Robust Dialect Identification System using Spectro-Temporal Gabor Features

Nagaratna B. Chittaragi*, Siva Krishna P Mothukuri, Pradyoth Hegde, Shashidhar G. Koolagudi
Department of CSE, National Institute of Technology Karnataka, Surathal, India
*Department of ISE, Siddaganga Institute of Technology, Tumkur, Karnataka, India
Email: {nbchittaragi, msivakrish, pradyothhegde }@gmail.com, koolagudi@nitk.edu.in

***Abstract*—Automatic identification of dialects of a language is gaining popularity in the field of automatic speech recognition (ASR) systems. The present work proposes an automatic dialect identification (ADI) system using 2D Gabor and spectral features. A comprehensive study of the five dialects of a Dravidian Kannada language has been taken up. Gabor filters representing spectro-temporal modulations attempt in emulation of the human auditory system concerning signal processing strategies. Hence, they are able to well perceive human voices in tern recognize dialectal variations effectively. Also, spectral features Mel frequency cepstral coefficients (MFCC) are derived. A single classifier based support vector machine (SVM) and ensemble based extreme random forest (ERF) classification methods are employed for recognition. The effectiveness of the Gabor features for ADI system is demonstrated with proposed Kannada dialect dataset along with a standard intonation variation in English (IViE) dataset for British English dialects. The Gabor features have shown better performance over MFCC features with both datasets. Better recognition performance of 88.75% and 99.16% is achieved with Kannada and IViE dialect datasets respectively. Proposed Gabor features have demonstrated better performances even under noisy conditions.**

***Index Terms*—Dialect Classification, 2D Gabor features, MFCC, Kannada Dialects, SVM, Extreme Random Forest**

I. Introduction

A dialect is a unique pattern of a language spoken across a specific region followed among a social group. The dialectal variations of a language can be identified in the phonemes, syllables, words, sentences along with the unique style of speaking, etc. Dialects of a language may have formed due to the influence of neighbouring language, socio-economical changes, educational & cultural background and many other such reasons. However, every dialect shares common grammatical characteristics associated with the language [1].

Classifying dialects can be very useful in the improvement of the ASR systems. The dialect identification broadens the area of speaker identification and verification systems. One of the major application of dialect identification of a language is digital forensics field. Since, speaker profiling, speaker identification and verification tasks boost the criminal investigation process. So that they can derive nativity, cultural back ground, and the socio-economic group the person belongs to [2]. Voice controlled devices can enhance their recognition rate through identification of the dialects. Interactive voice response systems (IVRS) can be benefited from dialects of a language.

Majority of existing systems available for dialect identification are addressed using only spectral and prosodic features [3]. Recently, 2D Gabor (spectro-temporal) features have been considered extensively in development of speech processing applications. However, systems reported with use of these spectro-temporal features for classification of languages and dialects is decidedly less. Gabor features are said to model specific stimuli to which the neurons of the mammalian auditory cortex are sensitive. Both spectral and temporal modulation frequencies do exists in these stimuli. [4]. This paper investigates the use of Gabor features those are biologically-inspired features and derived from a filter bank of two-dimensional Gabor functions for classification of Kannada dialects. Likewise, vocal tract variations during pronunciation can also convey few dialects relevant cues, and these are explored through MFCC features. Later, the complementarity of both features is analyzed through combination. Explored features performances are evaluated through single classifier based SVM method also with ensemble ERF algorithm. Five-fold cross-fold validation approach is used to make the model more stable.

The content flow of this paper is proceeded with following sections: Section 2 briefly discuss existing work carried out with spectral and 2D Gabor features. Section 3 presents the proposed Kannada dialect dataset details and IViE dataset. ADI system along with procedure of feature extraction and classification methods details are covered in Section 4. Section 5 discusses the experimental setup and the discussion of obtained results. Section 6 provides the precise summary and conclusions of the present work along with future research directions.

