CSIT (January 2015) 2(4):243–252
DOI 10.1007/s40012-015-0063-y

ORIGINAL RESEARCH

# Fusion of multi-stream speech features for dialect classification

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

Received: 12 September 2014 / Accepted: 31 May 2015 / Published online: 18 June 2015

**Abstract** Current research in the area of voice recognition has entered a new stage. It does not only concentrate on the correct evaluation of linguistic information embodied in the speech signal, it also works towards identification of variations naturally present in speech. Undoubtedly, the focus is to enhance the accuracy and precision of the developed technique. Speaker's accent due to his native dialect is one of the major source of variability. Prior knowledge of the spoken dialect will help in the creation of multi-model speech recognition system and can enhance its recognition performance. This paper focusses on applying some of the established dialect identification techniques to identify speaker's spoken dialect among dialects of Hindi. Fusion of multiple streams obtained as a combination of phonetic and prosodic features is implemented to exploit the acoustic information. The work presented here also exploits the ability of AANN to capture the distribution of data points in a reduced number and further to classify them into groups. System performance for different level of fusion is recorded for Hindi dialect classification. It is observed that Duration as prosodic feature is an important cue for automatic dialect identification systems.

✉ Shweta Sinha
meshweta_7@rediffmail.com

Aruna Jain
arunajain@bitmesra.ac.in

S. S. Agrawal
ss_agrawal@hotmail.com

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

[2] KIIT Group of Institutions, Gurgaon, India

**Keywords** Dialect identification · Auto-associative neural network · Feature combination · Hindi dialects · Prosodic features · Multi-stream fusion

## 1 Introduction

Current research in the area of automatic speech recognition (ASR) is focusing on understanding and modeling variations in spoken language. Speaker's native dialect, accent, and socioeconomic background highly influence their speaking style. The speaker will carry the trait of this style when speaking some other language or even the standard form of his language. The differences so caused introduce difficulties for modeling input speech in the development of speaker-independent systems. Automatic identification of speaker's dialect from the input speech can improve ASR performance. Apart from this, identification of dialect in spoken utterance can be used as biometric information for speaker's identity.

Dialect of a given language is a variety of that language with systematic phonological, lexical and grammatical characteristics followed by speaker's belonging to same geographical area. The task of automatic dialect identification (ADI) is to identify speaker's native dialect of a predetermined language using the acoustic signal alone. Dialectal studies can be based either on acoustic approach or the phonotactic approach. The former approach deals with the differences between the distribution of sounds in dialects, and the latter is concerned with the sequence of occurrence of sounds in different accents. It is a difficult problem as speakers from same dialect show major variations among themselves. Also, intra-speaker variations can be observed due to a different state of emotions. Due to this problem ADI is treated as more challenging than that of language identification (LID).

Due to the similarity of this task with LID task most of the research in this area takes the lead from research done in LID. Several successful approaches of LID field have been extended to this problem. Zissman et al. [1] have extended the approach of parallel phone recognition followed by language modeling (PRLM) to the classification of two Spanish dialects. They have used phonotactic information regarding the language. Kumpf and King [2] proposed accent specific HMMs and phoneme-based bigram language model for accent classification. Trained HMMs have been used for segmentation of accented speech for further training the classifier. This system gave good recognition performance. Torres-Carasquillo et al. [3] have used discriminatively trained Gaussian-mixture models-Universal background models (GMM-UBM) with shifted delta cepstral (SDC) features. It is combined with vocal tract length normalization for the recognition of English, Chinese and Arabic dialects. Huang et al. [4] proposed word based modeling technique for recognition of dialects. They have used GMM-HMM approach for acoustic modeling and have shown that small set of words accounts for large training database. Due to this they have advocated that only this small set of word will suffice for acoustic model training. Wells [5] in his findings outlined that variation in accent stretch out in phonetic as well as speaker's prosodic characteristics. Arslan et al. [6] used mel-frequency cepstral coefficients (MFCC) and energy as speech features for English accent classification. Relevance of prosodic information was evaluated by Rouas et al. [7] for LID also. Motivated by the improved performance of such systems showing that low-level features such as MFCC alone cannot provide sufficient discriminating information, Yan et al. [8] have applied formant vectors instead of MFCC to train GMM and HMM for accent identification of American, Australian and British English. Ma et al. [9] explores MFCC features with multi-dimensional pitch flux. They have used GMM for distinguishing three Chinese dialects. Alorfi [10] explores ergodic HMMs to model phonetic differences between two Arabic dialects.

