Int J Speech Technol (2016) 19:593–609
DOI 10.1007/s10772-016-9351-7

Analysis and modeling of acoustic information for automatic dialect classification

S. S. Agrawal[1] · Aruna Jain[2] · Shweta Sinha[2]

Received: 16 January 2016 / Accepted: 6 July 2016 / Published online: 22 July 2016
© Springer Science+Business Media New York 2016

Abstract A primary challenge in the field of automatic speech recognition is to understand and create acoustic models to represent individual differences in their spoken language. Individual's age, gender; their speaking styles influenced by their dialect may be few of the reasons for these differences. This work investigates the dialectal differences by measuring the analysis of variance of acoustic features such as, formant frequencies, pitch, pitch slope, duration and intensity for vowel sounds. This paper attempts to discuss methods to capture dialect specific knowledge through vocal tract and prosody information extracted from speech that can be utilized for automatic identification of dialects. Kernel based support vector machine is utilized for measuring the dialect discriminating ability of acoustic features. For the spectral feature shifted delta cepstral coefficients along with Mel frequency cepstral coefficients gives a recognition performance of 66.97 %. Combination of prosodic features performs better with a classification score of 74 %. The model is further evaluated for the combination of spectral and prosodic feature set and achieves a classification accuracy of 88.77 %. The proposed model is compared with the human perception of dialects. The overall work is based on four dialects of Hindi; one of the world's major languages.

Keywords Support vector machine · Gaussian mixture model · ANOVA test · Dialect identification · Human perception

1 Introduction

The inherent advantage of speech communication due to its variability, convenience and speed along with our increasing requirements to communicate with machines has driven the attention of researchers towards mechanical recognition of speech. Speech recognition accuracy is the most desired feature of speech-enabled applications. Technological advancements and improvements in the fundamental approaches have shown a successful transition from small vocabulary isolated word recognition to large vocabulary continuous speech recognition. With this the domain and the application of speech recognition systems have expanded. The demand for interactive voice response based interfaces has enjoyed remarkable escalation as part of internet revolution in recent years. These demands extend far beyond primitive forms of man–machine communication.

Performance of automatic speech recognition (ASR) systems degrades due to variability inherent in the signal. Apart from linguistic variability, speaker and channel variability are the two major inherent variability in speech. Inter and intra speaker variability is considered as one of the major issue in ASR performance. While substantial work has been done in this direction; speaker variability is still a major concern. Possible solutions to improve the performance of ASR systems require modeling techniques

✉ Shweta Sinha
meshweta_7@rediffmail.com

S. S. Agrawal
ss_agrawal@hotmail.com

Aruna Jain
arunajain@bitmesra.ac.in

[1] KIIT Group of Colleges, KIIT Campus, Sohna Road, Gurgaon, Haryana, India

[2] Department of Computer Science and Engineering, Birla Institute of Technology, Mesra, Ranchi, India

that can capture all inter and intra speaker variations. Dialect of a speaker is one of the major factors that influence the speech characteristics. Dialect is a form of language followed by people in a particular geographical area. They have a peculiar pattern of pronunciation and grammar rules. Even when speaking the standard form of a language, influence of speaker's native dialect can be seen on the acoustic characteristics of the spoken utterance. Presence of these variability veil the intended message in the signal with uncertainty. System performance degrades if these issues are not handled in advance. Automatically identifying the spoken dialect prior to the ASR engine will help improve the system performance.

Dialectal study can be based on phonotactic knowledge of the language that deals with language phonemes and their sequences or can be based on acoustic aspect. The acoustic approach for dialect identification is concerned about the identification and modeling of speech features that can discriminate the speech sounds belonging to speakers from different dialects. This approach makes use of acoustic phonetic features of speech signal that include spectral and prosodic features of speech. The present work deals with the pronunciation issues related to the dialects and hence follow acoustic approach for this study. Most of the research for ASR has been done for languages like English, Chinese, Japanese, Arabic and Spanish. These are merely a fraction of existing thousands of languages. English (British and American) followed by Chinese are two well-matured languages in this research area. Contrary to these languages, ASR research in Indian languages (Hindi, Tamil, Telugu, Punjabi and Bengali) are still in its inception stage. The deficit of available resource for these languages can be ascribed as one of the major reason for their slow progress. The paucity of annotated speech corpora for training and testing of acoustic models persists for these languages. Most of the research in the field of Indian language concentrates on feature extraction and modeling aspects (Raman 1985; Rao 1993; Sekhar and Yegnanarayana 2002; Kumar et al. 2004). While efforts have been made for performance enhancement of ASR in last several years, still there is a vast gap between the performance of man and machine in this area. Variability in the speech signal, which is one of the major constraints for system performance, has not yet attracted Indian researchers working in speech technology field. Aggarwal and Dave (2012) have paid their attention to cater the channel variability due to background noise, but has not tackled speaker variability issues. Recently efforts have been made for study of Hindi dialects. Comparative phonological study of Hindi dialects has been presented by Mishra and Bali (2011). Study of acoustic characteristics of Hindi dialects on a small database of 10 speakers highlight the prosodic differences among dialects of Hindi. These differences are due to vowel duration and intensity (Kulshreshtha and Mathur 2012).

This paper addresses an important aspect of ASR, which is the question of how to incorporate resilience to dialect speech into ASR systems. The present contribution attempts to cater speech variability due to dialect. Identifying dialect prior to the speech recognition will allow the use of restricted pronunciation dictionary, thus reducing the search space in any ASR. The importance of the work is enhanced by its focus on a major world language, Hindi. About 45 % of the Indian population speak Hindi which itself has several variety (dialects) spoken in different parts of Hindi speaking region. Today, computers and limited resource devices have become absolutely indispensable for the people of urban India. But for the development of a country where maximum population (68 %) belongs to the rural area, the technology has to reach them as well. The input device like keyboard and mouse attached to the computers require certain level of expertise as well as proficiency in English language to handle them. For the digital technology to reach to the unreachable, these constraints can be overcome by using speech based communication technology as a medium of interaction between man and machine. Multi-modal ASR system is required to handle the dialectal diversity in Hindi. One approach toward handling this diversity is to identify dialect prior to ASR engine there by reducing the search space of the matching algorithm. This paper presents automatic dialect classification for Hindi. This is the first work based on regional dialects of Hindi. To identify the acoustic correlates for these dialects, ANOVA test is run on the Hindi vowel sound units. Segmental and supra-segmental features are used for identification of dialects. This paper is arranged as follows: Sect. 2 discusses corpus development for Hindi dialects, statistical analysis of acoustic variations of Hindi vowel sounds in different dialects is summarized in Sect. 3; Sect. 4 identifies speech features for identification of dialects, Sect. 5 outlines modeling and evaluation of spectral and prosodic features, Sect. 6 discusses human perception of dialects and the paper is concluded in Sect. 7.

2 Development of speech corpus for Hindi dialect identification

Indian languages belong to different language families. 75 % of the Indian population speaks languages belonging to the Indo-Aryan family (Hindi, Urdu, Bengali etc.), 20 % population speak Dravidian language (Telugu, Tamil, Kannada) and, the rest speak languages of the Austro-Asiatic and Sino-Tibetan family. Indian constitution has identified 22 languages as scheduled language. Although there is no such language declared as national language by

the constitution of India, Hindi is considered as the mother tongue of India. It is spoken by almost 45 % population of India and covers a huge portion of world population. In recent times, this language has attracted the attention of several research communities with different literary and linguistics based scientific approaches. Efforts have been made towards the development of automatic speech recognition (ASR) systems for Hindi but literature survey highlights that not much has been achieved in the area of Indian languages, especially Hindi. Initiating any research specific to a language requires corpora for the language that spans the domain of research. The first and foremost requirement for starting Hindi speech project is to build corpora that cover the research domain.

As with the other major languages of the world, Hindi too has several varieties termed as dialect, followed across the Hindi belt of India. This language is mainly spoken by people in North and Central India. The Indian states like Delhi, Haryana, Bihar, Chhattisgarh, Himachal Pradesh, Jharkhand, Madhya Pradesh, Rajasthan, Uttarakhand and Uttar Pradesh constitute the Hindi belt of India. The dialects of Hindi are broadly categorized as the Eastern dialect and the Western dialect. Indian states of Bihar, Jharkhand, Chhattisgarh, East Uttar Pradesh, and North Madhya Pradesh follow Eastern Hindi while the Western Hindi is spoken in the rest of the Hindi belt. All across the Hindi speaking region, the written communication is done following some standard form of Hindi that is governed by grammar rules of the language but the spoken communication is done in the regional form of the language. There is no formal education on the dialects of Hindi and human beings at a very early age develop these dialectal styles of speaking due to their geographical and social surroundings. The major dialects of Eastern Hindi are Awadhi, Bhojpuri, Bagheli and Chhattisgarhi and those of the Western Hindi dialects are Braj bhasha, Haryanvi, Bundeli, Kannauji and Khari boli. Huge dialectal diversity exists among these varieties.

