Arab J Sci Eng (2018) 43:4289–4302
https://doi.org/10.1007/s13369-017-2941-0

RESEARCH ARTICLE - SPECIAL ISSUE - COMPUTER ENGINEERING AND COMPUTER SCIENCE

Dialect Identification Using Spectral and Prosodic Features on Single and Ensemble Classifiers

Nagaratna B. Chittaragi[1,2] · Ambareesh Prakash[3] · Shashidhar G. Koolagudi[1]

Received: 3 May 2017 / Accepted: 26 October 2017 / Published online: 17 November 2017

Abstract In this paper, investigation of the significance of spectral and prosodic behaviors of speech signal has been carried out for dialect identification. Spectral features such as cepstral coefficients, spectral flux, and entropy are extracted from shorter frames. Prosodic attributes such as pitch, energy, and duration are derived from longer frames. IViE (Intonational Variations in English) speech corpus covering nine dialectal regions of British Isles has been considered, to evaluate the proposed approach. Since corpus is available in both read and semi-spontaneous modes, the influence of spectral and prosodic behavior over these datasets is distinguishably articulated. Further, two distinct classification algorithms, namely support vector machine (SVM) and an ensemble of decision trees along with the SVM are used for identification of nine dialects. Dialect discriminating information captured from both features are used for constructing feature vectors. Experiments have been conducted on individual and combinations of features. A better dialect recognition performance is observed with ensemble methods over a single independent SVM.

✉ Nagaratna B. Chittaragi
nbchittaragi@gmail.com

Ambareesh Prakash
ambareesh.prakash@gmail.com

Shashidhar G. Koolagudi
koolagudi@nitk.ac.in

[1] Department of Computer Science and Engineering, National Institute of Technology Karnataka, Surathkal 575025, India

[2] Department of Information Science and Engineering, Siddaganga Institute of Technology, Tumkur, India

[3] Department of Mechanical Engineering, National Institute of Technology Karnataka, Surathkal, Karnataka 575025, India

Keywords Dialect recognition · Spectral features · Ensemble learning · Random Forests · Extreme random Forests · Gradient boosting

1 Introduction

Dialects represent the unique pronunciation patterns of a language, spoken among the community of speakers who belong to a particular geographical area. Dialects mainly exhibit grammatical, phonological, and prosodic differences among them. Pronunciation variations that occur in an individual's speaking styles are influenced by many surrounding factors related to the speaker, such as socioeconomic status, cultural background, geographical locations, education [1].

State-of-the-art systems that are capable of characterizing and identifying dialects would supply valuable inputs in the process of improving the performance of interactive speech systems. Dialectal traits are important factors in degrading the performance of automatic speech recognition (ASR) and human–computer interaction (HCI) systems [2]. Automatic dialect identification (ADI) can enhance the performance of ASR and HCI systems. Automation of dialect processing helps in the characterization of pronunciation patterns from socio-linguists [3]. In forensics, the tasks such as speaker identification, characterizing speaker traits, and speaker profiling are benefited by ADI systems [4]. ADI systems are helpful in developing robust speech systems, for portable devices. ADI can be used as an interpreter in call centers for an effective region-based customer call attention. Dialect identification is also useful in native language identification, medical applications, indexing of historical spoken documents and their retrieval, entertainment media, and so on [5].

ADI can be considered as a particular case of Language Identification (LID) problem, which is now drawing the

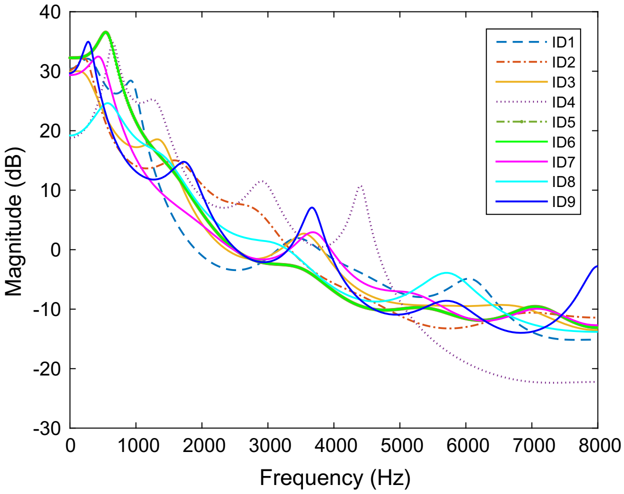

Fig. 1 Average LP spectra for utterance of "A" from female speakers of 9 dialects

attention of researchers from speech and language processing communities. Despite several approaches to design a robust ADI systems, it is considered as a challenging task, since ADI involves identification of the dialects within the same language class. Hence, the majority of the models developed are constrained to be language dependent. Applying these models to other languages may not be conducive, as there are many fundamental differences across languages regarding pronunciation patterns, phonology, grammatical structure, etc. [6,7]. Unavailability of standard datasets and difficulty in identifying the boundaries between the dialects are other challenges.

Several authors have proposed the use of variations in pronunciation of phonemes, consonants, or syllables across the dialects for characterization [8]. These variations have been explored successfully by analyzing acoustic–phonetic features that are represented with spectral and prosodic characteristics [9]. For instance, the spectral changes observed in the pronunciation of a vowel "A" by female speakers of nine different British English are shown in Fig. 1. This represents the averaged LP spectra for different dialects. Different energy levels, the width of spectral peaks, sharpness and position of formants are observed among nine dialects. These variations indicate the influence of spectral features in dialect identification. These features characterize the vocal tract by considering its shape and size during the production of various sound units. Speech cannot be simply defined as a collection and concatenation of audio units. Rather, the imposition of prosodic variations such as intensity patterns, intonations, and duration of the sound units add naturalness to the uttered speech [10]. The change in pitch, duration, intonation, rhythmic pattern, and stress features render a kind of melody to speech. A unique pattern of pronunciation style is observed in several dialects of any language.

Spectral and prosodic features have been considered for dialect recognition in this work. Performance of single learned SVM and ensemble approaches are compared. Spectral features such as Mel frequency cepstral coefficients (MFCC), shifted delta coefficients (SDC), spectral flux, and spectral entropy are extracted since they are usually involved in modeling the pronunciation variations. Prosody differences among dialects are captured through energy, duration, and pitch features. IViE speech corpus with nine dialects of British English is used for evaluation. Experiments have been conducted with single SVM classifier, three tree, and SVM-based ensemble classification techniques. Cross-dataset results and analysis are performed over possible combinations of two types of datasets (read and semi-spontaneous).

The remaining paper is organized as follows: Sect. 2 gives details of the existing work in dialect identification. Brief details of IViE speech corpus and various features extracted for dialect identification are given in Sect. 3. Section 4 describes the implementation details of ADI systems. Section 5 discusses the experiments, results and performance analysis. Section 6 concludes with an appropriate summary of the present work and further research directions.

2 Literature Review

In this paper, existing literature on automatic processing of dialects has been reviewed with respect to language models, acoustic–phonetic methods, phonotactic approaches, and classification techniques. These are discussed briefly in this section.

Few authors have observed that the boundaries between languages are distinct and have proposed the same for dialects as well, hence applied LID techniques for ADI [11,12]. Commonly, dialects are treated as subclasses of the languages and most common LID methods such as language modeling and phone recognition approaches are applied to achieve significant results [6,13]. In contrast, many other studies have been suggested that a majority of dialects cannot be treated as independent languages, as differences between dialects are very close unlike languages. Informally, there is a lot of overlap among the dialects [1]. Since dialects belong to the same language, they basically use the same vocabulary, syntax, and semantics, whereas variations may be observed only in pronunciation patterns and phonology, and a slight variation is observed in grammar. Hence, approaches applied for LID may not be suitable straight away for ADI [14]. Many studies on dialect processing are proposed at the sentence, word, syllable, or phoneme levels. Text-dependent and text-independent scenarios have been considered with read and spontaneous speech for dialect processing [15,16]. Further, it has been identified that dialect recognition problem has been addressed most commonly with acoustic–phonetics [12,15,17] and phonotactic methods [8,18,19].

