy features.

In general, prosodic and spectral features extracted in this work are said to carry non-overlapping dialect specific features Rao and Koolagudi (2011). This is because, spectral features try to exploit dialectal cues from the vocal tract system (speech production) and prosodic features from speech perception point of view. For instance, higher energy parameter generally indicate higher pitch values and slower speaking rate. For example, dialects like HYDK and STHK have shown higher energy and pitch range with reduced speaking rate. From this observation it may be considered that prosodic features as the co-relates with dialects. Further, spectral features also play a major role in discriminating dialects Chittaragi et al. (2018). Hence, combining spectral features along with prosodic features are expected to improve accurate recognition of dialects. Indeed, it is also required to combine features from multiple levels such as segmental and supra-segmental levels Soorajkumar et al. (2017), Nagesha and Kumar (2010).

5.2 Performance evaluation using single SVM classification method

In this paper, a series of experiments are conducted using a single SVM classification method on proposed KDSC. It includes the text independent, spontaneous speech recorded from native speakers. Four different ways of spectral and prosodic features are followed for implementation of ADI systems. They are: (1) using 13 MFCC, (2) using 13 MFCC, spectral flux and spectral entropy, (3) using pitch and energy features, and (4) the combination of all features. The experiments have been carried out with fivefold cross-validation approach. Dialect recognition performance obtained from all experiments are tabulated in Table 5. In addition, the experimentation is conducted with raw (original features extracted from speech) and derived (post-processed to derive two statistics from the original) 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 5, it is observed that, 78.50% of recognition accuracy is obtained with 13 MFCC features. Addition of spectral flux and entropy features to MFCCs

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
	CENK	CSTK	HYDK	MUBK	STHK

Fig. 7 Confusion matrix achieved with both spectral and prosodic features using SVM on Kannada dialect identification. Average recognition rate: 84.41%

has resulted with 79.29% performance. From these results, it may be noticed that spectral features alone are relatively successful in the classification of Kannada dialects. Where, MFCC features have characterized the unique vocal tract shape variations followed across five Kannada dialects. However, slightly lesser performance is seen with only prosodic features on KDSC. Even though prosodic features play a significant role in discrimination of dialects, prosodic features are not able to distinguish Kannada dialects correctly. It is noticed from the results that, the combination of both features has shown slight increase of about 80.75% of accuracy. 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 conducted 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. The hypothesis made earlier is found to be true with these, as derived features have performed better in all four ways. Also, 10.25% increment is noticed with the use of proposed derived features against raw features.

In general, SDC features are added to incorporate additional temporal information regarding differential and acceleration coefficients to feature vector Liu and Hansen (2011). Further, to verify the influence of SDC features on Kannada dialects, similar experiments have been conducted by adding a 26-dimensional feature vector (13 deltas and 13 delta-delta) to 13 MFCCs (total 39). Performances obtained with the addition of SDC features are presented in Table 5. Some interesting observations are accomplished, as decrease in performances is noticed after addition of SDCs. In all four ways, the performances obtained are lesser over the use of 13 MFCCs with Kannada language. From these results, it hints that

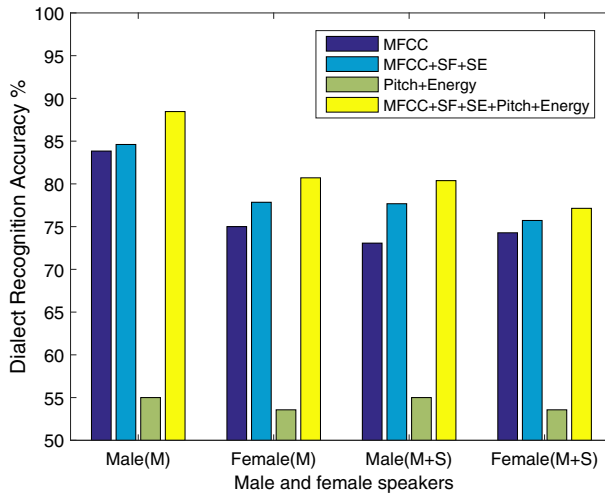


Fig. 8 Comparison of dialect recognition performance for male and female speakers, SF-Spectral Flux, SE-Spectral Energy, M-13 MFCC and M+S-MFCC+SDC features

temporal dynamics captured through SDCs may not be much useful on Kannada dialect classification.