Speech related research are based on data-driven technology and requires a large amount of labeled data. These data are used for training acoustic models. Contrary to major European and American languages with huge speech corpus in the public domain Hindi has no standard text and speech corpora for researchers. Individuals or research groups working in this field have created databases for fulfilling their requirements. Hindi is a rich language with approximately double the characters as compared to English. Varieties of speaking patterns are followed by speakers all across Hindi speaking belt. Influence of speaker's native tongue is widely observed in the spoken utterances. A multi-modal ASR system that can cope with these dialectal diversities is required to provide robustness to the system. This generates the need for speech corpus that covers all aspects of dialectal variations in the language. For the study of regional Hindi dialects, no pre-recorded speech corpus is available in public domain. Lack of such resources for Hindi is the major hurdle in speech processing research for this language.

This research is focused on four Hindi dialects; two from eastern Hindi and two from western Hindi have been selected for this study. The Khari boli (KB) which is one of the major western dialects is also known as standard Hindi and has the maximum number of speakers. It is spoken in the rural surroundings of Delhi, north-western Uttar Pradesh, as well as in neighbouring areas of Haryana and Uttarakhand. The other western dialect considered for this research is Haryanvi (HR) which is spoken by the people of Haryana, Punjab, parts of Rajasthan and Delhi. Among the eastern dialects Bhojpuri (BP) and Bagheli (BG) dialects have been selected. Bhojpuri is one of the widely spoken Eastern dialects of Hindi. It is spoken by approximately 33 million people and is a prominent dialect in eastern Uttar Pradesh, Bihar and Jharkhand. Bagheli is the variant of Hindi language spoken in Baghelkhand (central India). It is used for intra-group and inter-group communication among people of Rewa, Satna, Jabalpur districts of Madhya Pradesh and Koriya district of Chhattisgarh. There are about 2.8 million people of India who speak this dialect.

The purpose of this research is to study the influence of native dialects on acoustic characteristics and in turn exploit them to automatically identify the spoken dialect. It necessitates the development of speech and text corpus to cover all aspects of this research. In lieu to achieve the goal, text prompt sheets were prepared in standard Hindi using Devanagari script. The corpus consists of 600 words and 300 meaningful sentences capturing all the phonemes of the language. The text data used for recording consists of words and sentences from travel domain. Length of sentences in the text corpus varies from 7 to 13 words and the number of syllable per word varies from 1 to 5. The corpus is recorded using 50 native speakers (30 male and 20 female) from each of the four dialects. All the speakers were of the age group 18 to 51 years and had at least 15 years of formal education. Each speaker has to read 600 isolated words and 300 continuous sentences in one session. The total number of utterances in the database is (300 sentences + 600 words) × 50 speakers × 4 dialects) 180000. Each dialect has 45000 (27000 male utterance + 18000 female utterance) utterances. Approximately 1.10 h of read speech samples are obtained from each speaker. All recording was done in the office environment using a single microphone. Considering the factors like, intelligibility, perceptual quality and information storage in the speech signal the recorded signals are sampled at 16 kHz using the software GoldWave and are represented as 16 bit number.

2.1 Measures to reduce recording variability

Along with inter-dialectal diversity intra-dialectal diversity can be seen for most of the Hindi dialects. To reduce the intra dialectal variability geographic propinquity was maintained during the selection of informants. Standard Hindi in general is the common language for inter-group communication. It was not feasible to record the samples by physically going to the dialectal regions; instead the informants of those regions residing near the place of recording were selected for this task. Since subjects were not necessarily residing in their native places, it was assumed that they may not be always using their regional dialect for communication. To get the influence of their native dialect on the spoken utterance, the speakers were asked to speak for few minutes in their native dialect before the actual recording. Only speakers, who have studied their native language at least up to school level and are well verse, frequent in use of their native dialects with no articulatory defects were selected. To reduce the session and channel variability recording of each speaker was done in one single session with single microphone.

2.2 Phonological differences among the dialects

Dialect influences individuals speaking style. Insight into the phonological differences among the dialects can outline the factors that affect the acoustic properties. The native language shapes individuals sense of language. Khari boli is referred as standard Hindi dialect and has ten native vowels. Gemination is a feature of Khari boli dialect that gives its distinctive sounds. There is much similarity between Bagheli and Khari boli dialect. Some of the phonological differences between these dialects are as under:

- In Bagheli dialect /o/ and /u/ sounds occur in complementarity. (e.g., /d̪ono/ as /d̪uno/).
- Bagheli and Bhojpuri speakers are not able to differentiate /r/ and /rh/ sounds. (e.g., /kʰɑ:rhi:/ as /kʰɑ:ri:/).
- /ʋ/ is often replaced by /b/ in almost every dialect. (e.g., /d̪ɑ:ʋət̪/ as /d̪ɑ:bət̪/).
- In Bhojpuri dialect /u:/ occurs freely in place of /o/. (e.g., /d̪o/ as /d̪u:/).
- The sound of /ə/ is often pronounced clearly by Bhojpuri speakers even at the end of words that is not customary in standard Hindi (e.g., /kəməl/ as /kəmələ/). Speakers of Bhojpuri dialect have slow rate of speech production. The vowel /ə/ is more rounded in Bhojpuri dialect.
- In Haryanvi dialect, most of the vowels are more open as compared to other dialects. The sounds of /ə/ often occurs in free variation with /ɑ:/. Retroflection is the marked feature of Haryanvi dialect. (e.g., /kIɖ̪ʰər/ as /kIɖ̪ʰɑ:r/).
- The sounds like /t/ and /d/ at the end are often stressed and pronounced as clustered sound. (e.g., /təmIl nɑ:d̪u:/ as /təmIl nɑ:d̪d̪u:/).
- Except Khari boli, in almost every dialect the speakers do not distinguish sounds of the fricatives /s/ and palatal /ʃ/. Most of the speakers prefer the former. (e.g., /kələʃ/ as /kələs/).
- No differentiation is done by Bhojpuri speakers for /s/ and /ʂ/ sound. Most of the speakers prefer the former. (e.g., /bʰɑ:ʂɑ:/ as /bʰɑ:sɑ:/).
- Differentiation among the sound of /pʰ/ and /f/ and /dʒ/ and /z/ does not exist in any dialect except Khari boli. (e.g., /gədʒəl/ as /gəzəl/ and (/pʰəl/ as /fal/).

Influence of these speaking variations may be observed during communication while speaking standard Hindi. Study of acoustic characteristics requires acquaintance with language phonology. Co-articulation is a very important phenomenon that has motivated the researchers for segment or context dependent modeling approaches. But, pronunciation effect appears stronger when spoken in relaxed style and often leads to reduced articulation. In many cases these may be language dependent or, may be related to regional origin controlling the accent.

3 Statistical analysis of acoustic variations of Hindi vowel sounds in different dialects

Triggered by the belief, that incorporating knowledge about dialects in pronunciation dictionary and acoustic training can increase efficiency of speech-based systems have attracted the attention of speech researchers towards the study of acoustic characteristics influenced by regional variability. Since the differences in dialects account for differing phonetic realization of vowels and consonants of a language, analysis of their acoustic characteristics can further accentuate the differences between them. A variety of acoustic–phonetic cues for dialect discrimination have been explored in previous research on dialect or accent recognition. Such cues mainly include vocal characteristics represented by spectral features and paralinguistic information represented as prosody features. Even though use of prosodic features is not very common in ASR research for recognizing spoken utterances, linguistics believes that its use will help perform the system better. It is difficult to outline exactly which of these characteristics influence the speaking style. To analyze the effect of speakers native dialect ANOVA test have been performed on formant frequencies, pitch slope, intensity, and duration.

3.1 Speech data preparation

This research is the first of its kind for Hindi language. Active research has been carried out for dialect/accent based acoustic characteristics during past few years in other languages. Rabiner and Juang (Rabiner and Juang 1993) points out that although, for the written text identification vowels have very low relevance, their high recognition performance is essential for the reliability of any automatic speech recognition system. The reason behind this is the mechanism of vowel production. Vowels are produced by exciting an essentially static vocal tract shape with a quasi periodic excitation signal. They usually are of long duration and are spectrally well defined. Arslan and Hansen (Arslan and Hansen 1996) highlights in their study that vowels carry more information than consonants in accent assessment. Most of the studies in dialect or language identification concentrate on vowel sounds of the language. Unfortunately, vowels are more often distorted than consonants in accented speech. Wells (Wells 1982) in his book on accents outlines that accent variations stretch out not only in phonetic characteristics but also influences the prosodic characteristics. This section seeks to determine the effect of dialect on Hindi vowels acoustic characteristics, there by, identifying features that differentiate them.

Hindi language has 11 vowels, that are categorized as short vowels (/ə/, /I/, /ʊ/) and long vowels (/ɑ:/, /i:/, /u:/, /e:/, /ɛ:/, /o/, /ɔ:/). Out of these 11 vowels, nine are monophthongs and 2 vowels (/ɛ:/, /ɔ:/) can be pronounced as pure vowels as well as diphthongs. Out of all these, ten vowels are transcribed in two distinct forms; the independent form and the dependent form. In general, vowels in its independent form appear alone, at the beginning of any word or, immediately following another vowel. Dependent form of vowels is used when they follow any consonant. The first vowel schwa (/ə/) does not appear as dependent form and is assumed to be implicit in each Hindi consonant unless some other vowel in its dependent form is present. But, this schwa is deleted obligatorily at the end of words.