Spectral acoustic differences existing among dialects have been studied extensively by extracting MFCCs and SDC features. Both spectral cues and temporal variations are captured and evaluated for classification of dialects with Gaussian mixture models (GMM) [7,11,20]. Significant dialect recognition performance is reported with an addition of i-vectors to MFCC-SDC features [11,21]. Feature i-vector is obtained through a data-driven approach, which includes mapping of a sequence of frames of speech onto a fixed low-dimensional vector space, called total variability space [22]. Kullback–Leibler divergence-GMM (KLD-GMM)-based methods are applied to obtain the most discriminating GMM mixtures for dialects. Further, frame selection decoding (FSD) is used to enhance the classification accuracy by avoiding confusing acoustic regions [14].

Several attempts are being made in the literature to use prosodic features, since they contribute additional features for dialect recognition. The variations in prosodic features are extracted from duration, pitch, and energy contours of the phoneme, syllable, word, or pseudo-syllables [23,24]. Pseudo-syllables represent the patterns matching the $C \times V$ structure that includes a cluster of optional consonants with a single vowel segment.

A majority of phonotactic approaches are suitable for dialect identification if transcriptions of the utterances are available. Unavailability of transcriptions makes the dialect identification a challenging task [25]. Many authors focus on using phone recognition language modeling (PRLM), parallel PRLM (PPRLM), and parallel phone recognition (PPR) models for dialect processing. PPRLM approach is applied for recognizing four colloquial dialects with modern standard Arabic (MSA) dialects [3,18].

Literature has many references to ADI systems that are being addressed using generalized classifiers such as GMM, hidden Markov model (HMM), SVM, artificial neural networks (ANN). Among all, GMM-based models are widely seen in many cases as baseline systems. GMMs aid in typically modeling the standard acoustic properties that are supposed to be normally distributed and uncorrelated [7,8,15,20]. Later, the GMMs combined with universal background model (UBM) are used for dialect classification in the case of multiple languages [9,22]. The combinations of local and global prosodic features extracted from four Arabian dialects are proposed for dialect classification using HMM with GMMs [8].

SVM models are found to be very powerful prediction and classification schemes, designed for handling high-dimensional input spaces. These models provide an opportunity for working with a large feature representation of speech [26]. MFCC features combined with prosodic features to form large dimensional feature vector and have been used with SVM for classifying dialects in the Hindi, an Indian language [23]. MFCC features are used in classifying dialects of American English with SVM hyperplanes [27]. A GMM-SVM hybrid classifiers are have been proposed for classifying three dialects of Spanish. Experiments are 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 [28]. 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 [29].

Recently, the concept of combining multiple classifiers is being proposed for the improvement of the performance over individual classifiers. Very few attempts can be found addressing dialect identification problems with ensemble techniques [12,30]. Rotation forest an ensemble of decision trees has been used to explore robustness issues among all dialects [9]. Similarly, AdaBoost ensemble algorithm has been used in word-based dialect identification problem by applying in the probability space, rather than the features space [12]. These two methods have reported a significant improvements when compared with single classifiers. The majority of works have been found and are relying mainly on use of n-gram features for identification of dialects from text-based datasets for natural language processing [31].

3 Details of Speech Corpus and Feature Extraction for Dialect Identification

3.1 IViE Speech Corpus

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 dataset was recorded from nine urban varieties. The experimenters introduced the complete process of recording to the speakers before the start of recording. In the interactive tasks, the subjects/speakers spoke to each other. Recording was done in both read and semi-spontaneous mode from 12 subjects (6F + 6M adolescents) representing each dialect [32]. Both read and semi-spontaneous datasets are available separately. In the rest of the paper, the semi-spontaneous dataset is referred as a semi-read dataset. Details of IViE corpus are given in Table 1.

3.2 Feature Extraction for Dialect Identification

In the present work, dialect discriminating characteristics are captured through the spectral and prosodic cues. Preprocess-

Table 1 Details of IViE dataset used

Sl. no.	Region	Dialects	Read mode (in min)	Semi-read mode (in min)
1	Belfest	ID1	52	32
2	Bradford	ID2	49	31
3	Cardiff	ID3	49	35
4	Cambridge	ID4	51	37
5	Dublin	ID5	48	33
6	Leeds	ID6	51	31
7	Liverpool	ID7	48	26
8	London	ID8	50	38
9	Newcastle	ID9	53	31
Total duration			∼8 h	∼5 h

ing is carried out to remove longer silence regions from the input audio [33].

3.2.1 Spectral Features

Dialect-relevant cues exist in the behavior of the vocal tract system concerning specific sequences of vocal tract shapes for a speech utterance [23]. The changes in the shape of the vocal tract occur due to pronunciation variations. These have been observed in the utterance of similar linguistic units (vowels, consonants, syllable, word, and sentence) of different dialects. Articulatory configuration corresponding to sound unit, co-articulation effect, the context of the units, gender variability, and emotional status are also important reasons of acoustic variability [34].

MFCCs The human auditory system follows the nonlinear model to process speech signal. It is mentioned in the literature that lower frequency components of speech signal always carry most of the phonetic information. Hence, a nonlinear mel-scale filter is used to weigh a lower frequency components [35]. In this work, RASTA (Relative Spectra) processed MFCC features (12 + 1 frame energy) with 13 delta and 13 delta-delta values representing SDC features are extracted [36]. MFCC features are derived from 40 filter banks. RASTA filter-based processing helps in suppressing noisy portions of the speech. The details of the RASTA filtering process is covered in this paper by excluding other regular steps of MFCC feature extraction. The RASTA filter is a band-pass filter that uses the transfer function H_{RASTA},

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

In this filter, each log filter-bank magnitude component $f[m,i]$, where i takes values $1, 2, \ldots N$ (No. of channels) is filtered by using $H_{\mathrm{RASTA}}(z)$ function and producing the RASTA-filtered log filter-bank magnitudes $f_{\mathrm{RASTA}}[m,i]$. 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 [37].

Spectral Flux The spectral change between two successive frames is measured as spectral flux. It is possible to distinguish between two sounds, whether they are similar or not based on loudness and pitch, through auditory sensation feature called timbre. The spectral flux is commonly used for measuring timbre of the utterance. The quick changes in the power spectra of a signal are measured through spectral flux. Flux is calculated by comparing the power spectra of one frame against the same of the previous frame [33]. Following equation is used to estimate the flux.

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

where $EN_i(k) = \frac{X_i(k)}{\sum_{l=1}^{Wf_L} X_i(l)}$, here $EN_i(k)$ is the kth normalized DFT coefficient at the ith frame, Wf_{L} is the frame size.

Spectral Entropy The short-term entropy of energy, measures the variations occurring in the energy profile of a speech signal. Since every dialect directly shows differences in intonation, a correlation between pitch and energy profile helps to discriminate dialects [33].

First, a division of spectrum of a frame into L sub-bands (bins) is done. The energy E_f of the fth sub-band, $f = 0, \ldots, L-1$ is calculated and each 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 is then calculated with the following Eq. (3)

$$H = -\sum_{f=0}^{L-1} nf \cdot \log_2(nf) \quad (3)$$

where $L = 10$, i.e., each frame is divided into 10 bins for computing spectral entropy in this work.

3.2.2 Prosodic Feature

Features at frame level represent limited local information present in the signal. However, some features need to be extracted from the longer span of the speech, to represent changes within and among the sequence of sound units. Intonation, intensity patterns, and the different speaking rate features induce the naturalness to the speech during a conversation [38]. These represent the prosodic cues and assists in exploiting the particular speaking patterns of each dialect. Usually, pitch, intensity variations, stress patterns, and rhythmic production are explored to represent the prosodic attributes. These additional features of the speech units contribute additive information for identification of dialects.

Spontaneous and read speech have demonstrated significant variations in both acoustic and linguistic properties. These are observed even with two types of speech datasets available in IViE corpus. Spontaneous speech exhibited the existence of clear prosodic cues such as varying rate of speaking, corrections, filled pauses, repetitions, use of partial words during the pronunciation [39], whereas read speech of all nine dialects of English has shown rich spectral information [40]. Hence, in this work, the focus is given on exploring both spectral and prosodic features.