The confusion matrix obtained with the combination of two features is given in Fig. 7. It is drawn for the ADI system with highest accuracy of 84.41% obtained using a 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 has a unique pronunciation style with higher pitch and energy values than the remaining four dialects. Even more interesting is, mis-classification 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 comparatively lower pitch and energy features. Hence, speakers from this region with higher pitch and energy values are misclassified with other dialects, slightly more with Mumbai dialect. About 83.33% of STHK dialect samples are classified correctly and 16.66% of misclassification with Hyderabad region. This is primarily due to use of similar speaking styles with high pitch and energy properties.

Further, during the recording process, dialect-specific differences have been investigated earlier during formal interaction among male and female speakers. Hence, gender dependent analysis is done individually with female and male speakers on five Kannada dialect. An analysis performed over spontaneously produced pronunciations in a larger corpus has revealed that female speakers tend to follow standard pronunciations than male speakers. Male speakers show larger proportions with filled pauses and repetitions Clopper and Smiljanic (2011). Kannada dialects are classified more accurately with the male over female speakers in every scenario considered in this paper. Comparison of dialect recognition

Table 6 Average dialect recognition performance using ESVM classifier method on KDSC: SF-Spectral flux, SE-Spectral entropy, size of feature vector is given in brackets()

Accuracies in %			
Sl. No.	Feature set	Raw features	Derived features
1	MFCC	79.06 (13)	81.36 (26)
2	MFCC+SF+SE	81.25(15)	83.12 (30)
3	Pitch+Energy	38.75 (02)	44.52 (04)
4	MFCC+SF+SE+Pitch+Energy	84.56 (17)	86.25 (34)

performances with derived features on male and female speakers is presented in Fig. 8. An interesting observation made from this is that, a slight reduction in performance is seen with female speakers when 39 MFCCs are used over 13 MFCCs features. However, a significant reduction is noticed with male speakers as the addition of SDC features. Further, Kannada dialects are classified more accurately with combined feature vectors and have also demonstrated lower performance with prosodic features.

5.3 Performance evaluation using ESVM classification method

It is proposed in the literature that, an ensemble of various classifiers work better over traditional single classifier based algorithms (a). In this paper, four different ADI systems are implemented with ESVM method on the KDSC. Table 6 shows the results obtained using ensemble SVM classifier with the RBF kernel and 13 MFCC features without addition of SDCs. With all experiments conducted, ESVM has performed comparatively better. The highest of 86.25% average accuracy is obtained with combined feature vector. Moreover, prosodic features have reflected lower performance using ESVM which is lesser than single SVM. Reason for lower performance may be due to the availability of very few features to construct larger SVM in ESVM method. Moreover, complete explanation given for single classifier SVM holds true even with ESVM. It is evident from obtained these results, where the use of ensemble SVM classifier can be a better choice for the dataset that contains a fewer number of features, than the numbers of data samples per class Kim et al. (2002).

5.4 Performance comparison with UBM based i-vector features

This section focuses on comparative analysis of the proposed features over the state-of-the-art UBM based i-vector features. Results obtained using GMM-UBM, and i-vector features on KDSC are given in Table 7. Experiments carried out with KDSC using MFCC features using traditional GMM-UBM have shown 75.5% dialect recognition accuracy. Use of UBM based i-vector features in terms of total variability matrix have shown a slightly better recognition rate of 81.40%, which is

Table 7 Average dialect recognition performance on Kannada dialect dataset using UBM i-vector features