3.2 Vowel extraction

A syllable is one such acoustic unit that has a close connection with human speech perception and articulation. Also, acoustic features associated with prosody are suprasegmental features, and can only be properly extracted from syllables. The speech must be segmented and transcribed to extract waveform corresponding to each vowel from the syllables. To guarantee accuracy, speech segmentation in this work is done manually using Praat software (version 5.1.04). The waveform along with spectrogram is used to segment recorded utterances into words, words into syllables and further, from the syllables of interest vowels were extracted. Articulation has the impact of structural variables like phonetic context and position in the phrase (Cho and Keating 2001). To capture the co-articulation effect in vowels at different prosodic positions, they were extracted from word-initial, word-medial and word-final position. Figure 1 shows an example of segmentation. The top panel represents the waveform of a recorded utterance; the middle panel represents the utterance spectrogram, and the bottom panel has three layers corresponding to the word, syllable and phoneme segmentation.

3.3 Acoustic feature extraction

Out of several consonantal contexts in Hindi, CV syllables are the most frequent syllables followed by CVC structure. In this study, the vowels were extracted from these two syllabic structures. Acoustic parameters, such as the first, second and third formant frequencies F1, F2 and F3 respectively, the fundamental frequency F0, vowel duration and intensity of speech are extracted from these vowels using speech analysis software Praat.

Autocorrelation based pitch extraction procedure was used to extract the first three formants. Corresponding to each of the three positions, for every speaker, syllables of interest were considered for extraction of 10 vowels used in the analysis. Their acoustic parameters were obtained, and finally individuals mean for each of the ten vowels were computed. From these individual means, dialectal means are obtained separately for male and female speakers.

3.4 Acoustic feature analysis

A hypothesis of this research is that acoustic analysis of vowels can help in defining synchronic differences between regional dialects of Hindi relative to some baseline variety. To justify our assumption One-way ANOVA (analysis of variance) was calculated for ten Hindi vowels using four important dialects (Khari boli, Bhojpuri, Haryanvi and Bagheli) as between-subjects factor for formant analysis (F1, F2 and F3), pitch (F0) vowel duration and intensity analysis at three positions in the word.

3.5 Analysis of formant frequency

For the spectral features cepstral based approach are the ones that have always been researcher's choice. Formants have also been used by several researchers in identification of dialects. Yan and Vaseghi (2003) have modeled formant space of three English accents. Their experiment concluded with the finding that second formant is the most influential formant frequency for providing dialect specific

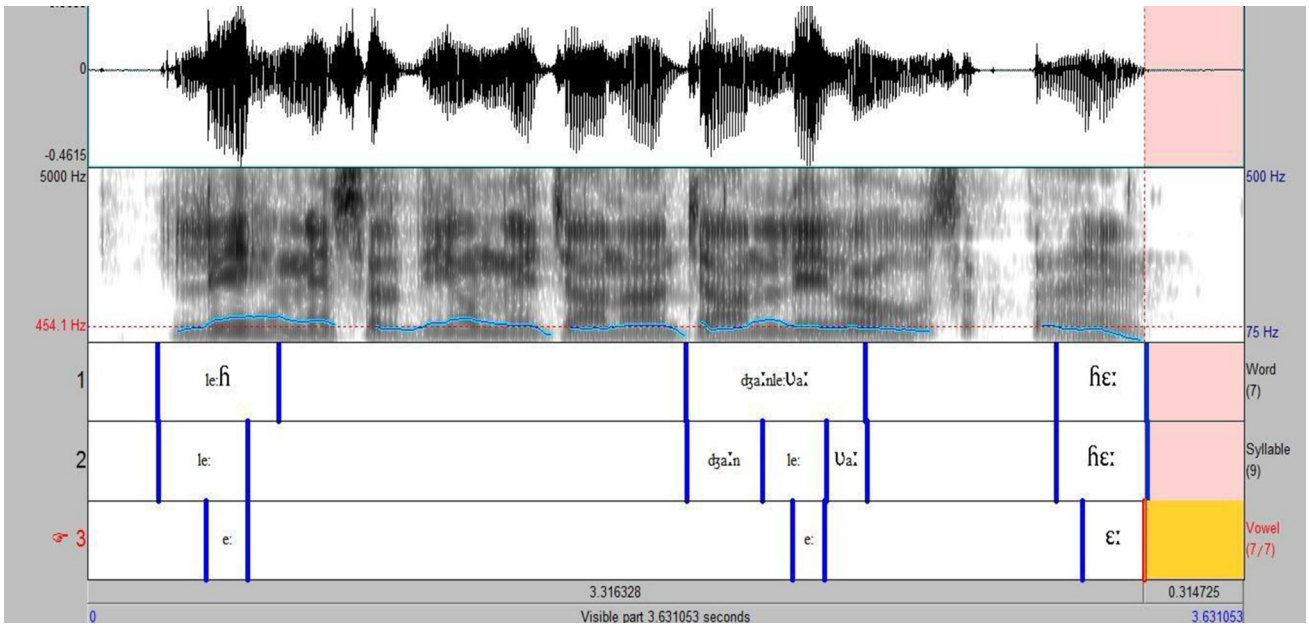

Fig. 1 Manual segmentation of Khari boli utterance using Praat software

information. Study was performed by Adank et al. on regional acoustic varieties of Dutch vowel systems. Results highlight that vowel duration vary in different regions. They further submitted that the regional impact on the second formant frequency is much prominent than that on the first formant frequencies of vowels (Adank et al. 2007). Arslan and Hansen posited the importance of second and the third formant for accent identification with the emphasis on their relation with tongue movements (Arslan and Hansen 1996).

Considering Khari boli as the standard Hindi dialect (Pandey 1989) for analysis, the F1-F2 plot for four Hindi dialects show that, second formant values for Bhojpuri dialect speakers are higher for back vowels (/ɑ:/, /ʊ/, /u:/, /o/, /ɔ:/) for Bagheli speakers F2 is higher for all but/ɑ:/. Haryanvi speakers have an approximately same value of second formant except for /ɑ:/ where it is higher as compared to Khari boli speakers. It can be further observed that for front vowels (/i:/, /I/, /e:/, /ɛ:/), F2 for Bhojpuri speakers are low compared to Khari boli speakers. F1 for Haryanvi speakers are high for close front vowel (/I/, /i:/). F2 value for all front vowels except for open front vowel (/ɛ:/) is high for Bagheli speakers as compared to speakers of Khari boli dialect. Figure 2 represents vowel triangle for four Hindi Dialects.

Similar trends were seen for formants values of vowels obtained from female speakers of these dialects, except for they possess higher F1, F2 values as compared to their male counter part. To draw some statistical inferences from the experimental data one-way ANOVA was performed on the set of three formant values for ten vowels with Pearson correlation coefficient(p) set to .05.

The results of one-way ANOVA for the formant frequencies of male speakers revealed a significant main effect of dialects for third formant frequencies of almost all the vowels except for the vowels, /I/ ($p = .685$), /i:/ ($p = .137$) and /ʊ/ ($p = .059$). Significant main effect of the dialects for the first formant frequencies were only observed for vowel /I/ [$F(3,116) = 11.168$, $p = .024$], /e:/ [$F(3,116) = 18.712$, $p < .001$] and /o/ [$F(3,116) = 16.024$, $p = .004$]. For male speakers, the second formant frequencies show statistically significant difference between the dialects for the vowels /ʊ/ [$F(3,116) = 19.671$, $p < .001$], /u:/ [$F(3,116) = 18.096$, $p < .001$], /ɛ:/ [$F(3,116) = 10.702$, $p = .044$], /o/ [$F(3,116) = 19.031$, $p < .001$] and /ɔ:/ [$F(3,116) = 12.061$, $p = .021$]. To measure the pair-wise differences among the dialects Tukey post hoc test was run for these formant frequencies. The results of this test further revealed that F1 for vowel /I/ and /o/ were significantly low ($p < .001$) for KB and BP dialects only. It was further observed that F2 for vowels /ʊ/, /u:/ and /o/ showed significant results for the dialects KB and BP only (all $p < .05$). Maximum variations in vowel /ə/ (short, neutral vowel) for the four dialects are observed, followed by vowel /ɛ:/ (mid-front vowel). No significant differences were further observed between any other two dialects. Hence decision based on these formants could not alone be used for distinguishing these dialects. The test further revealed that for F3 obtained from the male speakers, no significant difference between any two

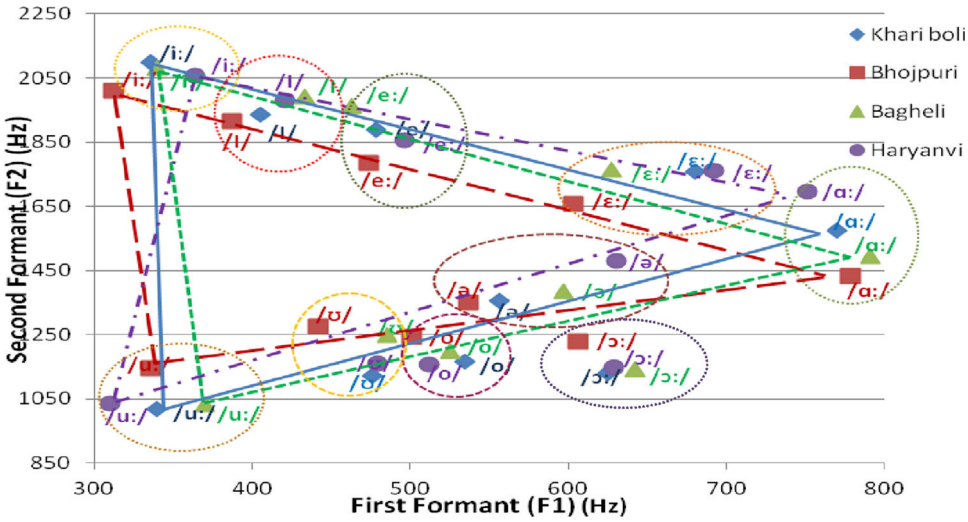

Fig. 2 Vowel space diagram approximated by male speakers of four Hindi dialects

dialects for /o/ exists. Significant main effect on F3 value of vowel /ɑ:/ ($p < .001$) is observed for dialects KB and BG.