II. Literature Review

Existing literature demonstrates the majority of proposed systems for dialect processing are mainly addressed through the use of acoustic-phonetic and/or phonotactic features [5], [6]. Among these, acoustic features such as spectral and prosodic features are most extensively applied for the majority of the languages. However, it can be observed that several dialects have demonstrated variations w.r.t. spectro-temporal domain features responsible for characterizing dialects. Spectro-temporal features can be effectively modeled through 2D Gabor features. Whereas, these are majorly employed in image

978-1-5386-5457-6/18/$31.00 ©2018 IEEE

TABLE I: Kannada dialect dataset

	CENK	CSTK	HYDK	MUBK	STHK
No. of speakers	28	30	33	26	24
Gender (Male+Female)	18+10	19+11	25+8	12+14	16+8
Duration (in min.)	102	103	100	125	124

TABLE II: IViE dialect dataset

	D1	D2	D3	D4	D5	D6	D7	D8	D9
No. of speakers	12 (M+F) for each dialect								
Duration (in min.)	32	31	35	37	33	31	26	38	31

processing applications. Although, few authors have made attempts of applying these features for ASR and speaker recognition tasks of speech processing [7], [8].

Kleinschmidt and Gelbart were the first to propose the use of Gabor feature extraction for ASR and shown increase in robustness of the system towards extrinsic variabilities [4]. They used two dimensional Gabor features with a combination of a simple classifier and hence proved the robustness. A study is conducted and investigated MFCC features performs best with clean speech whereas, Gabor features performs best with both clean and speech with additive noise. This activity also has shown that MFCC and perceptual linear prediction (PLP) performances degrade, but Gabor features exhibited different sensitivity towards both extrinsic and intrinsic variabilities [9].

Spectro-temporal and spectral features alone are said to carry complementary information and it is evaluation has been done [9]. Numerous filters were used to extract spectro-temporal modulation features those are organized in streams. These streams were processed with multi-layered perceptions (MLPs), and individual streams were combined with standard features [10]. In the case of acoustic event classification, optimized Gabor filter bank is applied to the datasets together with HMM classifier and resulted better performances [11].

These have suggested the significances of Gabor features for the speech and speaker recognition tasks. No systems are observed in the literature for language and/or dialect classification with the use of Gabor features. This has motivated to consider 2D Gabor features for classification dialects. Further, when Indian languages are concerned, very few standard systems are proposed for classification of dialects of regional languages. Due to unavailability of standard speech dataset for Kannada language work done is comparatively less and they used small dataset for evaluation [12]. Hence, in this work, a new dialect dataset is proposed with five dialects of Kannada language. Apart from these, neural network, Gaussian mixture models and HMMs are commonly used in the classification of dialects [13]. Ensemble algorithms are gaining more popularity nowadays. They show better performance through a combination of predictions and also with the small-sized dataset.

III. Dialect Speech Corpora

In this work, Kannada dialect dataset is used for evaluation of the proposed Gabor features Further, all experiments are carried out with standard dataset available with nine dialects of English language.

1) Kannada Dialect Speech Corpus: Proposed Kannada dialect dataset includes five prominent dialects of Kannada language. It is an official language spoken by more than 8 crores of the population across Karnataka and outside. Kannada is one among the Dravidian language spoken in southern part of India. Due to unavailability of standard speech datasets, very few works related to language and dialects identification are being observed in the literature. In this work, five prominent Kannada dialects are considered. These dialects represents the unique distinguishing pronunciation patterns followed among the group of people across the Karnataka state. The speech dataset is collected from all five regions. Five different Kannada dialects identified are Central (CENK), Coastal (CSTK), Hyderabad (HYDK), Mumbai (MUBK) and Southern (STHK). The speech data was recorded in interview style, and speakers were made to answer few questions to get their personal and work information. Also, speakers are made to involve in a friendly and informal discussion such that their speaking style is retained. Details regarding number and gender information of the dataset are given in Table. I. The recording process is carried out in a fairly clean environment using Sony voice recorder with 44100 Hz sampling frequency. Further, recorded speech is pre-processed to remove interviewer voice and noise from speech. The dataset proposed is with text-independent spontaneous data with approximate size 10 hours [14].

2) IViE Dataset: Apart from this, standard speech dataset available for English dialects is also used for system evaluation. IViE dataset is recorded from nine various dialectal regions of British Isles representing nine dialects of English language. Dialectal regions are London, Cambridge, Cardiff, Liverpool, Bradford, Leeds, Newcastle, Belfast in Northern Ireland and Dublin in the Republic of Ireland. Dataset is recorded from 12 (6 Male and 6 Female adolescents) native speakers of dialects of English, and they are made to read a story of the Cinderella from the printed material. It is a text-dependent studio recorded dataset with approximately 8 hours [15]. Details regarding number and gender information of the dataset are given in Table. II.