Features	Models	Accuracies in %
MFCC (13*2)	GMM-UBM	75.50
MFCC-SDC (39*2)		74.90
MFCC (13*2)	UBM - i-vector based system	81.40
MFCC-SDC (39*2)		78.54

Table 8 Average dialect recognition performance using single SVM classifier method on IViE dialect dataset: SF-Spectral flux, SE-Spectral entropy, size of feature vector is given in brackets ()

Accuracies in % using derived features			
Sl. No.	Feature set	Single SVM	Ensemble SVM
1	MFCC (26)	84.48	86.72
2	MFCC+SF+SE (30)	85.59	87.85
3	Pitch+Energy (04)	61.09	46.31
4	MFCC+SF+SE+Pitch+Energy (34)	89.91	91.38

marginally better over the GMM-UBM system. The results achieved with the use of derived spectral and prosodic features on KDSC using single SVM and ESVM algorithms are comparatively better over the results obtained with existing UBM-i-vector feature based dialect recognition systems. Hence, it is evident from the overall results obtained, that, derived spectral and prosodic features extracted in this work have significant role in classification of five Kannada dialects.

5.5 Performance evaluation and comparison with IViE speech corpus

Majority of the time it is required to compare the results obtained with newly proposed system with the standard datasets available. Hence, all above mentioned experiments are conducted on the standard IViE dialect speech corpus to study the significance of spectral and prosodic features on English dialects also. An average ADI recognition obtained with derived spectral and prosodic features is presented in Table 8. Overall results obtained with both single SVM and ESVM methods are comparatively better. Meanwhile, it can be noted that spectral and prosodic combination of features have shown the highest 91.38% recognition rate. Indeed, prosodic features have also shown lower performance with IViE dataset. Further, IViE dataset includes nine dialects which are found to be clearly distinguishable. Nine dialects represent different parts of British Isles, the discrimination is implicitly present in all nine dialects of English in IViE dataset. Increased recognition rate may be due to IViE is a clean studio recorded semi-spontaneous speech corpus with limited speaker variabilities. Also, the size of the IViE is small when compared to KDSC. Whereas, five Kannada dialects that have been presented are very closely spaced with overlapping properties. Hence, there is a lot of

similarity among five dialects. Moreover, the proposed KDSC with larger speaker variabilities among all dialects and quality of the dataset has impacted in a slight reduction in performance.

6 Conclusions and future works

In this paper, the ADI system is proposed by exploring spectral and prosodic features on five dialects of Kannada language. A new text independent and spontaneous speech corpus is collected exclusively from native speakers from five prominent dialects of Kannada. Significant spectral MFCCs, SDC, spectral flux, entropy features and prosodic features are extracted for dialect characterization. Further statistical processing of these features is carried out to obtain derived features. In this paper, ADI systems are developed using single SVM and ensemble based ESVM classification methods. Later, results are verified with individual and combinations of the spectral and prosodic features. The existence of non-overlapping dialectal traits among these features and the reduction in performance with the addition of SDC features has been investigated on KDSC. Spectral features alone have demonstrated better performances over the prosodic features. Combined feature vector has shown better ADI recognition rate of 84.41% when compared to spectral (83.02%) and prosodic features (47.75%) alone with derived features using single SVM method on Kannada dataset.

Similarly, a slightly increased 86.25% recognition rate is obtained with a combination of spectral and prosodic derived features using ESVM method. Standard IViE English corpus is also used for validating the results obtained on KDSC. Proposed feature vector evaluated on IViE has shown a increased dialect recognition performance of about 91.38% with derived feature vectors. Further, dialect recognition performances that have been obtained from proposed features are slightly better over standard UBM based i-vector features on Kannada language. Dialect recognition studies can be further extended in future by identifying more dialect specific prosodic features. Investigation of acoustic-phonetic information from various levels of spoken units can be explored to improve the performance of ADI systems. Hyper-parameters of ensemble based classification algorithms can be further fine-tuned to build a better model to achieve an improved dialect recognition rate.