From the results obtained by the execution of one-way ANOVA on female speakers' formant frequencies a statistically significant main effect of dialects on the third formants of almost all the vowels, except for /I/ ($p = .721$) and /i:/ ($p = .426$) is seen. Tukey post hoc test revealed that F3 for /o/ is significantly lower KB and BG dialects ($p < .001$). For all other vowels, Tukey post hoc test revealed a significant difference between F3 (all $p < .05$) of speakers from different dialects. Significant differences between the second formant frequency of back vowels are observed for the dialects (all $p < .05$). ANOVA for F1 showed a very little significance of dialects on the vowels. Furthermore, the post hoc test revealed that these differences in F1 are selective in nature.

3.5.1 Pitch and pitch slope

F0 value for the start and end position of the vowels along with the average for whole vowel duration was extracted to study the influence of pitch and its variations. To evaluate the regional effect on F0, One-way ANOVA was carried out for the male and female speakers of each dialect separately. The results of one-way ANOVA for both male and female speakers highlight the significant main effect of dialects on the fundamental frequency at different positions of most of the vowels. For both male and female speakers, no significant effect of dialect was observed for /u:/ at word initial position and for /I/ and /u:/ at word middle position (all $p > .05$). Also, no significance of dialect for female speakers were observed for /o/ ($p = .058$) at word middle position. Tukey post hoc test further revealed that the pair wise differences were significant for long vowels (all $p < .001$) among all the dialects. This may be influenced by the long duration of these vowels that influences the rhythm. Further analysis of post hoc test revealed that no significant effect of dialect on F0 for other vowels exists, and results obtained were selective in nature. It can be analyzed that no concrete conclusion regarding the spoken dialect can be obtained based only on the average pitch of speakers over the vowel duration and variations in pitch over time may be useful for this study.

The previous result shows that pitch plays a significant role in accent identification. Slope of pitch contour was verified by Grover et al. (1987). They posited that German, French and English speakers significantly differ in their intonation slope. Pitch slope has been computed as the variations of the fundamental frequency divided by the duration of the vowel in seconds. The variation is defined as the difference of pitch value at the end vowel position and pitch at the start of the vowel. This same method of pitch slope computation is followed by Zheng et al. (2012) in their study on two British dialects. The slope thus computed reflects steepness and variations over the whole vowel. The results of pitch slope for 10 Hindi vowels obtained at three word positions (Initial, mid and final) in four Hindi dialects are represented in Figs. 3 and 4.

These values were obtained from the mean of 30 male data. Similar fall and rising trends were obtained for the female speakers of these dialects. From the figures, it can be seen that pitch slope is negative for /ə/ in Haryanvi and Khari boli dialect; but is positive for Bhojpuri and Bagheli dialects at word initial position. Also, the negative slopes are much steeper than the positive slope. The slope of /I/ is negative in all dialects steeper fall in pitch is observed in Bhojpuri dialect. /ʊ/ in every dialect has a negative slope but the fall for Bhojpuri and Khari boli dialect is very sharp. For /ɔ:/, Khari boli has a negative slope at the word initial position whereas, all other dialects have a negative slope.

In /ɛ:/, the steep rise in slope for Haryanvi is observed at the word initial position. Also, this vowel has a positive

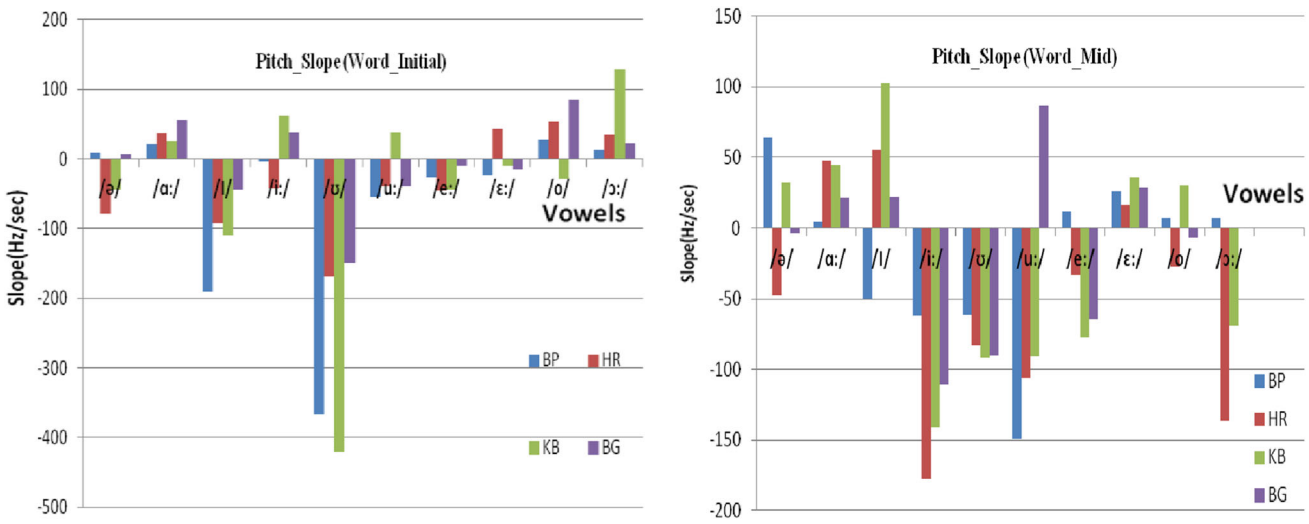

Fig. 3 The mean pitch slope of 10 vowels over 30 male speakers at word initial and mid positions

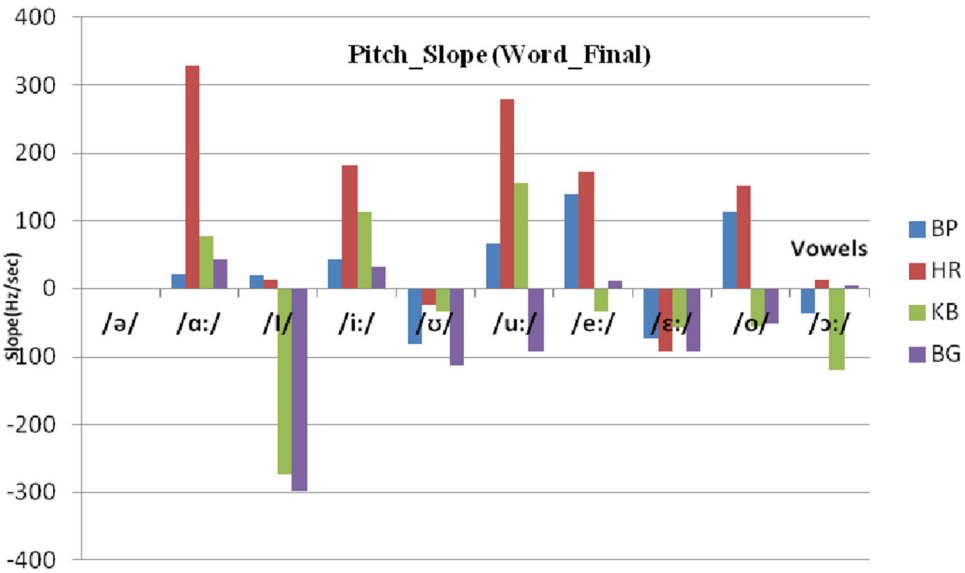

Fig. 4 The mean pitch slope of 10 vowels over 30 male speakers at word final positions

slope in all dialects at the word mid position but is negative at the word end position. At the word mid position more steep rise and fall can be observed for all the vowels as compared to an initial position. Most of the vowels show similar characteristics in all the dialects. Few like, /u:/ is negative for all except Bagheli. Further, Haryanvi and Khari boli has a sudden fall in pitch whereas, a slow rise of the pitch is observed for Bagheli dialect speakers. In the word end position, for the vowel /i:/ all dialects have a positive slope, but Haryanvi has a steep rise as compared to others. Much steeper change is observed for /o/ in all the dialects at the word end. /ɔ:/ shows slow fall or rise at the end position. In Hindi, vowel /ə/ is not pronounced at the end (Schwa deletion) and hence was not studied for this position.