IV. Proposed ADI System

In this paper, dialect identification system is developed using Gabor and Spectral features individually. This section includes the brief details of stages involved in the development of ADI system. Fig. 1 presents the flow diagram of the proposed ADI system with all stages.

A. Gabor Feature Extraction

In this section, 2D Gabor feature extraction procedure is discussed. Intrinsic factors such as dialects, gender and vocal tract contribute to the variabilities of the human auditory

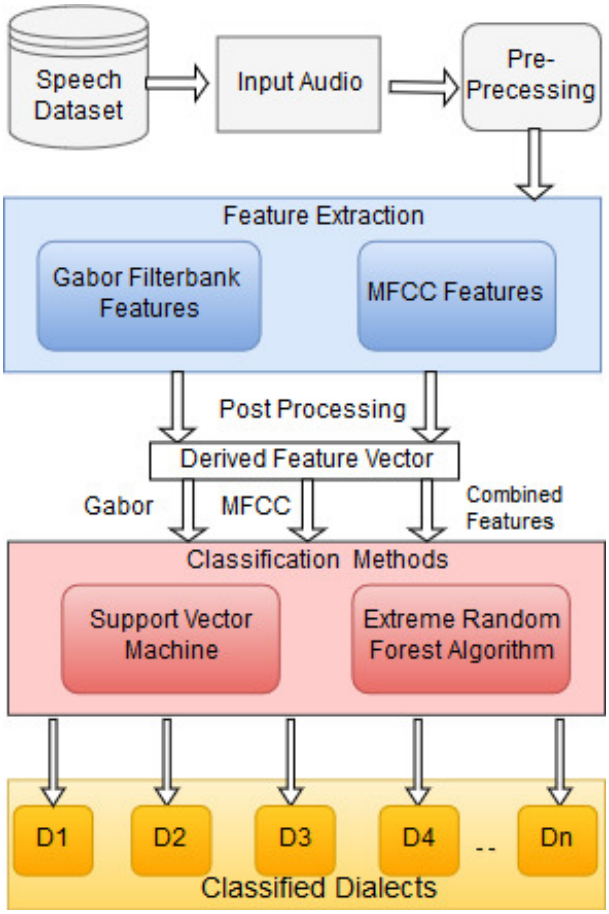

Fig. 1: Proposed model for dialect identification system

system. In the present study, dialect discriminating characteristics are retrieved from the speech. An attempt is made to reduce the gap between human and automatic language/dialect identification by transferring auditory processing principles from the human auditory system to automatic recognition systems.

2D Gabor features are computed through convolution of the log-Mel spectrogram of the speech signal (input) with a set of 2D Gabor filters. Logarithmic compression and Mel-frequency scale are considered since it represents a straightforward approach to auditory processing. A 2D Gabor filter bank is designed so that it provides approximately uniform coverage of the spectro-temporal frequency modulations. Each Gabor filter mathematically represented as a product of complex sinusoidal carrier function with corresponding envelope function. Generally, two most prominent envelope functions such as Gaussian and Hann functions are used. In present work, Hann envelop function is used in designing the Gabor filter. Eq. 1 gives the Gabor filter.

$$g(m,q,m0,q0,\omega_m,\omega_q,\nu_q) = s(m,q).h(m,q); \quad (1)$$

where, carrier function and envelope function

$$S(m,q) = s_{\omega_m}(m-m_0).s_{\omega_q}(q-q_0) \quad (2)$$

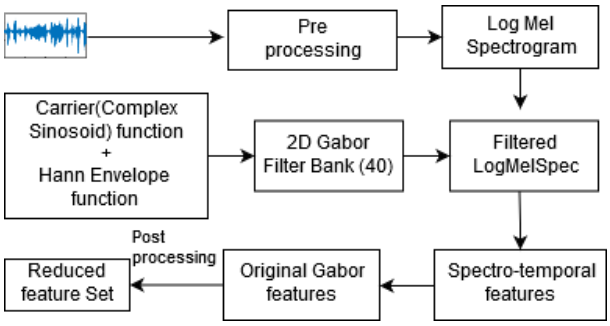

Fig. 2: Gabor feature extraction steps

$$h(m,q) = \frac{h_{\nu_m}}{2\omega_m}(m-m_0).\frac{h_{\nu_q}}{2\omega_q}(q-q_0) \quad (3)$$

the ω_m and ω_q are in terms of spectral and temporal modulation frequencies, ν_m and ν_q are the number of semi-cycles of envelope function in both spectral and temporal dimensions. The complex Sinusoidal carrier function is

$$S_{\omega_i}(x) = exp(j\omega_i x) \quad (4)$$

and Hann envelope function is given by

$$h_a(x) = \begin{Bmatrix} 0.5 - 0.5\cos(\frac{2\pi x}{a}) & ; \frac{-a}{2} < x < \frac{a}{2} \\ 0 & Otherwise \end{Bmatrix} \quad (5)$$