References

- Ahuja, P., & Vyas, J. M. (2018). Forensic speaker profiling: The study of supra-segmental features of Gujarati dialects for text-independent speaker identification. *Australian Journal of Forensic Sciences*, 50(2), 152–165.
- Arslan, L. M., & Hansen, J. H. L. (1996). Language accent classification in American English. *Speech Communication*, 18(4), 353–367.
- Bahari, M. H., Dehak, N., Van hamme, H., Burget, L., Ali, A. M., & 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.

- Behravan, H., Hautamäki, V., & Kinnunen, T. (2015). Factors affecting i-vector based foreign accent recognition: A case study in spoken Finnish. *Speech Communication*, 66, 118–129.
- Biadys, F. (2011). Automatic dialect and accent recognition and its application to speech recognition (PhD Thesis, Columbia University).
- Biadys, F., & Hirschberg, J. (2009). Using prosody and phonotactics in Arabic dialect identification. *Interspeech*, 9, 208–211.
- Biadys, F., Hirschberg, J., & Ellis, D. P. W. (2011). Dialect and accent recognition using phonetic-segmentation supervectors. In *Interspeech* (pp. 745–748).
- Bougrine, S., Cherroun, H., & Ziadi, D. (2017). Hierarchical classification for spoken Arabic dialect identification using prosody: Case of algerian dialects. arXiv preprint [arXiv:1703.10065](https://arxiv.org/abs/1703.10065).
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Campbell, W. M., Campbell, J. P., Reynolds, D. A., Singer, E., & Torres-Carrasquillo, Pedro A. (2006). Support vector machines for speaker and language recognition. *Computer Speech & Language*, 20(2), 210–229.
- Canavan, A., & Zipperlen, G. (1996). Callfriend American English-non-southern dialect. *Linguistic Data Consortium, Philadelphia*, 10, 1.
- Chambers, J. K., & Trudgill, P. (1998). *Dialectology* (2nd ed.). Cambridge: Cambridge University Press.
- Chandrasekaran, K. (2012). Indeterminacies in Howatch's St. Benet's Trilogy. *Language in India*, 12(12), 382–389.
- Chang, C. C., & 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., & Campbell, J. P. (2010). A linguistically-informative approach to dialect recognition using dialect-discriminating context-dependent phonetic models. In *IEEE international conference on acoustics speech and signal processing (ICASSP)* (pp. 5014–5017). IEEE.
- Chen, N. F., Tam, S. W., Shen, W., & Campbell, J. P. (2014). Characterizing phonetic transformations and acoustic differences across English dialects. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1), 110–124.
- Chen, T., Huang, C., Chang, E., & Wang, J. (2001). *Automatic accent identification using Gaussian mixture models* (pp. 343–346). IEEE workshop: In automatic speech recognition and understanding.
- Chittaragi, N. B., Koolagudi, S. G. (2017). Acoustic features based word level dialect classification using SVM and ensemble methods. In *Tenth international conference on contemporary computing (IC3)* (pp. 1–6). IEEE.
- Chittaragi, N. B., 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]* (pp. 1–6). Berlin: Springer.
- Chittaragi, N. B., Prakash, A., & Koolagudi, S. G. (2018). Dialect identification using spectral and prosodic features on single and ensemble classifiers. *Arabian Journal for Science and Engineering*, 43(3), 4289–4302.
- Clopper, C. G., & Pisoni, D. B. (2006). The nationwide speech project: A new corpus of American English dialects. *Speech Communication*, 48(6), 633–644.
- Clopper, C. G., & Smiljanic, R. (2011). Effects of gender and regional dialect on prosodic patterns in American English. *Journal of Phonetics*, 39(2), 237–245.
- D' Arcy, S., Russell, M. J., Browning, S. R., Tomlinson, M. J. (2004). The accents of the British Isles (ABI) corpus. In *Proceedings Modélisations pour l'Identification des Langues* (pp. 115–119).
- Darwish, K., Sajjad, H., & Mubarak, H. (2014). Verifiably effective arabic dialect identification. In *Empirical methods in natural language processing* (pp. 1465–1468).
- Dehak, N., Torres-Carrasquillo, P. A., Reynolds, D. A., & Dehak, R. (2011). Language recognition via i-vectors and dimensionality reduction. In *Interspeech* (pp. 857–860).
- Dietterich, T. G. (2000a). Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp 1–15). Berlin: 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.
- Etman, A., & Beex, A. L. (2015). Language and dialect identification: A survey. In *SAI intelligent systems conference (IntelliSys)*, (pp. 220–231).
- Ferragne, E., & Pellegrino, F. (2007). *Automatic dialect identification: A study of British English. In Speaker classification II* (pp. 243–257). Berlin: Springer.