Pitch Pitch also known as fundamental frequency (F0) represents the perceptual property of sound, commonly described as a perception of the relative altitude of sound. Fundamental frequency represents the physical correlate of the pitch and the rate of vibration of the vocal fold. However, a different F0 is perceived among the speakers of dialects. The auditory pitch perception influenced by the harmonic structure and amplitude plays a role in distinguishing various pronunciation styles of nine dialectal regions [38,41]. The subharmonic-to-harmonic ratio-based pitch estimation algorithm is used in this paper by which, accurate pitch perception results and reduction in gross error rate is achieved [42].

Energy Both the dialect speech corpora have reflected variations in energy profile since every dialect follows varying stress patterns during pronunciation. Hence, frame level energy is considered as a feature for identification of dialects. Equation (4) is used for calculating energy.

$$E(i) = \sum_{n=1}^{w_\mathrm{L}} |x_i(n)|^2 \tag{4}$$

here $x_i(n)$, $n = 1, \ldots, W_\mathrm{L}$ be the audio samples in the ith frame, where W_L is the length of the frame.

$$E(i) = \frac{1}{W_\mathrm{L}} \sum_{n=1}^{w_\mathrm{L}} |x_i(n)|^2 \tag{5}$$

Energy is normalized by dividing it with W_L to remove the dependency on the frame length.

3.3 Post-processing of Features

The process of feature extraction and post-processing of features carried out in this work is shown in Fig. 2. Single feature vector has been constructed for a speech signal for further processing. Initially, the speech signal is divided into M frames where each frame is of length 20 ms with a 10 ms overlap. The spectral features that are mentioned above have been computed at every frame forming a $M \times N$ dimensional matrix, where M represents the number of frames, and N represents the length of feature vector. Further, midterm processing is carried out to represent the statistical values of the feature vectors, rather than considering the derived feature vectors. Midterm processing is done for every two consecutive frames F_i and F_{i+1} to compute statistical mean and standard deviation. Due to this, the dimensionality will be doubled (2N) for each frame. Further, the mean of all M frames has been taken to form a single feature vector of size 2N. Later, the time taken for the complete utterance of each sound clip is appended as a duration feature (ms) to prosodic and combined features. At the end, feature vector size is $82((39 + 1 + 1) \times 2)$ for spectral features, sizes of prosodic and combined features are $5((2 \times 2) + 1)$ and $87((43 \times 2) + 1)$ respectively when duration feature is appended.

4 Dialect Identification System Using Spectral and Prosodic Features

In this work, spectral and prosodic behaviors and their combinations are used for implementing dialect identification system. A single classifier-based multi-class SVM, three tree-based ensemble methods (viz: Random Forest (RF), Extreme Random Forest (ERF), Extreme Gradient Boosting (XGB)) and bagging classifier based on SVM (BSVM) are used. The primary focus of the present work is to examine the behaviors of single and ensemble classifiers for designing dialect identification using spectral and prosodic features.

In general, individual classifiers use statistical methods to estimate class-conditional probabilities. Later, they are converted into posterior probabilities. SVM is a powerful pattern classification method that separates the classes by constructing a hyperplane. SVMs are designed to work on high-dimensional input spaces, for example, language and dialect classification problems [43], whereas in ensemble classifiers the prediction results of multiple base models are combined due to which accuracy is expected to increase [44]. Instead of relying on decisions by a single expert (base learner), they attempt to decide by utilizing the collective input from a committee of experts. Either independent or dependent approach is followed for the selection of appropriate base models. Bagging algorithms combines predictions from independent base models derived from bootstrap samples of the original data [45]. Usually boosting algorithms, follow dependent fashion in the growth of ensembles. Base models are improved iteratively, depending on training to reduce the errors of the current ensemble [46]. Figure 3

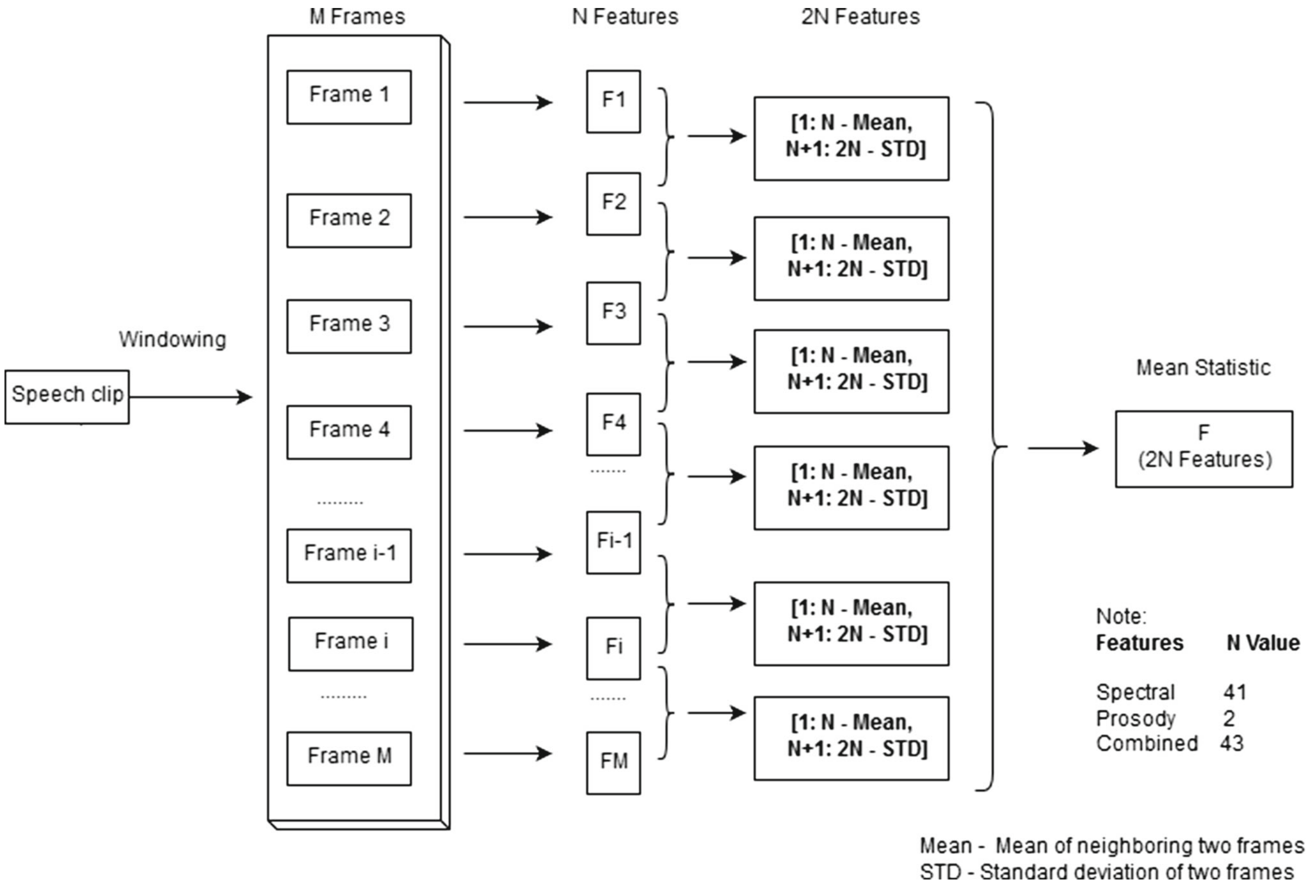

Fig. 2 Statistical midterm processing of short-term features

describes the details of the general workflow of single and ensemble classification methods.