In this work, 23 frequency channels and 40 spectro-temporal filters are considered resulting feature 920 components. The datasets Kannada and IViE are used, as the filters with the extensive spectral extent results in small changes in feature values when one frequency channel shifts them, the redundant outputs are thus avoided by selecting the specific feature channels and reduced the dimensionality of the feature vector to 311. Hence, the centred frequency channel got chosen for each modulation filter and the channels having an overlapping of neighboring Gabor filters are included to get the reduction in feature dimension [16]. Real components of 2D filters in Gabor filter bank with spectral and temporal modulation frequencies are used. Various stages involved in the process of feature extraction are given in Fig. 2.

In addition to these exclusive spectral features are captured from vocal tract region through MFCC features. RASTA processed 13 MFCCs features are extracted from 40 filter banks from voiced activity detected components with 20 ms frame with 10 ms overlap [17]. Speaker normalization is imposed using the cepstral mean subtraction method on all the feature vectors across the dialects. Next, frame-wise feature vectors are further post-processed to obtain 26 (13*2) derived MFCCs features by computing two statistical parameters namely; mean and standard deviation from two consecutive frames [18]. Similarly, 622 derived final Gabor features are used for classification.

B. SVM and Extreme Random Forest Algorithm

SVMs are one among majorly considered machine learning algorithms used for classification. SVMs define the separable hyperplane to classify the data patterns. An optimal separable hyperplane is chosen from the many decision boundaries with which have the substantial margin differences. In this paper, SVM classification method used to capture dialect specific cues and it is designed with the one-versus-rest approach to perform 5-class and 9-class pattern classification. Kernel function plays a significant role in constructing the SVM classifier. Hence, in this work, radial basis function (RBF), for separating hyperplane with the maximal margin in a highly dimensional feature space [19].

Ensemble algorithms are gaining more popularity nowadays due to a combination of several base classification models. Always it is better to consider the predictions and opinions from several models rather from a single. In this regard, ERF a variant of the random forest algorithm fits a number of randomized decision trees over sub-samples of the dataset and uses the averaging to improve the predictive accuracy thereby controlling overfitting is employed. Decision tree-based ERF is built with totally randomized trees whose structures are independent of the output values of the learning sample. This work used the bootstrapping approach with a collection of 2048 classifier decision trees constructed from the training data. The resultant heuristic analysis concluded better accuracy using 2048 classifier trees with the IViE and Kannada Dialect Speech Corpora. While constructing the tree, node splitting is regulated on picking best split decided by Gini criterion over a random subset of features [20], [21].

V. EXPERIMENTS AND RESULTS

This section focuses on the performance evaluation of Gabor and MFCC features with the development of ADI system. Experiments are carried out using SVM and ERF ensemble algorithm with proposed Kannada dialects. Similarly, experiments are conducted to analyze and compare obtained results with a standard IViE dataset. A Series of experiments carried out with five-fold cross-validation approach by dividing the dataset into 80:20 ratio for training and testing respectively.

Performance evaluation of two different features is carried out individually and in combined form. Dialect recognition average accuracies reported with both single and ensemble ERF with clean speech are presented in III. From these results, it is observed that Gabor features have shown better performances over MFCC features. Indeed, ERF has demonstrated overall better performances with baseline and also with Gabor features. With Kannada dialect dataset, 88.25% classification performance is obtained, which is slightly better than the MFCC features. Gabor features are said to carry both spectral and temporal variations in the speech have also successfully identified the discriminating cues across Kannada dialects. Score level fusion of features has shown improved recognition since both features carry complementary information. The highest recognition accuracy of about 88.75% is obtained with ERF algorithm.