- Giannakopoulos, T., & Pikrakis, A. (2014). *Introduction to audio analysis: A MATLAB approach*. Cambridge: Academic Press.
- Grabe, E., & Post, B. (2002). Intonational variation in the British Isles. In *Speech prosody*.
- Hanani, A., Russell, M. J., & Carey, M. J. (2013). Human and computer recognition of regional accents and ethnic groups from British English speech. *Computer Speech & Language*, 27(1), 59–74.
- Hansen, J. H. L., & 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., & Miglio, V. G. (2014). Prosody and its application to forensic linguistics. *LESLI: Linguistic Evidence in Security Law and Intelligence*, 2(2), 11–29.
- Hermansky, H., & Morgan, N. (1994). Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4), 578–589.
- Huang, R., & 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., & Angkititrakul, P. (2007). Dialect/accent classification using unrestricted audio. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2), 453–464.
- Jain, D., & Cardona, G. (2007). *The Indo-Aryan languages*. Abingdon: Routledge.
- Jiao, Y., Tu, M., Berisha, V., & Liss, J. M. (2016). Accent identification by combining deep neural networks and recurrent neural networks trained on long and short term features. In *Interspeech* (pp. 2388–2392).
- Kim, H. Chul, P., Shaoning, J., Hong M., Kim, D. & Bang, S. Y. (2002). Support vector machine ensemble with bagging. In *First international workshop on pattern recognition with support vector machines* (pp. 397–408).
- Lei, Y., & 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.
- Lim, B. P., Li, H., & Ma, B. (2005). Using local & global phonotactic features in Chinese dialect identification. In *International conference on acoustics, speech, and signal processing (ICASSP)* (Vol. 1, pp. I–577). IEEE
- Liu, G. A., & Hansen, J. H. L. (2011). A systematic strategy for robust automatic dialect identification. In *Nineteenth European signal processing conference* (pp. 2138–2141).
- Liu, G., Lei, Y., & Hansen, J. H. L. (2010). Dialect identification: Impact of differences between read versus spontaneous speech. In *Eighteenth European signal processing Conference* (pp 2003–2006). IEEE.
- Malmasi, S., & Dras, M. (2015). Language identification using classifier ensembles. In *Proceedings of the joint workshop on language technology for closely related languages, varieties and dialects*, (pp. 35–43).
- Mannepilli, K., Sastry, P. N., & Suman, M. (2016). MFCC-GMM based accent recognition system for Telugu speech signals. *International Journal of Speech Technology*, 19(1), 87–93.
- Mary, L., & Yegnanarayana, B. (2008). Extraction and representation of prosodic features for language and speaker recognition. *Speech Communication*, 50(10), 782–796.
- Ma, B., Zhu, D., & Tong, R. (2006). Chinese dialect identification using tone features based on pitch flux. *International Conference on Acoustics Speech and Signal Processing Proceedings (ICASSP)*, 1, 1029–1032.
- Mehrabani, M., & Hansen, J. H. L. (2015). Automatic analysis of dialect/language sets. *International Journal of Speech Technology*, 18(3), 277–286.
- Nagesha, K. S., & Kumar, G. H. (2010). *Acoustic-phonetic analysis of Kannada accents*. Mumbai: Tata Institute of Fundamental Research.
- Pedersen, C., & Diederich, J. (2007). Accent classification using support vector machines. In *Sixth international conference on computer and information science (IEEE/ACIS)* (pp. 444–449).
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Prahalad, K., Kumar, E. N., Keri, V., Rajendran, S., & Black, A. W. (2012). The IIIT-H Indic speech databases. In *Thirteenth annual conference of the international speech communication association*.
- Rajapurohit, B. B. (1982). *Acoustic characteristics of Kannada* (Vol. 27). Central Institute of Indian Languages.
- Ramus, F., & 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.