3.5.2 *Vowel duration*

Factors such as the location of pauses, rhythm, the number of syllables, manner of articulation and speaking style all influence the duration of a vowel as well as language phonemes (Sinha et al. 2013). Since articulation manners in each dialect are unique, differences in phonetic duration are realized from one style to another. The average duration of ten vowels at the three positions in a word by male and female speakers is represented in a stacked graph in Fig. 5a, b respectively. From the Fig. 5a it is clear that the male speakers of Haryanvi and Bagheli do not show much difference in the spoken duration of vowel /ə/ at the middle word position, but have a significant average duration difference of 42 ms at the initial word position. For the

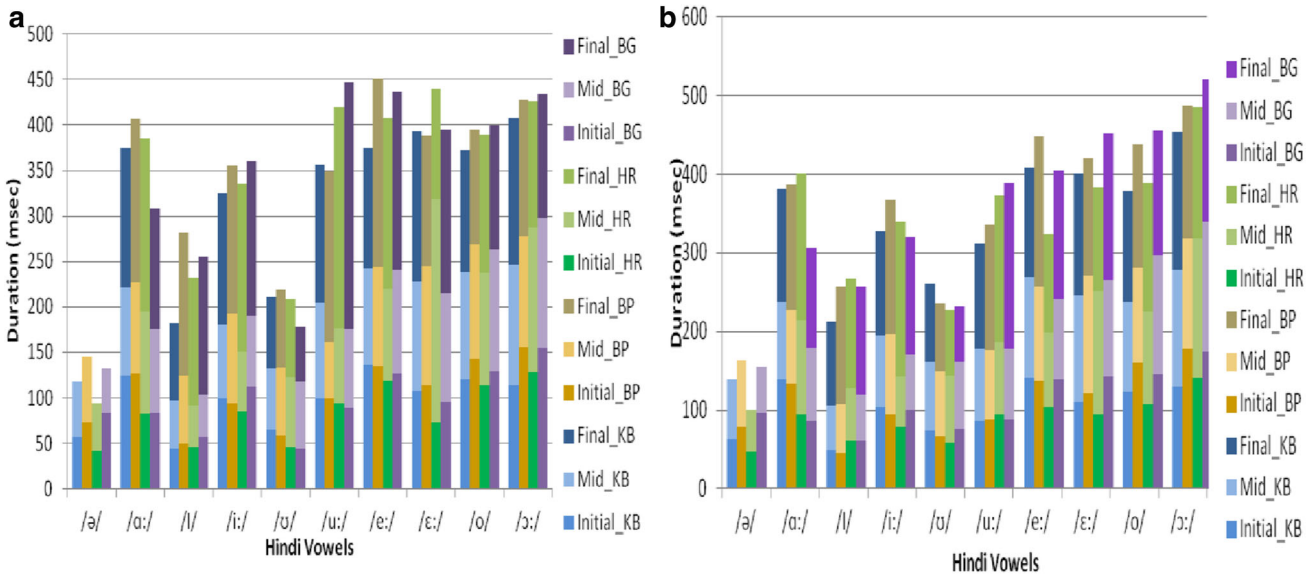

Fig. 5 Average duration of vowels spoken by **a** Male **b** female speaker at all three word position

vowel /ɑ:/, average duration in Khari boli dialect differs from Haryanvi and Bagheli dialect speakers at the initial word position and is approximately equal to the Bhojpuri, dialect speaker. Also, it is observed that the average duration of Bagheli speakers is quite low for the vowel /ɑ:/ as compared to others. For both /I/ and /i:/ vowels, the duration is very small at the final word position in Khari boli dialect than any other dialect. It is further observed that Haryanvi speakers and Bagheli speakers take equal time to utter /ʊ/ and /u:/ at the word-initial and mid position. At the end word position, average duration for /ʊ/ is more for Haryanvi speakers and for /u:/ it is more for Bagheli speakers. For /o/ and /ɔ:/, not much difference in the duration due to dialects can be observed at any word position. For the female speakers, Fig. 5b outlines the significant difference in the average duration /ə/ and /ɑ:/ for Bagheli and Haryanvi dialect. For the vowel /u:/, the average duration of Khari boli speakers is the smallest and is highest for Bagheli speakers. Similar to their male counterpart average duration of vowel /e:/ is the longest for Bhojpuri speakers. Further analysis shows that the duration for /o/ in Khari boli dialect at all word position is smallest and is maximum in Bagheli dialect.

One-way ANOVA with $p = .05$ revealed that significant main effect ($p < .05$) of dialects on average duration of most of the vowels is observed. Only vowels, /i:/ and /ʊ/ with $p = .089$ and $p = .184$ respectively at the final word position and the vowels /ɛ:/ and /o/ with $p = .616$ and $p = .898$ respectively at the middle word position shows no statistically significant difference due to dialects. Post-hoc test show that even though initial result of ANOVA shows significant main effect of dialect on vowels /ɔ:/ ($p = .002$) and /u:/ ($p = .004$) at the initial position, the post hoc test reveals that no significant difference between Bhojpuri and Bagheli dialect exist for /ɔ:/ ($p = .983$) and Khari boli and Bhojpuri differ significantly for /u:/ ($p = .613$) It is further observed that the vowel /ʊ/ at the middle word position has a significant difference ($p = .516$) in the dialect Bhojpuri and Bagheli.

The statistical analysis of female data for the duration of vowel outlines almost similar characteristics of vowels as for male speakers. However, even though ANOVA findings show a non significant effect of dialect on the duration of vowels /u:/ ($p = .253$) at word initial position and, /i:/ ($p = .108$) and /ʊ/ ($p = .306$) at the word final position post hoc test revealed that Khari boli and Haryanvi ($p = .041$), Haryanvi and Bhojpuri ($p = .002$) are statistically different for /i:/ at word final position. Also, vowels /ɑ:/ and /u:/ at mid positions are not significant for Bhojpuri-Bagheli and Khari boli-Bagheli dialects.

The results of ANOVA and post hoc test performed for the duration of ten vowels indicated clearly that for most of the vowels one group of speakers were different from another group of speakers. It can be analyzed that for most of the vowels, the duration can work as a distinguisher for identification of dialects.

3.6 Intensity analysis

The perceived loudness of any speech signal is its intensity. It is measured as the sound power per unit area. Considering dialects as between-subject factor One-way ANOVA was executed to obtain the significance of the intensity (mean intensity) on ten Hindi vowels.

The statistical analysis of mean intensity using one-way ANOVA revealed a significant difference between the

vowels /ə/ [F(3,116) = 37.852, p < .001] and /ɔ:/ [F (3,116) = 8.631, p < .001]. The post hoc test on these data revealed a similar result. The result obtained for the vowels at the middle word position revealed a considerable difference for the vowels /ə/ [F (3,116) = 20.707, p < .001], /ɑ:/ [F (3,116) = 16.215, p = .042], /i:/ [F (3,116) = 15.536, p = .033], /u:/ [F (3,116) = 9.346, p = .022] and /ɛ:/ [F(3,116) = 9.764, p = .036]. Most of these are long vowels and influences stress differently due to regional accent. At the final word position the vowels /I/ (p = .001), /ʊ/ (p = .001) and /e:/ (p < .001) show intensity based distinction due to dialects. Similar results for these vowels were obtained from the analysis of female speakers' data.

The statistical findings disclose that the stress on long vowels is generally put differently at the mid of the word by different dialect speakers. Also, the short vowels (/I/, /ʊ/, /e:/) are stressed differently by dialects at the final word position. No major distinction due to dialects is obtained for stress at other word positions on most of the vowels. The results of analysis draw attention to the peculiarities in vowel duration due to dialects. It is further obtained that though average pitch value can not distinguish the dialect of the spoken utterance; its dynamics gives notable results. It was observed on the existing speech corpus that the second and the third formant frequency are a better candidate than F1 for discriminating the dialects, but formant frequencies and intensity are selective in nature and behave differently for different vowels and dialects.

4 Speech feature for dialect identification

Speech signal not only endows linguistic messages but also several paralinguistic attributes defining spoken aspect is contained in the signal. Features extracted at different levels of speech can be effectively used to identify some of the speaker's characteristics. At the segmental level, dialect particular data can be seen as an arrangement of distinctive sequence of the vocal tract shapes for delivering different sound units. These distinctive sequences are characterized by spectral envelope of the speech signal that is represented by the spectral features. The spectral feature represents the linguistic content of the signal but the overall speech quality can be represented in terms of intonation, energy, duration, loudness and so on. All these attributes are affected by speaking style of speakers. These attributes are meaningful only if they are extracted from longer segments of speech, may be sentences, words or syllables. These long segment features are termed as prosodic features. The spectral and prosodic features extracted at different level of speech is modeled and fed to the classifier for identification of speaker's dialect. In this section we evaluate the efficiency of different spectral and prosodic features for ADI and measure the data distribution capturing ability our proposed models.

4.1 Spectral features

Acoustical analysis highlights that Hindi dialects differ in vowel space. It is also highly likely that they will significantly differ in their spectral distribution and thus can be exploited for automatic identification of dialects. For the spectral features 13 static MFCC coefficients are extracted by dividing speech segment into successive overlapping frames of 20 ms with an overlap rate of 10 ms. From all the obtained frames, the silence frames are removed based on amplitude threshold obtained from the available samples. From these frames static features are obtained. To capture temporal variations SDC features are obtained over the combination of multiple frames. The SDC parameter used in this task is 13-1-2-2. This value has been achieved by running several passes to obtain the best performance. These SDC features were combined with 13 MFCC features. Total 39 dimensional feature set was obtained as spectral features. tenfold cross validation is used for evaluating the systems, where each fold consists of 3 male and 2 female speakers from each of the four dialects. The final output is obtained as the average of the scores obtained from all the folds.