TABLE III: Dialect classification performances with Kannada and IViE dataset in (%) with clean speech

Features	MFCC		Gabor Features		Combined Features	
Models	SVM	ERF	SVM	ERF	SVM	ERF
Kannada Dataset	83.5	85.5	84.5	88.25	87.5	88.75
IViE Dataset	90.66	92.56	96.12	96.69	97.52	99.18

TABLE IV: Confusion matrix with Kannada dialect performance with Gabor and MFCC features. Average highest recognition rate: 88.75%

Accuracy	CENK	CSTK	HYDK	MUBK	STHK
CENK	88.47	11.52	0	0	0
CSTK	0	100	0	0	0
HYDK	0	0	88.66	13.33	0
MUBK	0	6.25	18.75	75	0
STHK	0	0	6.36	0	93.64

Meanwhile, a confusion matrix obtained with the highest performance with five Kannada dialects is presented in Table IV. From the confusion matrix, it is noticed that CSTK (coastal dialect) is correctly classified since this dialect speakers follow a unique pronunciation pattern those vary in the spectral temporal domain. MUBK dialect is highly misclassified with HYDK dialect. These two dialects are spoken in neighboring regions, and speaking patterns merely vary with rate and energy patterns. Similarly, the HYDK dialect is also mutually confused with MUBK. STHK dialect is slightly confused with HYDK dialect since they both follow similar patterns with energy and speaking rate parameters. CENK dialect is confused with CSTK dialect. These two dialects of Kannada are spoken in a very similar style due to use of similar phoneme styles. Also, these two dialects do match with the standard written and spoken form of Kannada.

ADI performance achieved with IViE dataset is comparatively high. Gabor features have produced an accuracy of 96.69 % and 99.18 % when they are combined with MFCC features with ERF algorithm. Significant improvement is observed with Gabor features; also English dialects are correctly classified over Kannada dialects. Reason for this is, British English dialects are with clear, distinct characteristics with reduced overlapping among them. However, Kannada dialects are not distinguishable and even identifying clear boundaries between Kannada dialects is a challenging task. Hence, dialects are misclassified with lower performances. Further, IViE dataset is recorded from the limited number of speakers. However, a large number of the speaker are involved with Kannada dialect dataset recording. Hence, speaker variabilities are also contributed for reduction of performance.

A. Analysis of noise robustness of ADI

For deploying an ADI system in a noisy environment, it is necessary to verify the effectiveness and robustness of the proposed systems under various noisy conditions. Hence, for evaluation of the system behavior under noisy conditions,

TABLE V: Dialect classification performance with Kannada dataset in (%) in the noisy environment

With Babble noise with four different SNR ratios						
Features	MFCC		Gabor Features		Combined Features	
Models	SVM	ERF	SVM	ERF	SVM	ERF
5 db	71.25	70	77	81.25	78.25	83.75
10 db	74.25	76.25	82.25	83.75	82.5	85
15 db	78.50	81	83.25	85.25	84	86.25
20 db	79.50	84.75	84.5	86.75	85.25	87.75

noisy speech signals are simulated through the inclusion of a challenging multi-speaker or babble noise. In this noise interference is a speech from multiple speakers in the locality. The robust performance of ADI system on Kannada dataset is evaluated with four different speech-to-noise ratios (SNR). Results observed are shown in Table V. Features are evaluated individually and with combined form. Gabor features have shown better performances even under noisy environment.

Also it is revealed that the proposed features performed slightly better than baseline MFCC features with all four SNRs considered (5dB, 10dB, 15dB and 20dB). 3 shows the robustness performance of the system with Kannada dialect dataset. MFCC features are resulted with 70% & 84.75% and Gabor features with 81.25% & 86.75% accuracy with 5dB and 20dB SNR ratios respectively. The slight increment in accuracy is seen with Gabor features with increase of SNR values, however higher increase is observed with MFCC features with the increase of SNR ratios. Whereas, proposed features and combined feature vectors have shown very similar performances with all SNR ratios.

VI. CONCLUSIONS

This work demonstrated the applicability of 2D Gabor features for dialect identification problem. These are well established features in the field of image processing. An attempt is made to apply 2D Gabor features for classification of five prominent dialects of Kannada language. Spectro temporal variations during pronunciations across five dialects of Kannada language. Also, vocal tract spectral features are extracted using traditional MFCC features. Further, non overlapping dialectal cues existing from both are explored through score level fusion. Explored features are evaluated on a newly proposed Kannada dialect dataset and with a standard IViE dialect corpus. Gabor features have shown sustained improvement in performance over baseline MFCC features with both SVM and ERF classification methods. The highest classification performance of about 88.75% and 99.18% is achieved with Kannada and IViE datasets respectively. The performance of Gabor features is also evaluated for noise robustness. Indeed, they resulted with improved performance over MFCC. Further, the size of the proposed Kannada dialect dataset needs to be increased so that that state-of-the-art deep neural network approaches can be applied. As of now, they are providing very less accuracy.