4.1 SVM-Based Dialect Identification System

In the proposed work, SVMs are trained on nine dialects with one-versus-rest approach to handle 9-class pattern classification into nine two-class classification problems. SVM uses a linear separating hyperplane with the maximal margin between support vectors using linear kernel function [47]. Dialectal cues at spectral and prosodic levels are extracted separately and nine SVM models are trained with individual and combination of features. Training inputs from all nine dialect 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 nine labels $k = 1, 2, \ldots 9$. All these are used to train the SVM model for nine classes. The SVM for the dialect class k is constructed 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 representing the negative example. For evaluation of the developed dialect identifica-

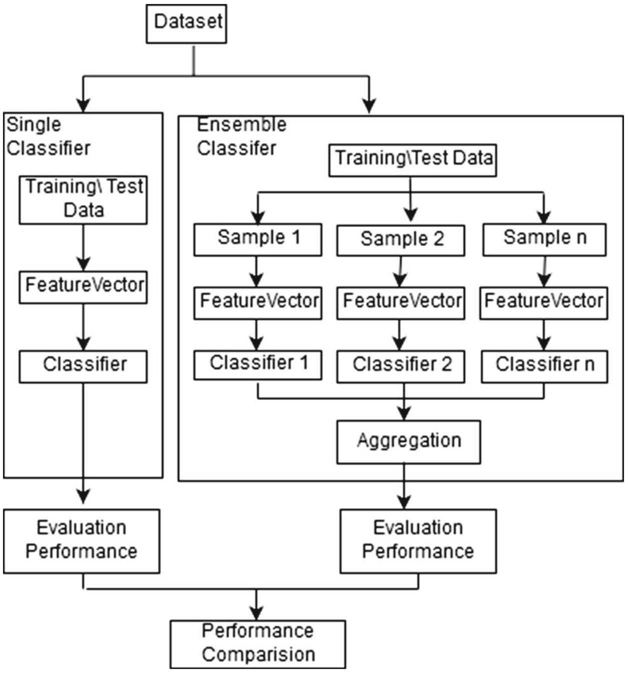

Fig. 3 Workflow of single vs. ensemble classifiers: derived from the work [26]

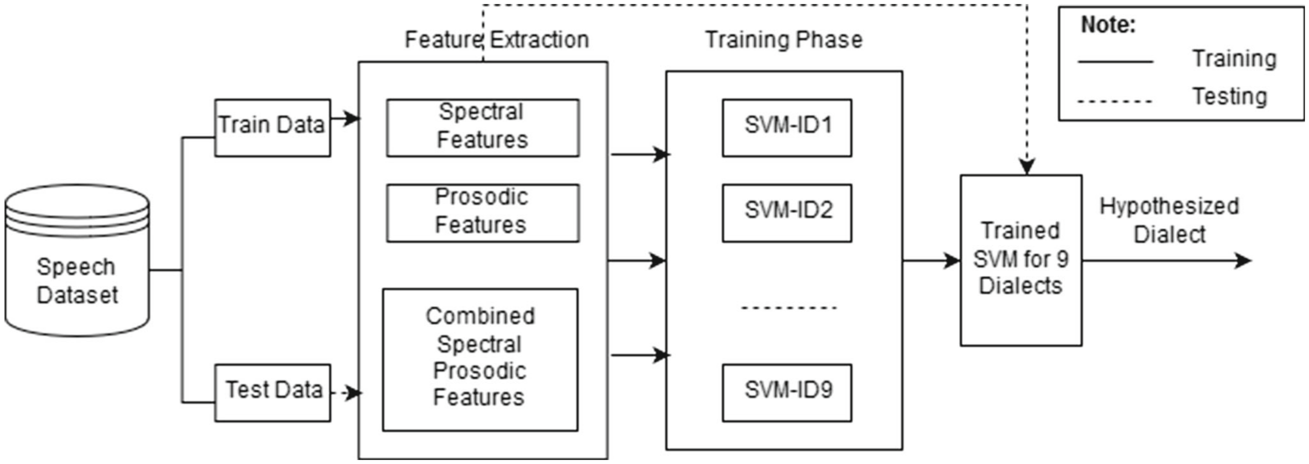

Fig. 4 Dialect recognition system with SVM

tion system, feature vectors derived from test speech samples are given to all SVM models. For a given test pattern x, the evidence $TS_k(x)$ is obtained from all nine 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., $C(x) = \text{argmax}_k(TS_k(x))$. The block diagram of proposed dialect identification system with SVM is given in Fig. 4.

4.2 Ensemble Learning-Based Dialect Identification System

Recently, ensemble methods resulting from combining the predictions of several classifiers have proven to be the most successful approaches for speech recognition tasks. However, use of these state-of-the-art techniques, for dialect identification is rarely observed. In this work, both bagging (RF, ERF, BSVM) and boosting (XGB)-based methods are used in designing dialect identification systems. For each feature type, a separate ensemble system is developed in two stages. In the first stage, evidences are obtained for all models under three configurations in four scenarios (see Table 2). In the second stage, the proof of individual ensemble component models is compared with each other [26,48].

Decision tree-based RF classifier is implemented in this work, with a forest that includes 2048 decision trees constructed by bootstrapping approach (samples are drawn randomly with replacement) from the training dataset. Empirical analysis has yielded better accuracy with the use of 2048 decision trees with IViE speech corpus. During the construction of the tree, splitting a node is controlled by picking best split decided by Gini criterion among a random subset of features.

$$\text{Gini} = N_L \sum_{k=1\ldots K} p_{kL}(1 - p_{kL}) + N_R \sum_{k=1\ldots K} p_{kR}(1 - p_{kR}) \tag{6}$$

Table 2 Four combinations (four scenarios) of two different IViE speech dataset, R—read dataset, S—semi-read dataset

Sl. no.	Type	Description
1	RR	Disjoint sets of read dataset are used in training and testing (80:20)
2	RS	Complete read dataset for training and a semi-read dataset for testing
3	SS	Disjoint sets of semi-read dataset are used in training and testing (80:20)
4	SR	Complete semi-read dataset for training and read dataset for testing

where p_{kL} represents the portion of the k class in a left node of the tree, p_{kR} represents the portion of in right node of the tree, N_L and N_R indicates the number of nodes in left and right part of a tree. $\sqrt{n}$ features are used to split a node, where n is the size of feature vector [49,50]. When the complete forest is constructed with decision trees, classification is done by combining voting from predictions of different trees trained on different parts of the training set.

An ERF is a small variant of RF, that differs in the way of the splitting of trees. Forest is constructed with 2048 trees by sub-sampling done with replacement. Instead of using an optimized split for tree construction as in RF method, ERF chooses the best out of randomly generated thresholds for each candidate feature. In this model also, maximum features parameter is selected as $\sqrt{n}$, where n is the size of feature vector [49,51].

XGB Boosting iteratively improves the base learner prediction in a greedy fashion, such that each additional base learner improves the accuracy by further reducing the