Intonational cues are considered to be useful indicators for humans for identification of regional dialect. Many works have been done to exploit the influence of dialect on intonational cues [11, 12, 13]. Artificial neural networks (ANN) for their ability to capture the inherent non-linearities in the input signal have also been used in many of the dialect recognition work [14, 15].These systems are easy to implement and perform well on a small set of data. Recent research in this direction is focused on kernel-based approach, where the features are modeled using GMM and support vector machine (SVM) is used as a classifier. In Biadsy et al. [16] used GMM-super vectors extracted for each phone type with SVM classifier for identification of English accent. Multi-layer feature combination with SVM classifier is used for Chinese accent identification [17]. In this work, they have used both the segmental and the supra-segmental information to capture the diversity of the variations within accent. MFCC and log energy with higher order derivatives is modeled using GMM with multi-class SVM classifier in [18].

The work presented in this paper focus on applying some of the established dialect identification techniques to identify speaker's spoken dialect among dialects of Hindi. Most of the work done for AID deals with languages of western countries and are based on the study of influence of pronunciation of L1 on L2. This paper deals with the regional dialects of Hindi and its influence on utterances of standard Hindi. Fusion of multiple streams obtained as a combination of phonetic and prosodic features is implemented to exploit the acoustic information. In [19], Auto-associative neural networks (AANN) have been proven to be the substitute of GMMs for pattern classification problem. This paper also exploits the ability of AANN to capture the distribution of data points in a reduced number and further to classify them into groups. System performance for different level of fusion is recorded.Few dialect based studies using suprasegmental features of speech have been done. In [15], classification of accents of speakers have been done using a database collected from people who are non-native speakers of Hindi. Study of the impact of Hindi due to their mother tongue; i.e. L1 is studied. In [20], dialect classification of isolated utterances is done and multilayer feed forward neural network is used as a classifier.

Following this introductory section, rest of the paper is arranged as follows: Section 2 describes speech database for this work. Front end Speech features used for dialect identification is explained in Sect. 3. Multi-stream feature fusion is presented in Sect. 4. Section 5 describes the classification model, Sect. 6 evaluates system performance, and conclusion of the work is discussed in Sect. 7.

## 2 Speech database

Speech related research are based on data-driven technology and requires a large amount of labeled data. These data are used for training acoustic model. Contrary to major European and American language with huge speech corpus in the public domain Hindi has no standard text and speech corpora for researchers. Individuals or research group working in this field have created databases for fulfilling their requirements. For the study of regional Hindi dialects, no such database is available for use. Lack of such resource for Hindi is the major hurdle in speech processing research for this language. Hindi is mainly spoken by people in North and Central India. There are around fifty dialects of

Hindi. Huge dialectal diversity can be seen in these dialects. Due to geographical and lingual background variations can be easily observed in dialects also. Recording considerable number of speech samples to cover the variations from all the dialects is itself a challenge. This work is based upon four major dialects of Hindi, namely; Khari Boli dialect (KB) (Delhi and parts of neighbouring states), Haryanvi (HR) (Haryana and part of Delhi), Bhojpuri (BP) (East Uttar Pradesh, Bihar and Jharkhand) and Bagheli (BG) (Madhya Pradesh and parts of Chhattisgarh). To minimize the inter dialectal variations, the speakers are selected from close geographical propinquity. A database of 300 continuous sentences spoken by 28 male and 20 female speakers from each of the four dialect is created.