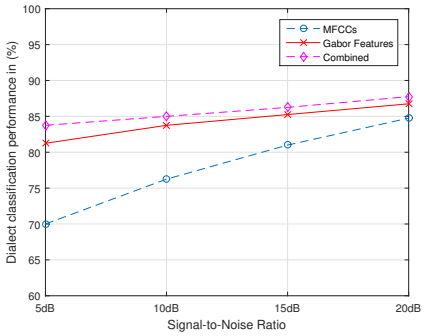

Fig. 3: Noise robustness with Kannada dataset (Recognition performance in %)

ACKNOWLEDGMENT

This work is supported by DST-GOI (Department of Science and Technology, Government of India) sponsored project entitled *Characterization and Identification of Dialects in the Kannada Language*.

REFERENCES

[1] J. K. Chambers and Peter Trudgill, *Dialectology*, 2nd ed. Cambridge University Press, 1998.

[2] M. J. Harris, S. T. Gries, and V. G. Miglio, "Prosody and its application to forensic linguistics," *LESLI: Linguistic Evidence in Security Law and Intelligence*, vol. 2, no. 2, pp. 11–29, 2014.

[3] J. H. L. Hansen and G. Liu, "Unsupervised accent classification for deep data fusion of accent and language information," *Speech Communication*, vol. 78, pp. 19–33, 2016.

[4] M. Kleinschmidt and D. Gelbart, "Improving word accuracy with Gabor feature extraction," in *Seventh International Conference on Spoken Language Processing*, 2002.

[5] R. Huang, J. H. Hansen, and P. Angkititrakul, "Dialect/accent classification using unrestricted audio," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 453–464, 2007.

[6] M. A. Zissman, T. P. Gleason, D. Rekart, and B. L. Losiewicz, "Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech," in *Acoustics, Speech, and Signal Processing, ICASSP*, vol. 2, 1996, pp. 777–780.

[7] Meyer, Bernd T and Kollmeier, Birger, "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition," *Speech Communication*, vol. 53, no. 5, pp. 753–767, 2011.

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

[9] B. T. Meyer and B. Kollmeier, "Complementarity of MFCC, PLP and Gabor features in the presence of speech-intrinsic variabilities," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[10] S. V. Ravuri and N. Morgan, "Using spectro-temporal features to improve AFE feature extraction for ASR," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[11] J. Schröder, S. Goetze, and J. Anemüller, "Spectro-temporal Gabor filterbank features for acoustic event detection," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 12, pp. 2198–2208, 2015.

[12] R. Soorajkumar, G. Girish, P. B. Ramteke, S. S. Joshi, and S. G. Koolagudi, "Text-Independent Automatic Accent Identification System for Kannada Language," in *Proceedings of the International Conference on Data Engineering and Communication Technology*. Springer, 2017, pp. 411–418.

[13] F. Biadsy, J. Hirschberg, and D. P. Ellis, "Dialect and Accent Recognition Using Phonetic-Segmentation Supervectors," in *Interspeech*, 2011, pp. 745–748.

[14] N. B. Chittaragi, A. Limaye, N. Chandana, B. Annappa, and S. G. Koolagudi, "Automatic Text-Independent Kannada Dialect Identification System," in *5th International Conference on Information System Design and Intelligent Applications, India (2018), Universit des Mascareignes, Mauritius (in-press)*. Springer, 2018.
[15] E. Grabe and B. Post, "Intonational variation in the British Isles," in *Speech Prosody*, 2002.
[16] B. T. Meyer, S. V. Ravuri, M. R. Schädler, and N. Morgan, "Comparing different flavors of spectro-temporal features for ASR," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.
[17] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE transactions on speech and audio processing*, vol. 2, no. 4, pp. 578–589, 1994.
[18] N. B. Chittaragi, A. Prakash, and S. G. Koolagudi, "Dialect identification using spectral and prosodic features on single and ensemble classifiers," *Arabian Journal for Science and Engineering*, vol. 43, no. 8, pp. 4289–4302, 2018.
[19] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, p. 27, 2011.
[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[21] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.