4.2 Prosodic features

In the literature it is shown that human being rely on intonational cues (Peters et al. 2002). Barkat et al. presented that eastern and western Arabic dialects can be distinguished significantly on the basis of intonation alone (Barkat et al. 1999). Hamdi et al. outlines that rhythmic differences occur between eastern and western Arabic dialects (Hamdi 2004). Comparing percentage of vocalic intervals and standard deviation of inter-vocalic intervals among the speakers of two dialects can give information regarding these characteristics. Ljolje and Fallside have used fundamental frequency, their derivatives and energy for discriminating the native and non-native speakers of English (Ljolje and Fallside 1987).

Prosodic model of dialect classification is based upon the hypothesis that dialects of any language differ in their prosodic distribution. Acoustic feature analysis of Hindi vowels highlight that Hindi dialects differ in their vowel space. It is also highly likely that they will significantly differ in their prosodic distribution and thus can be exploited for automatic identification of dialects. Syllables are assumed to have close connection with human speech perception and articulation (Ganapathiraju et al. 2001). The prosodic features in the present work are extracted from the

syllables. In order to further process the speech samples using syllables contained in it, the signal must be segmented at the syllable level and aligned with phonetic transcriptions. From these syllables silence frames are removed with the assumption that all the non-silent frames are valid and from the non-silent frames prosodic features are extracted.

4.2.1 Prosodic feature extraction

The acoustic realization of prosody can be observed and quantified using fundamental frequency, energy and duration (Rao et al. 2012). Analysis of pitch and pitch slope using one-way ANOVA shows that the four Hindi dialects differ for these features. Vowel duration is also dependent on location of pauses, the word and syllable boundaries, as well as manner of articulation. Since the manner of articulation in dialects is different, phonetic duration differences occur among dialects. Analysis of vowel duration has shown considerable differences among Hindi dialects. Energy level of the speech signal helps in identifying the voiced/unvoiced part of speech. Stress pattern of speakers can be represented by combining energy with pitch and duration. In literature (Biadsy et al. 2011; Koolagudi et al. 2009; Sreenivasa and Yegnanarayana 2009) these features have been shown as are good correlates of prosodic features. In this research local prosodic features extracted from syllables have been considered for the classification of dialects and the decision is based upon the cumulative score obtained over all the syllables.

For every syllable four pitch based features; f_{max}, f_{min}, f_{mean} and f_{slope} are extracted. The f_{max} and f_{min} values are the maximum and minimum pitch values obtained over the syllable, f_{mean} is the mean pitch obtained over all the frames in the syllable and the f_{slope} over the syllable is computed as the absolute difference of f_{max} and f_{min} divided by the time duration between the two points. Energy of each overlapping frames of syllable is obtained by summing the squared amplitude of each sample. For the energy feature four values; E_{max}, E_{min}, E_{mean} and E_{range} is extracted from the syllables as above. E_{range} is obtained as the difference between the maximum and the minimum energy over the syllable. Syllable duration is measured in milliseconds.

5 Modeling and evaluation of speech features

Assuming that the acoustic features for each dialect are different, the acoustic models exploit these differences to categorize the input data into groups. These models for each dialect are created from the speech features, be it spectral or prosodic or both. The training samples from each dialect are used to estimate the parameters of the model. The dialect dependent models are further used to produce scores for their classification.

Success of statistical methods based on hidden Markov models (HMM) in tasks in speech and natural language domain has laid to a new quest for more powerful recognition methods with total dedication towards increasing robustness of classifier while reducing error in classification. Discriminant approaches followed by SVM for pattern classification has gained prominence in this respect. Torres-Carasquillo et al. (Torres-Carrasquillo et al. 2004) have used discriminatively trained Gaussian-mixture models-Universal background models (GMM-UBM) with shifted delta cepstral (SDC) features. Due to reduced number of parameters in GMM, its training and testing is faster compared to HMMs. GMM is used for spectral feature modeling and multi-class SVM classifier is implemented for accent classification task by Lazaridis et al. (1998). Hanani et al. (2013) have applied LID technique to British English accent classification. Recent research in this direction is focused on kernel-based approach, where the features are modeled using GMM and SVM is used as a classifier. Biadsy et al. (2011) used GMM-super vectors extracted for each phone type with SVM classifier for identification of Arabic English accent. SVMs are discriminative classifier that depends upon the number of support vectors with discriminative characteristics. They are suitable for less number of feature vectors. In reality, there is no strong recommendation of any of these classifiers to be used as accent classifier. Owing to the fact that each has their own merits and demerits it is advisable to combine multiple classifiers to obtain the final results.

5.1 SVM-GMM model for spectral feature evaluation

Support vector machines are supervised learning models. Apart from performing linear classification, SVMs are capable of doing non-linear classification using kernel functions. SVMs have been applied successfully on several kinds of classification problems and have consistently performed better than other non-linear classifiers like neural networks and mixtures of Gaussians (Robinson 1989). Due to their inefficiency to model unequal length input data its usage has been very limited in speech classification. To deal with this limitation of SVM, help of a generative model: GMM is utilized. The varying length speech features are modeled using GMM to produce a fixed length data. For using GMM-SVM approach we utilize the GMM-UBM method for model adaptation. Since most of the work done for accent and language classification show no improvements in the results by adapting the GMM weights and covariance (Rifkin 2008), therefore only

Table 1 Dialect classification performance of RBF kernel based on MFCC + SDC features

gamma(γ) C = 20	Classification error (%)				
	Khari boli	Haryanvi	Bhojpuri	Bagheli	Average error (%)
.4	41.3	40.6	43.2	42.8	41.98
.5	39.7	41.1	40.6	42.3	40.93
.6	38.4	36.6	37.8	38.1	37.73
.7	33.7	33.4	34.2	33.1	33.6
.8	33.7	32.1	33.5	32.8	33.03
.9	34.2	33.6	34.1	34.8	34.18
1.0	37.9	40.7	36.6	41.7	39.23
2.0	46.6	47.3	46.4	45.3	46.40

GMM mean vectors, are adapted and concatenated to form supervectors. All the dialect data available for this research is used to ML-train the UBM using EM Algorithm. A dialect dependent 512 component GMM is created for entire training data corresponding to the dialect *m* by MAP adapting the means of the UBM using a relevance factor $r = 16$. The value of r keeps the balance between old and new estimates. A GMM supervector is then obtained by concatenating the 512 mean vectors corresponding to each utterance. These supervectors are then used with the SVM kernels to create dialect dependent SVM model. While SVMs reduces the complexity in data by converting it to high dimensional feature space, it also introduces the computational and generalization problems. These problems are handled by introducing kernel-tricks for classification. The Radial Basis Function (RBF) (Eq. 1) is

$$K(x_i, x_j) = \exp\left[-\frac{1}{2}\left(\frac{\|x_i - x_j\|}{\phi}\right)\right]^2 \quad (1)$$

investigated for its performance in this work.

5.1.1 Score computation

By using any of the one-vs-one or one-vs-all approach SVM can be utilized as a multi-class classifier. Although the number of classifiers (n(n − 1)/2) is far more in one-vs-one approach as compared to n number of classifier in one-vs-all approach, the former is assumed to be faster and memory efficient as compared to the other approach (Rifkin 2008). One of the reasons for this is that the training data size for each classifier is reduced and requires less resource during training. This work utilizes one-vs-one approach for the classification of four dialects and has implemented six binary classifiers. All these SVM classifiers are trained with GMM supervectors to generate dialect dependent models. During testing, each test utterance is represented as supervector by MAP adapting the means of UBM. These supervectors are fed into dialect dependent SVM models and final decision is taken based on 'Max-Win' strategy.

5.1.2 Evaluation of spectral features

For the tuning of RBF Kernel the parameter of interest is γ; the variance of the kernel and the parameter C used to penalize the errors associated with training. It is not known in advance that what combination of these values will give the best result of the problem in hand. The goal is to obtain (C, γ) such that any unknown data can accurately be predicted. Applying grid-search on C and γ, the final values were obtained. C was changed repeatedly, starting from 1 in the steps of 10 to obtain the best result and then classification accuracy for different γ was obtained. The performance reported is the Table 1 is the values averaged over ten folds. With the RBF kernel the best average percentage classification for MFCC + SDC features is obtained to be 66.97 %.

The best result is obtained for the Bhojpuri dialect. This may be due to strong geographic proximity among the speakers of this dialect. The results obtained by kernel methods outperform the results obtained on the same database using auto-associative neural network AANN (Sinha et al. 2015).

5.2 SVM for prosodic feature modeling

Support vector machine takes fixed size input in each iteration. For prosodic distribution based dialect identification fixed size input data is obtained from each syllable. This revokes the requirement for GMM. For the prosodic features only SVMs are used. The classifier for each of the prosodic feature set is trained individually to capture the prosodic distribution of each dialect. As with the spectral features, for the prosodic features also evaluation is done using RBF kernels.

5.2.1 Score computation

One-vs-one approach is used to implement the classifier. During testing, syllables are assigned one of the two classes

by all the classifier. The class identity for any syllable is decided on 'Max-Win' strategy. And the class for the test utterance is based upon the count of syllables belonging to each class. The class with maximum number of syllables is selected as the class identity for the test input.