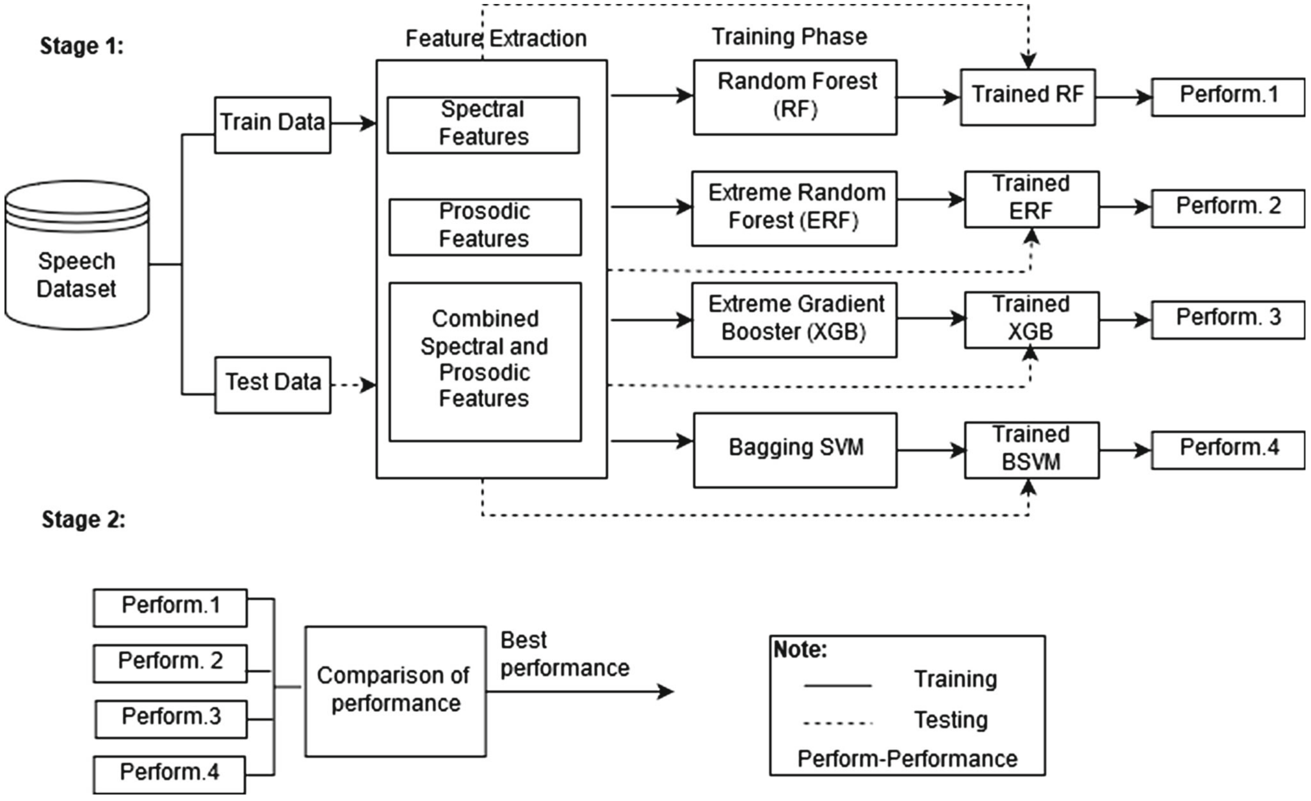

Fig. 5 Ensemble-based dialect recognition system

selected loss (error) function. Multi-class logloss function is used in this work.

$$\text{logloss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j}\log(p_{i,j}) \qquad (7)$$

where N is size of 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.

In this work, decision tree classifier is used as the base learner. Decision trees are constructed as follows: η (eta) representing learning rate parameter is assigned with 0.2, which controls the shrinking of feature weight to make the boosting process more conservative. Maximum depth of a tree is limited up to 6, subsample ratio of the training inputs are limited, such that 0.6 of data instances are used to grow trees. Objective function softmax for handling nine classes is used in this work. These few parameters are fine-tuned get better recognition accuracy. The XGBoost library is used for implementation [52,53].

In addition, ensemble technique with SVM as a base learner is implemented for dialect identification system using bagging classification function. The bootstrap aggregating (bagging) method similar to that of RF is used with 2048 SVM classifiers as the weak learner instead of decision trees. It fits the base classifiers each on random subsets obtained from the dialect dataset and final predictions are computed by averaging all classifiers [49,54]. The block diagram of proposed dialect identification system using three decision tree and SVM-based ensemble methods is given in Fig. 5.

5 Experiments, Results and Discussion

In this work, the performance is evaluated with SVM and four ensemble classifiers under three configurations. Since the IViE dataset is small and available in read (8 h) and semi-read mode (5 h), every experiment is carried out in four different scenarios to realize the influence of spectral and prosodic features with both speech corpora. Details are given in Table. 2.

For evaluating the developed ADI systems, feature vectors derived from test speech samples are given to all the ADI systems. The model giving highest evidence is hypothesized as the dialect corresponding to the given speech sample.

5.1 Performance Evaluation Using SVM

Nine SVM models trained on the datasets of nine English dialects are used to handle 9-class pattern classification problems with extracted feature vectors.

Table 3 Dialect recognition performances using SVM-based systems

Features	Recognition accuracy in %			
	RR	RS	SS	SR
Spectral	86.33	67.44	85.59	74.13
Prosodic	76.25	39.19	61.09	34.33
Spectral + prosodic	90.93	69.14	89.9	76.00

Table 4 Performance of SVM-based dialect identification using prosodic features: ID1:ID9 9 dialectal regions. Average recognition performance: 39.19% in RS scenario

	Confusion matrix for prosodic features								
	ID1	ID2	ID3	ID4	ID5	ID6	ID7	ID8	ID9
ID1	31	0	16	0	4	0	11	3	7
ID2	3	35	2	2	6	1	6	17	0
ID3	0	0	62	0	0	7	2	0	1
ID4	2	7	4	5	3	0	0	10	41
ID5	15	5	3	0	5	1	27	9	7
ID6	9	10	16	1	10	9	8	9	1
ID7	0	6	0	0	6	0	62	0	0
ID8	1	10	5	3	10	1	26	9	5
ID9	6	0	15	2	0	1	7	4	36

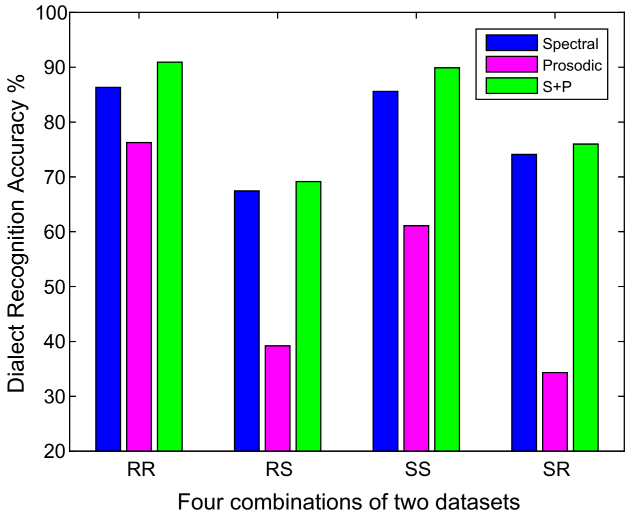

Fig. 6 Comparison of dialect recognition performance of spectral, prosodic and the combined, S + P: spectral + prosodic features, RR: read–read, RS: read–semi-read, SS: semi-read–semi-read and SR: semi-read–read

Series of experiments are conducted in four scenarios with 5-fold cross-validation approach. Table 3 shows the recognition performances obtained for nine dialects in four scenarios. Spectral features have shown better significance in discriminating dialects with all scenarios when compared to the prosodic features. The accuracy is comparatively good and almost similar when training and testing are carried out with same data set (RR and SS). With these two scenarios, in training phase, the machine has learned the complete patterns of pronunciation styles of both read and semi-read speech of all dialects. When samples with similar patterns are tested, it is leading to a better accuracy of 90.93% in RR scenario. The comparative results obtained using SVM classifier is shown in Fig. 6.

Read and semi-read speech is significantly different in acoustics and linguistics [40]. Read speech is a well pronounced controlled speech, indicating clear and correct phoneme pronunciation in the acoustic space [39]. Whereas, semi-read speech is with varying filled pauses, change in the rate of speech, and with partial words pronunciations. Due to these varying properties of two types of speech, reduction in accuracy is observed when training and testing sets are different (RS and SR). RS scenario shows a lower performance, due to the absence of specific properties of semi-read speech during the training phase. Whereas, all such attributes exist in testing patterns of semi-read speech. Also, due to spectral space shrinkage in the semi-read speech, the individual phoneme recognition accuracy is reduced and hence the reduction in overall performance.

Prosodic features such as pitch, energy, and duration are used in this study. Table 4 shows the details of individual dialect identification performance with the help of confusion matrix for RS scenario. Results indicate that explored prosody features are not significant as the spectral features in all scenarios for dialect identification. The performance achieved is only 39.19% with prosodic features in RS scenario. ID3 and ID7 dialectal regions have been classified with the better classification of 86.11% (62 out of 72 testing samples) for prosodic features. Speakers representing these two regions have shown the unique pronunciation patterns and styles with high influence of pitch and energy features, when compared to other dialects [32], whereas dialectal regions representing ID4, ID5, ID6, and ID8 have shown high misclassification rate, indicating the lesser influence of considered prosodic features. ID4 (more than 50%) is highly misclassified with ID9 dialect, similarly, ID8 is misclassified with ID7 dialect.