- Rao, K. S., & 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.
- Reddy, V. R., Maity, S., & Rao, K. S. (2013). Identification of Indian languages using multi-level spectral and prosodic features. *International Journal of Speech Technology*, 16(4), 489–511.
- 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., & 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., & 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)*, (pp. 127–132). IEEE.
- Shen, W., Chen, N., & Reynolds, D. (2008). Dialect recognition using adapted phonetic models. In *Proceedings of the annual conference of the international speech communication association, INTERSPEECH* (pp. 763–766).
- Shon, S., Ali, A., & Glass, J. (2018). Convolutional neural networks and language embeddings for end-to-end dialect recognition. arXiv preprint [arXiv:1803.04567](https://arxiv.org/abs/1803.04567).
- Sinha, S., Jain, A., & Agrawal, S. S. (2015). Acoustic-phonetic feature based dialect identification in Hindi Speech. *International Journal on Smart Sensing & Intelligent Systems*, 8(1), 235–254.
- Sinha, S., Jain, A., & Agrawal, S. S. (2019). Empirical analysis of linguistic and paralinguistic information for automatic dialect classification. *Artificial Intelligence Review*, 51(4), 647–672.
- Snyder, D., Garcia-Romero, D., Povey, D., & Khudanpur, S. (2017). Deep neural network embeddings for text-independent speaker verification. In *Proc. Interspeech* (pp. 999–1003).
- Soman, K. P., Ramasamy, V., Antony, P. J., & Saravanan, S. (2011). A rule-based Kannada morphological analyzer and generator using finite state transducer. *International Journal of Computer Applications*, 27(10), 0975–8887.
- Soorajkumar, R., Girish, G. N., Ramteke, P. B., Joshi, S. S., & 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*, (pp. 411–418). Berlin: Springer.
- Torres-carrasquillo, P. A., Gleason, T. P., & Reynolds, D. A. (2004). Dialect identification using Gaussian mixture models. In *ODYSEY—The speaker and language recognition workshop*, (pp. 2–5).
- Utami, I. T., Sartono, B., & Sadik, K. (2014). Comparison of single and ensemble classifiers of support vector machine and classification tree. *Journal of Mathematical Sciences and Applications*, 2(2), 17–20.
- Vanishree, V. M. (2011). Provision for linguistic diversity and linguistic minorities in India (Master's Thesis, Applied Linguistics, St. Mary's University College, Strawberry Hill, London).
- Zhang, Q., & 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.
- Zhenhao, G. (2015). Improved accent classification combining phonetic vowels with acoustic features. In *Eighth international congress on image and signal processing (CISP)* (pp. 1204–1209).
- Ziedan, R., Micheal, M., Alsammak, A., Mursi, M., & Elmaghraby, A. (2016). A unified approach for arabic language dialect detection. In *Twenty ninth international conference on computers applications in industry and engineering (CAINE)* (pp. 165–170).
- Zissman, M. A., Gleason, T. P., Rekart, D. M., Losiewicz, B. L. (1996). Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech. In *Acoustics, speech, and signal processing, ICASSP* (Vol. 2, pp. 777–780).
- Zue, V., Seneff, S., & Glass, J. (1990). Speech database development at MIT: TIMIT and beyond. *Speech Communication*, 9(4), 351–356.