5.2.2 Evaluation of combined prosodic features

To evaluate the efficiency of prosodic features, feature level fusion of all the prosodic parameters were obtained by simple concatenation of values based on syllables. The feature stream was fed to the system for classification. System performance was evaluated for the RBF kernel. The results obtained for combined prosodic feature is presented in Table 2. The best performance is obtained at (20, .8) pair.

Results based on prosody information in the speech signal shows that four Hindi dialects studied in these research exhibit strong differences from one another in terms of their prosodic characteristics. Experimental results show that prosodic features, including pitch range, pitch slope, syllable duration and energy can automatically identify dialect of a speaker to good extent and gives accuracy up to 74 %. Such accuracy strongly indicates that prosody alone can guarantee good identification to the spoken utterances. Further improvements can be obtained by combining the spectral and prosodic aspects of dialect.

5.3 Combined spectral and prosodic features

Results obtained in previous sections show that prosodic features even when used alone give good recognition for Hindi dialects. These points to their value for distinguishing Hindi dialects. Spectral features have also shown their contribution toward the identification of dialects. Speech signal exhibit dialect or accent specific features at different levels. Although, for the present task spectral features do not seem to be as much promising as the combination of prosodic features, we need to check if these features add some value to the ability of prosodic features or not. In this section, we explore the combined effect of the two. The two models are tested for the combination of spectral and prosodic features. Figure 6 represents the SVM model for evaluation of combined effect of spectral and prosodic feature.

For the evaluation of the combination of spectral and prosodic features using kernel methods, SVMs discriminative ability is exploited. For processing the spectral features GMM supervectors are created. The prosodic features are treated directly by the SVM. For both the feature sets kernel function returns corresponding classification score by each dialect model. The final decision is based on 'Max-Win' strategy, giving more weight to prosodic features.

Table 3 presents the classification error for different values of gamma. With the RBF kernel the average percentage classification was obtained to be 88.77 % for ($C = 20$, $\gamma = .8$) pair.

Even though prosodic features can give good discrimination among the Hindi dialects, the empirical results show that combination of spectral and prosodic features give better evidence for discrimination of dialects as compared to using these features separately, also the results obtained by SVM-GMM model outperforms AANN model (Sinha et al. 2015).

6 Human perception of dialects

100 listeners were identified for the perception test of the recorded samples. The listeners were registered for the test by providing basic information about their name, age, gender, years of acquaintance with the dialect etc. Table 4 presents the listeners statistics. In communication through speech considerable variability comes into existence that may be controlled by regional or social belonging of speakers. For the essence of good communication it is desirable that human must cope with the variability. Listener's judgment regarding speaker's speaking style is

Table 2 Dialect classification performance of RBF kernel based on combined prosodic features

gamma (γ) C = 20	Classification Error (%)				
	Khari boli	Haryanvi	Bhojpuri	Bagheli	Average error (%)
.4	39.1	35.8	37.7	38.3	37.65
.5	35.3	37.0	36.3	35.7	36.08
.6	34.3	36.3	33.4	34.1	34.53
.7	30.7	33.2	31.1	32.8	31.95
.8	26.7	27.5	28.3	28.1	27.65
.9	32.3	30.2	30.5	31.3	31.08
1.0	36.5	35.3	37.1	39.2	37.03
2.0	37.5	38.2	41.4	40.6	39.43

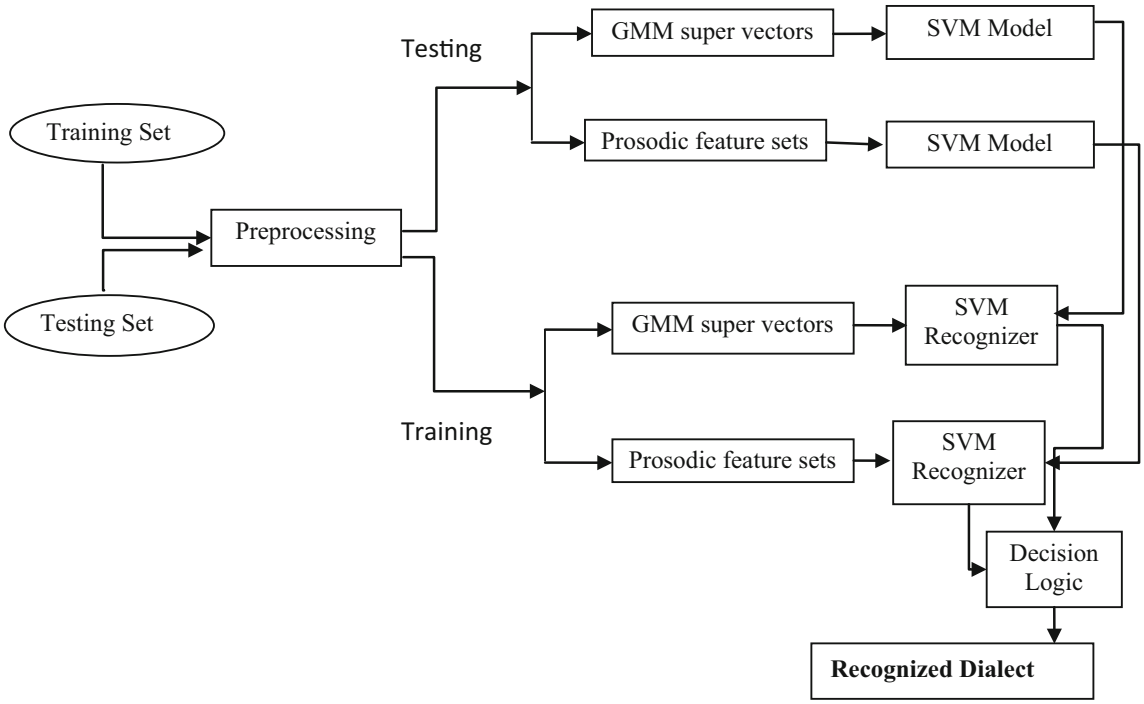

Fig. 6 SVM model for classification of Hindi dialects using spectral and prosodic features

Table 3 Dialect classification performance of RBF kernel for Hindi dialects based on spectral and prosodic features combined

gamma(γ) C = 20	Classification error (%)				
	Khari boli	Haryanvi	Bhojpuri	Bagheli	Average error (%)
.4	15.9	14.7	15.1	19.0	16.18
.5	14.6	13.8	13.2	17.4	14.75
.6	12.4	12.9	11.2	14.6	12.78
.7	11.7	12.2	9.4	14.1	11.85
.8	11.1	11.6	9.0	13.2	11.23
.9	11.3	11.9	9.0	13.4	11.4
1.0	12.6	13.1	10.8	15.2	12.9
2.0	16.7	19.4	18.6	18.0	18.12

Table 4 Summary of listener's information

Listener's Statistics				
	Khari boli dialect	Haryanvi dialect	bhojpuri dialect	Bagheli dialect
No. of listeners	26	23	28	23
Age	16-48 yrs	21-44 yrs	24-55yrs	19-41 yrs
Male listeners	14	10	12	12
Female listeners	12	13	16	11

guided by his/her perception. With the aim to quantify the acoustic features selected in this research for dialect identification, comparison of our computer based model is done with human performance.

Different sets of sound were played to different groups. Apart from playing some random sample for the listeners to help them acquire some knowledge about the speech and text corpus, no explicit training in recognition was given to the subjects. In order to be sure of their decision if the listeners demanded, same sounds were played more than once. The mix of the samples was created considering the dialect, gender and length of spoken utterances. The

Table 5 Dialect recognition performance of perception test by human

Hindi dialects	Human perception of dialects (%)			
	Khari boli dialect	Haryanvi dialect	Bhojpuri dialect	Bagheli dialect
Khari boli	84	08	03	05
Haryanvi	06	78	07	09
Bhojpuri	02	08	83	07
Bagheli	08	06	07	79