To understand the influence of pitch on dialects, the average pitch values of male and female speakers are considered and shown in the form of a box plot given in Fig. 7. Box plots are chosen as they represent detailed statistics of the distribution of data based on the five statistical measures namely median, minimum, maximum, first quartile and third quartile. A rectangle spans between first quartile and third quartile typically called as interquartile range, and it is observed to be large in all dialects except ID7 and ID9. Spans of first quartile and third quartile are not exactly divided by median value. Outliers that represent the high maximum and low minimums are found in ID7

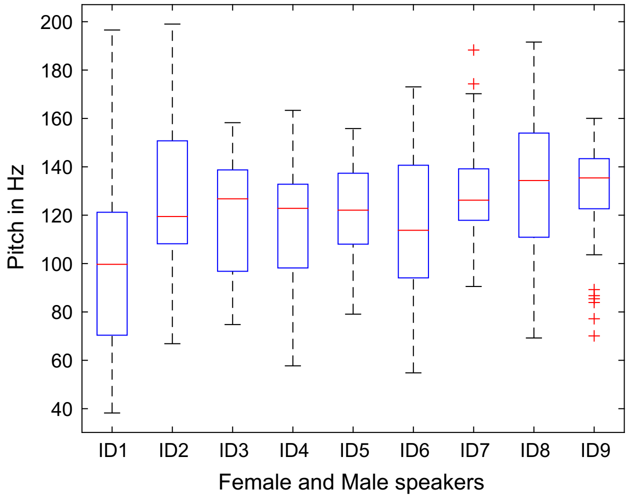

Fig. 7 Statistical range for Pitch values of all 9 dialects

and ID9 dialect speakers. Minimum and maximum values are distinct in all dialects. These patterns have shown the significance of prosodic features in discriminating the dialects. Prosodic information many times behaves complimentary to spectral information. Hence, combining spectral features with prosodic information increases the accuracies. Since all dialectal regions are from British English, they all most share similar speaking patterns with minor variations. Use of single SVM classifier has failed in identifying these minor differences effectively that exist among dialects [32,55].

5.2 Performance Evaluation Using Ensemble Classification Techniques

From the literature, it is observed that ensemble of various classifiers or learners work better rather than single classifiers [48]. In this regard, ADI systems are evaluated to understand the behaviors of three ensemble methods using decision tree base learner and ensemble method with SVM as a base or weak learners [48]. Each of these is evaluated using three different configurations namely simple validation (SV), cross-fold validation (CV) and hold-out voting (HOV).

Generally, cross-validation makes the model more stable by considering the variations across the dataset. A better approximation is achieved as it trains and tests on every part of the dataset by typically taking care of the expected prediction error. Similarly, CV is performed in this paper by considering the 4 out of 5 ($k = 5$) folds for training purpose and left over one fold for testing. Testing is carried out to observe the average behavior of the system each time by rotating the training and testing folds.

Further, to prevent the over fitting of hyper-parameters of all models on CV results and also to generalize the model, HOV approach is performed. Complete dataset is divided into 80:20 ratio, where 80% of data used for training the model and predicted the results for 20% of hold-out set. Similar procedure is followed and five predictions are being generated over five different combinations of 80% data. Final predictions for 20% data are computed by considering majority votes obtained out of all five predictions. HOV performances are referred for further analysis of the systems, as they are giving better accuracies, and the obtained results are considered to be more stable across the datasets than SV performances.

In simple validation with RR and SS scenarios, 80:20 ratio of the complete dataset is used for training and testing, respectively. The results obtained are considered to be less significant even though better results are obtained. Since the machine is biased and purely depends on samples involved in training and testing [49].

Results obtained from series of experiments conducted in this study, using decision trees and SVM-based ensemble methods are presented in Table 5 and 6 respectively. It is evident from the results that the spectral features such as MFCCs, flux and entropy have shown distinct values for all nine dialects of British English. Hence, spectral features have outperformed in all four ensembles and all scenarios. It is observed from the results that dialect recognition accuracy is comparatively high when training and testing is carried out with the similar data set (RR and SS) with both spectral and combined features. Among all ensemble methods, it is noticed that RR scenario with combined features has produced highest dialect recognition rate of 95.87% and 97.52% with ERF and BSVM ensemble methods, respectively. Results obtained have significantly proved the influence of both spectral and prosodic features among all nine dialect speakers of British English.

SS scenario has also given better recognition performance with spectral and combined features of 90% and 92.30% with XGB and BSVM ensemble methods, respectively. A slight reduction in accuracy is due to the fact that semi-read and read speech do not posses similar acoustic properties [39].

RS and SR scenarios yielded slightly less performances with spectral and combined features. Hence, these results suggest that even ensemble algorithms fail to recognize the dialect discriminating boundaries among the available feature vectors. Comparison of nine dialect recognition performance concerning spectral features is given in Fig. 8 and same for combined features in given Fig. 9.

From Table 5 and 6, among all ensemble models, it is noticed that RF, ERF, and XGB models have exhibited a degradation in performance for prosodic features in RS, SS and SR scenarios except in the RR scenario, whereas BSVM ensemble has shown the lower performance of about 33.59% for all scenarios (even with RR). Highest accuracy of about

Table 5 Dialect recognition rate performance in % three tree-based ensemble algorithms

Features	*Read-read scenario*								
	Random forest			Extreme random forest			Extreme gradient		
	CV	HOV	SV	CV	HOV	SV	CV	HOV	SV
Spectral	88.76	91.74	92.56	91.07	93.39	91.74	83.14	85.12	85.95
Prosodic	67.77	71.07	71.41	70.74	76.03	76.86	64.46	67.77	69.42
Spectral + prosodic	90.08	92.56	91.74	94.21	95.87	96.69	85.12	88.43	90.08
Read–semi-read scenario									
Spectral	63	62.96	64.97	66.57	67.74	67.59	67.31	68.83	69.59
Prosodic	40.83	41.20	41.36	42.99	43.67	44.59	42.16	43.05	41.97
Spectral + prosodic	64.19	65.28	64.97	69.25	69.44	71.45	67.41	66.97	68.51
Semi-read–semi-read scenario									
Spectral	83.58	85.38	86.92	84	86.15	86.6	85.23	90	89.23
Prosodic	47.69	46.92	50.77	50.77	52.31	53.08	48.15	50.06	50.12
Spectral + prosodic	85.23	86.92	87.69	85.69	86.15	86.92	85.29	90	90.77
Semi-read–read scenario									
Spectral	78.64	79.61	80.92	79.20	81.26	81.09	75.85	76.78	75.45
Prosodic	45.33	46.26	47.26	46.16	46.60	47.76	44.34	46.13	46.43
Spectral + prosodic	81.82	83.41	83.75	82.55	84.57	85.07	77.94	78.44	79.27

CV cross-fold validation, HOV hold-out voting, SV simple validation

Table 6 Dialect recognition performance in % using SVM-based ensemble

Features	*Read–read scenario*		
	CV	HOV	SV
Spectral	94.38	95.98	96.66
Prosodic	32.75	33.59	34.29
Spectral + prosodic	95.71	97.52	97.26
Read–semi-read scenario			
Spectral	69.47	69.99	70.53
Prosodic	32.74	33.17	36.25
Spectral + prosodic	70.70	70.52	71.19
Semi-read–semi-read scenario			
Spectral	87.85	88.46	89.74
Prosodic	46.31	49.23	47.69
Spectral + prosodic	91.38	92.30	93.08
Semi-read–read scenario			
Spectral	77.74	79.43	79.61
Prosodic	30.41	30.18	30.51
Spectral + prosodic	80.13	81.42	81.76

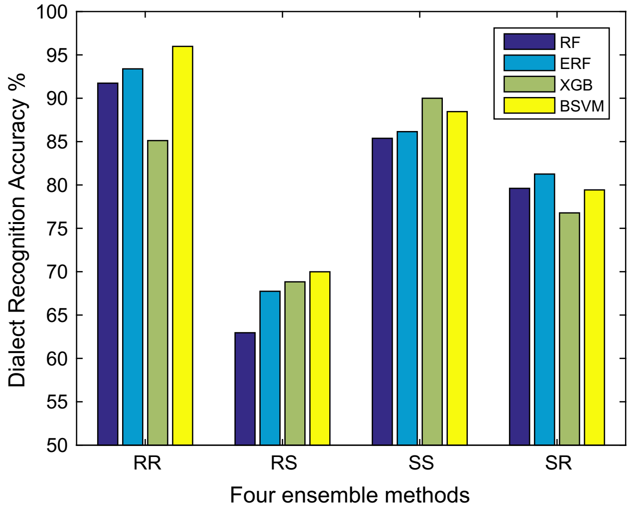

Fig. 8 Comparison of dialect recognition performance using spectral features

52.31% has been achieved with ERF method among RS, SS, and SR scenarios with prosodic features. Comparison of nine dialect recognition systems using prosodic features is given in Fig. 10.