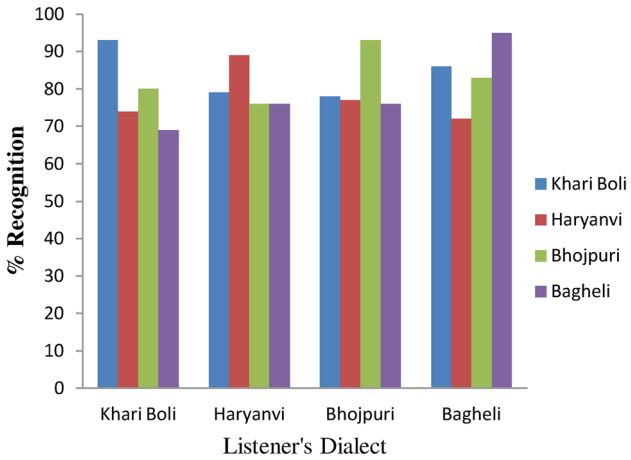

Fig. 7 Study of listener's dialectal acquaintance on perception of dialect

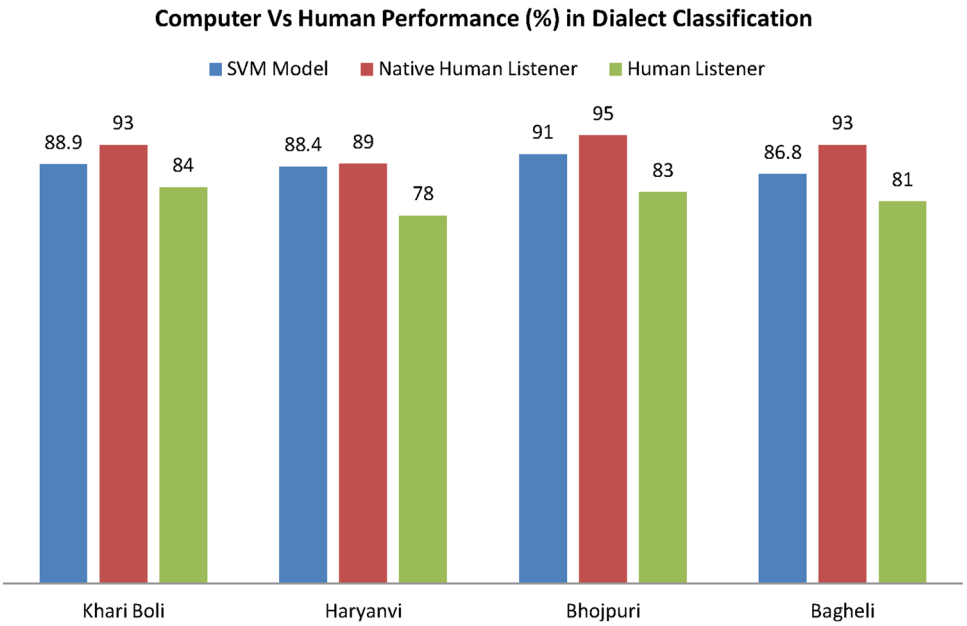

Fig. 8 Performance Comparison of man and machine for automatic dialect classification

average human perception of the data was obtained to be 81.5 % (Table 5).

Figure 7 represents the analysis result of the test conducted to study the influence of listener's dialectal acquaintance on his perception of dialect. The graph highlights that the native speakers of any dialect can identify dialect better than the non native speakers. From the analysis of the result it was further observed that the classification of isolated utterances by human was much error prone as compared to continuous sentences.

7 Conclusion and future work

In this paper the problem of automatic dialect classification from the spoken utterances of Hindi language is considered. In order to study the dialectal influence on the acoustic characteristics, 10 Hindi vowel sounds were first investigated to obtain an insight into the similarities and dissimilarities of these sounds in dialects of Hindi. Statistical analysis of the acoustic parameters; the first three formants, fundamental frequency, pitch slope, duration and

intensity was done using one way ANOVA. The results of this analysis showed that different formant frequencies of these vowels are influenced distinctly by different dialects. Results based on analysis of average pitch highlights that no crisp decision regarding the spoken dialect can be obtained by these values, but their dynamics results in promising outcome. Study of variations in pitch over time showed that pitch slopes for different vowels vary significantly in different dialects. Intensity being speaker dependent characteristics, the average intensity did not give any substantial information regarding the spoken dialect. The analysis further showed significant effect of dialect on the duration of vowel sounds. The distinctive features identified for distinguishing the dialects were evaluated for their efficiency using SVM-GMM model. RBF kernel is employed to check feature performance. The results highlights that the spectral feature MFCC with SDC are able to capture dialectal information from the speech signal to some extent. The best performance with this feature set was obtained to be 66.97 %. The results highlight that prosodic features were more efficient than the spectral feature in capturing dialectal characteristics from the speech signal. A recognition accuracy of 74 % was obtained with the combination of all prosodic features. System performance was compared with human perception of dialects. Human recognition score was obtained to be 81.5 %. Detailed analysis of the perception test highlight that human's acquaintance with the dialect influences their perceptual ability for that dialect. The results highlight that perception of dialect by the listeners of that dialect is noteworthy, but perception by listeners from different dialect does not give comparable result. Figure 8 represents the comparative performance chart for the computer models and human listeners.

For future study, glottal closure and other excitation source features can be explored for the dialectal distinction. The combination of these features with the spectral and prosodic features should be exploited. Including vocal tract length normalization technique to remove speaker dependent information can enhance system performance. As no standard database for studying dialectal characteristics of Hindi speech exists, this study is based upon self created database comprising of utterances from only four dialects of Hindi. This corpus should be extended to capture more Hindi dialects as well as increase the number of subjects for each dialect.

References

Adank, P., Van Hout, R., & Van de Velde, H. (2007). An acoustic description of the vowels of northern and southern standard Dutch II: Regional vari- etiesa. *The Journal of the Acoustical Society of America, 121*(2), 1130–1141.

Aggarwal, R. K., & Dave, M. (2012). Integration of multiple acoustic and language mod- els for improved Hindi speech recognition system. *International Journal of Speech Technology, 15*(2), 165–180.

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

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

Biadsy, F., Hirschberg, J. B. & Ellis, D. P. W. (2011). Dialect and accent recognition using phonetic-segmentation supervectors. In *INTERSPEECH* (pp. 752–756).

Cho, T., & Keating, P. A. (2001). Articulatory and acoustic studies on domain-initial strengthening in Korean. *Journal of Phonetics, 29* (2), 155–190.

Ganapathiraju, A., Hamaker, J., Picone, J., Ordowski, M., & Doddington, G. R. (2001). Syllable-based large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing, 9*(4), 358–366.

Grover, C., Jamieson, D. G., & Dobrovolsky, M. B. (1987). Intonation in English, French and German: perception and production. *Language and Speech, 30*(3), 277–295.

Hamdi, R., Barkat-Defradas, M., Ferragne, E. & Pellegrino, F. (2004). Speech Timing and Rhythmic structure in Arabic dialects: A comparison of two approaches. In *INTERSPEECH* (Vol. 4, pp. 1613–1616).

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.

Koolagudi, S. G., Maity, S., Kumar, V. A., Chakrabarti, S., & Rao, K. S. (2009). IITKGP-SESC: speech database for emotion analysis. In *Contemporary computing* (pp. 485–492). Springer.

Kulshreshtha, M., & Mathur, R. (2012). *Dialect accent features for establishing speaker identity: A case study*. New York: Springer.

Kumar, M., Rajput, N., & Verma, A. (2004). A large-vocabulary continuous speech recognition system for Hindi. *IBM journal of research and development, 48*(5.6), 703–715.

Lazaridis, A., Goldman, J.-P., Avanzi, M. & Garner, P. N. (2014). Syllable-based Regional Swiss French Accent Identification using Prosodic Features. In Nouveaux cahiers de linguistique francaise, Number EPFL-CONF-199821.

Ljolje, Andrej, & Fallside, Frank. (1987). Recognition of isolated prosodic patterns using Hidden Markov models. *Computer Speech & Language, 2*(1), 27–34.

Mishra, D. & Bali, K (2011). A comparative phonological study of the dialects of Hindi. In *Proceedings of ICPhS XVII*, *Hong Kong* (pp. 17–21)

Pandey, P. K. (1989). Word accentuation in Hindi. *Lingua, 77*(1), 37–73.

Peters, J., Gilles, P., Auer, P., & Selting, M. (2002). Identification of regional varieties by intonational cues: An experimental study on Hamburg and Berlin German. *Language and Speech, 45*(2), 115–138.

Rabiner, L., & Juang, B.-H. (1993). *Fundamentals of speech recognition*. Upper Saddle River: Prentice hall.

Raman, S. (1985). Speech recognition of Hindi stop consonants. PhD thesis, Indian Institute of Technology, Madras, 1985.

Rao, P. V. S. (1993). VOICE: An integrated speech recognition synthesis system for the Hindi language. *Speech Communication, 13*(1), 197–205.

Rao, K. S., & Koolagudi, S. G. (2012). *Emotion recognition using speech features*. New York: Springer.

Rifkin, R. (2008). Multiclass classification. http://www.mit.edu/9.520/spring09/Classes/. Accessed 20 Sept 2014.

Robinson, A. J. (1989). Dynamic error propagation networks. PhD thesis, University of Cambridge.

Sekhar, C. C., & Yegnanarayana, B. (2002). A constraint satisfaction model for recognition of stop consonant-vowel (SCV) utterances. *IEEE Transactions on Speech and Audio Processing, 10* (7), 472–480.

Sinha, S., Agrawal, S. S. & Jain, A. (2013) Dialectal influences on acoustic duration of Hindi phonemes. In *Conference on Asian spoken language research and evaluation (O- COCOSDA/ CASLRE)* (pp. 1–5). IEEE, Piscataway.

Sinha, S., Jain, A., & Agrawal, S. S. (2015). Fusion of multi-stream speech features for dialect classification. *CSI Transactions on ICT, 2*(4), 243–252.

Sreenivasa, K. S., & Yegnanarayana, B. (2009). Intonation modeling for Indian languages. *Computer Speech & Language, 23*(2), 240–256.

Torres-Carrasquillo, P.A., Gleason, T. P. & Reynolds, D. A. (2004). Dialect identification using Gaussian mixture models. In *ODYSSEY 04-the speaker and language recognition workshop* (pp. 297–300).

Wells, J. C. (1982). *Accents of English* (Vol. 1). Cambridge: Cambridge University Press.

Yan, Q. & Vaseghi, S. (2003). Analysis, modelling and synthesis of formants of British, American and Australian accents". In *Proceeding acoustics, speech, and signal processing* (Vol. 1, pp. I–712). IEEE, Piscataway.

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