Comparison of dialect recognition performances with respect to single classifier SVM and four ensemble methods are demonstrated for spectral, prosodic, and combined features in Fig. 11. It is observed that the use of ensemble approaches have reduced the training complexity and have shown high predictive accuracy over two datasets considered in this work [56], whereas single classifier SVM also gives better results due to the use of appropriate features, hyper parameter tuning and selection of suitable kernel function. Among all four ensemble methods, BSVM-based ensemble method has reported the highest accuracy of 97.52% for combined features and shown slightly better recognition performance than decision tree-based ensemble methods except for prosodic features in all four scenarios.

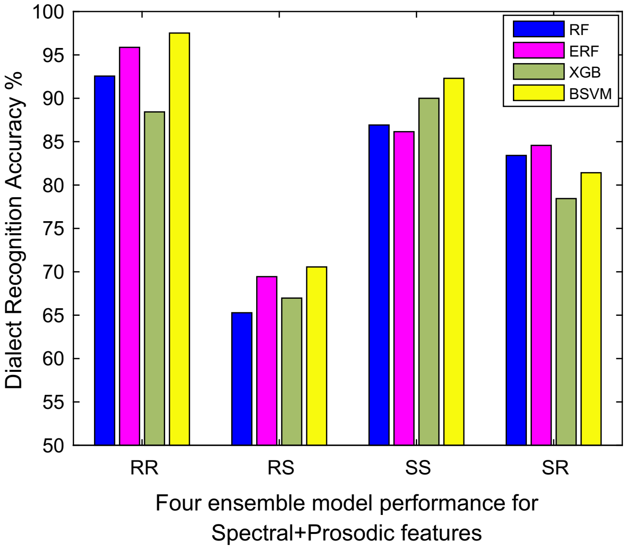

Fig. 9 Comparison of dialect recognition performance using both spectral and prosodic feature

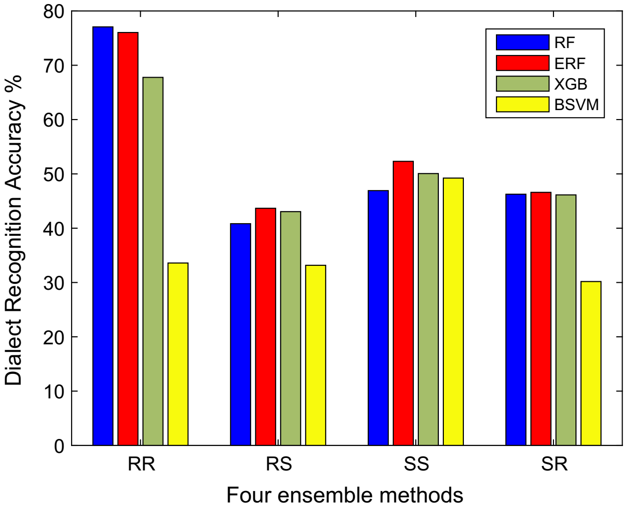

Fig. 10 Comparison of dialect recognition performance using prosodic features

6 Summary and Conclusions

In this paper, spectral and prosodic features are explored and analyzed individually and in combination from dialect discrimination perspective. Spectral features such as cepstral coefficients, SDCs, spectral flux, and entropy are extracted. Prosodic properties are also explored from the longer frame. In this paper, SVM, decision tree-based and SVM ensemble methods are used for developing the dialect identification system and results are verified with individual and combinations of the features. IViE speech corpus available in two modes is used to carry out experiments in four possible combinations of the dataset. Better dialect recognition is achieved with RR and SS scenarios. Spectral features have shown better performance with single classifier SVM and four ensemble methods. Selected prosodic features resulted in lesser significance with all dialects of British English. Single classifier SVM, decision tree and SVM-based ensemble methods have shown better recognition performance in RR scenario. Results have indicated the higher influence of spectral features with nine dialects and existence of minor non-overlapping behaviors with prosodic features proved with higher performances for combined features.

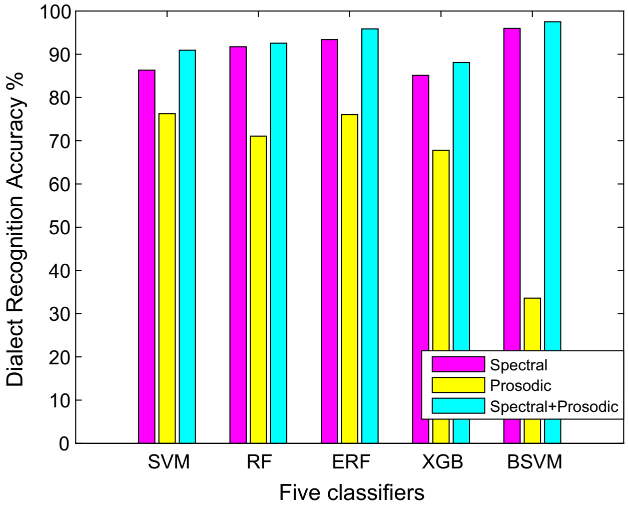

Fig. 11 Comparison of dialect recognition performance for RR scenario with all five classifiers

In future, language-specific dialect discriminating prosodic features can be explored. Prosodic cues among dialects exist in rhythmic patterns, stress and intonations can be further examined to make efficient dialect identification system. Further ensemble algorithms can be fine-tuned with specific parameters for efficient recognition of dialects. The ensemble of various classifiers can be implemented.

References

1. Chambers, J.K.; Trudgill, P.: Dialectology, 2nd edn. Cambridge University Press, Cambridge (1998)
2. Ferragne, E.; Pellegrino, F.: Automatic dialect identification: a study of British English. Speak. Classif. **II**, 243–257 (2007)
3. Chen, N.F; Shen, W.; Campbell, J.P: 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 (2010)
4. Harris, M.J.; Gries, S.T.; Miglio, V.G.: Prosody and its application to forensic linguistics. Ling. Evid. Sec. Law Intell. **2**(2), 11–29 (2014)
5. Gray, S.; Hansen, J.H.L.: An integrated approach to the detection and classification of accents/dialects for a spoken document retrieval system. In: Automatic Speech Recognition and Understanding, pp. 35–40 (2005)

6. Zissman, M.A.: Comparison of four approaches to automatic language identification of telephone speech. IEEE Trans. Speech Audio Process. **4**(1), 31–44 (1996)
7. Mehrabani, M.; Hansen, J.H.L.: Automatic analysis of dialect/language sets. Int. J. Speech Technol. **18**(3), 277–286 (2015)
8. Biadsy, F.: Automatic Dialect and Accent Recognition and its Application to Speech Recognition. PhD Thesis, Columbia University (2011)
9. Liu, G.A.; Hansen, J.H.L.: A systematic strategy for robust automatic dialect identification. In: 19th European Signal Processing Conference, pp. 2138–2141 (2011)
10. Sreenivasa Rao, K.; Yegnanarayana, B.: Modeling durations of syllables using neural networks. Comput. Speech Lang. **21**(2), 282–295 (2007)
11. Torres-carrasquillo, P.A.; Gleason, T.P.; Reynolds, D.A.: Dialect identification using Gaussian Mixture Models. ODYSSEY - The Speaker and Language Recognition Workshop, pp. 2–5 (2004)
12. Huang, R.; Hansen, J.H.L.; Angkititrakul, P.: Dialect/accent classification using unrestricted audio. IEEE Trans. Audio Speech Lang. Process. **15**(2), 453–464 (2007)
13. Zissman, M.A.; Gleason, T.P.; Rekart, D.M.; Losiewicz, B.L.: Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech. In: IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 777–780 (1996)
14. Lei, Y.; Hansen, J.H.L.: Dialect classification via text-independent training and testing for Arabic, Spanish, and Chinese. IEEE Trans. Audio Speech Lang. Process. **19**(1), 85–96 (2011)
15. Rouas, J.L.: Automatic prosodic variations modeling for language and dialect discrimination. IEEE Trans. Audio Speech Lang. Process. **15**(6), 1904–1911 (2007)
16. Chen, N.F.; Tam, S.W.; Shen, W.; Campbell, J.P.: Characterizing phonetic transformations and acoustic differences across english dialects. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(1), 110–124 (2014)
17. Sarma, M.; Sarma, K.K.: Dialect Identification from Assamese speech using prosodic features and a neuro fuzzy classifier. In: 3rd International Conference on Signal Processing and Integrated Networks (SPIN), pp. 127–132 (2016)
18. Shen, W.; Chen, N.; Reynolds, D.: Dialect recognition using adapted phonetic models. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 763–766 (2008)
19. Purnell, T.; Idsardi, W.; Baugh, J.: Perceptual and phonetic experiments on American English dialect identification. J. Lang. Soc. Psychol. **18**(1), 10–30 (1999)
20. Chen, T.; Huang, C.; Chang, E.; Wang, J.: Automatic accent identification using Gaussian Mixture Models. In: Automatic Speech Recognition and Understanding, IEEE Workshop, pp. 343–346 (2001)
21. Dehak, N.; Torres-Carrasquillo, P.A.; Reynolds, D.A.; Dehak, R.: Language recognition via i-vectors and dimensionality reduction. In: Interspeech, pp. 857–860 (2011)
22. Hansen, J.H.L.; Liu, G.: Unsupervised accent classification for deep data fusion of accent and language information. Speech Commun. **78**, 19–33 (2016)
23. Sreenivasa Rao, K.; Koolagudi, S.G.: Identification of Hindi dialects and emotions using spectral and prosodic features of speech. Int. J. Syst. Cybern. Inform. **9**(4), 24–33 (2011)
24. Etman, A.; Louis, A.A.: American dialect identification using phonotactic and prosodic features. In: SAI Intelligent Systems Conference (IntelliSys), pp. 963–970 (2015)
25. Biadsy, F.; Hirschberg, J.; Habash, N.: Spoken Arabic dialect identification using phonotactic modeling. In: Proceedings of the Workshop on Computational Approaches to Semitic Languages Conducted by Association for Computational Linguistics, pp. 53–61 (2009)
26. Utami, I.T.; Sartono, B.; Sadik, K.: Comparison of single and ensemble classifiers of support vector machine and classification tree. J. Math. Sci. Appl. **2**(2), 17–20 (2014)
27. Pedersen, C.; Diederich, J.: Accent classification using support vector machines. In: Computer and Information Science, 6th IEEE/ACIS, pp. 444–449 (2007)
28. Chitturi, R.; Hansen, J.H.L.: Multi-stream dialect classification using SVM-GMM hybrid classifiers. In: IEEE Workshop on Automatic Speech Recognition Understanding (ASRU), pp. 431–436 (2007)
29. Lachachi, N.E.; Adla, A.: Two approaches-based L2-SVMs reduced to MEB problems for dialect identification. Int. J. Comput. Vis. Robot. **6**(1–2), 1–18 (2016)
30. Darwish, K.; Sajjad, H.; Mubarak, H.: Verifiably Effective Arabic dialect identification. In: Empirical Methods in Natural Language Processing, pp. 1465–1468 (2014)
31. Malmasi, S.; Dras, M.: Language identification using classifier ensembles. In: Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects, pp. 35–43 (2015)
32. Grabe, E.; Post, B.: Intonational variation in the british isles. In: Speech Prosody, International Conference (2002)
33. Giannakopoulos, T.; Pikrakis, A.: Introduction to Audio Analysis: A MATLAB Approach. Academic Press, London (2014)
34. Reetz, H.; Jongman, A.: Phonetics Transcription, Production, Aoustics and Perception. Wiley Blackwell, New York (2009)
35. Tsai, W.H.; Chang, W.W.: Discriminative training of gaussian mixture bigram models with application to chinese dialect identification. Speech Commun. **36**(3), 317–326 (2002)
36. Hermansky, H.; Morgan, N.: Rasta processing of speech. IEEE Trans. Speech Audio Process. **2**(4), 578–589 (1994)
37. Kotnik, B.; Vlaj, D.; Kacic, Z; Horvat, B.: Robust MFCC feature extraction algorithm using efficient additive and convolutional noise reduction procedures. In: ICSLP, **2**, pp. 445–448 (2002)
38. Ramus, F.; Mehler, J.: Language identification with suprasegmental cues: a study based on speech resynthesis. J. Acoust. Soc. Am. **105**(1), 512–521 (1999)
39. Liu, G.; Lei, Y.; Hansen, J.H.L.: Dialect identification: impact of differences between read versus spontaneous speech. In: 18th European Signal Processing Conference, pp. 2003–2006. IEEE (2010)
40. Nakamura, M.; Iwano, K.; Furui, S.: Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. Comput. Speech Lang. **22**(2), 171–184 (2008)
41. Wightman, C.W.: Automatic detection of prosodic constituents for parsing. Doctoral dissertation (1992)
42. Sun, X.: A pitch determination algorithm based on subharmonic-to-harmonic ratio. In: The 6th International Conference of Spoken Language Processing, pp. 676–679 (2000)
43. Campbell, W.M.; Campbell, J.P.; Reynolds, D.A.; Singer, E.; Torres-Carrasquillo, P.A.: Support vector machines for speaker and language recognition. Comput. Speech Lang. **20**(2), 210–229 (2006)
44. Paleologo, G.; Elisseeff, A.; Antonini, G.: Subagging for credit scoring models. Eur. J. Oper. Res. **201**(2), 490–499 (2010)
45. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
46. Freund, Y.; Schapire, R.: A short introduction to boosting. J. Jpn. Soc. Artif. Intell. **14**, 771–780 (1999)
47. Chang, C.-C.; Lin, C.-J.: Libsvm: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**(3), 27 (2011)
48. Dietterich, T.G.: Ensemble methods in machine learning. In: International workshop on multiple classifier systems. Springer, pp. 1–15 (2000)

49. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
50. Friedman, J.; Hastie, T.; Tibshirani, R.: The Elements of Statistical Learning, Volume 1. Springer Series in Statistics. Springer, New York (2001)
51. Geurts, P.; Ernst, D.; Wehenkel, L.: Extremely randomized trees. Mach. Learn. **63**(1), 3–42 (2006)
52. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**, 1189–1232 (2001)
53. Chen, T.; Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
54. Kim, H.C.; Pang, S.; Je, H.M.; Kim, D.; Bang, S.Y.: Support vector machine ensemble with bagging. In: Pattern Recognition with Support Vector Machines: First International Workshop, pp. 397–408 (2002)
55. Grabe, E.; Post, B.; Nolan, F.: The IViE Corpus. Department of Linguistics. University of Cambridge, Cambridge (2001)
56. Marc, C.; De Frank, S.; Johan, S.; De Bart, M.: EnsembleSVM: a library for ensemble learning using support vector machines. J. Mach. Learn. Res. **15**, 141–145 (2014)