The aim of this work is to study the influence of dialects on pronunciations of Hindi. A text corpus of continuous sentences, consisting of all Hindi phonemes is created using Devanagari script based on Khari Boli (considered as standard Hindi) dialect. Read speech samples are recorded in a soundproof room using Gold Wave tool and are sampled at 16 kHz. All the selected speakers are of the age group 18–50 years.

## 3 Feature front-end

Speech in its digital form is variable and is of very high dimension. Also, it is impractical to model them directly with available modeling techniques. Feature extraction process is required to convert the speech waveform in a form usable for further processing. The main goal of feature extraction in ASR is to preserve the key lexical information while suppressing the non-lexical variations.

Extraction of proper acoustic features that can efficiently characterize speaker's accent/dialect is an important issue in the design of ADI. Research in [21], highlights that quality and robustness of speech features control the performance of any ASR systems. Acoustic features can be categorized as phonetic and prosodic features. Both parametric (i.e. MFCC) and non-parametric (LPC, PLP) types of phonetic features have been used in several speech based identification task [6, 8, 15]. This paper uses MFCC and SDC as phonetic feature.

### 3.1 Acoustic features for ADI system

#### *3.1.1 Mel frequency cepstral coefficient (MFCC)*

MFCC, introduced by Davis and Mermelstein [23] is the most widely used feature based on filter bank analysis. It exploits auditory principles along with the decorrelating property of cepstrum. The procedure of MFCC computation is shown in Fig. 1.

#### *3.1.2 Shifted delta cepstral (SDC) coefficients*

Dialect and language identification research typically uses delta and double-delta coefficients with the cepstra obtained at each frame time t as spectral features. Significant improvements have been achieved in research [24, 25] using SDC features. This improved feature set is an extension of delta cepstral coefficients [26]. In contrast to the conventional delta coefficient, SDC captures long-term temporal features. They reflect dynamic characteristics of spectral features and can possess pseudo prosodic behavior. SDC features are obtained from multiple frames. Figure 2 represents SDC coefficients calculation method. These coefficients use four parameters represented as N-d-P-k. N represents the number of MFCC based cepstral coefficients from each frame of data. The parameter d represents the time shift of consecutive blocks over which deltas are calculated; whereas, the parameter P determines the gaps between consecutive delta computations and k is the number of blocks whose delta coefficients are stacked to form the SDC feature vector. For a given time t, we first obtain:

$$\Delta C(t,i) = C(t+iP+d) - C(t+iP-d) \qquad (1)$$

i varies from 0 to k−1; and finally, the SDC feature vector is obtained as k stacked version of SDC coefficients; represented by:

$$SDC(t) = \left[\Delta C(t,0)^t \Delta C(t,1)^t \ldots\ldots \Delta C(t,k-1)^t\right]^t \qquad (2)$$

Recent research [24, 27] have shown the effectiveness of this feature by achieving improved recognition score without any increase in computational cost due to dimensionality. This motivated the authors to use them for Hindi dialect identification task.

Previous study shows that intonation plays an important role in recognition of spoken accent; that is due to speaker's native dialect. Research in [6, 8, 28] shows that prosodic features such as formant frequency, pitch, phone duration, energy and intensity, all contribute to accents to some degree. Prosodic features are supra-segmental features. These features can be properly extracted from units greater than phones. To take the benefits of these features in this work syllables were used as processing unit. Also, syllable is one such acoustic unit that has a close connection with human speech perception and articulation [29]. In order to further process the speech samples using syllables contained in it speech signal must be segmented at the syllable level and aligned with phonetic transcriptions. The segmentation can be done manually to achieve high accuracy. But for large data set manual segmentation and labeling is extremely time-consuming task. Automatic segmentation of speech has become standard practice for

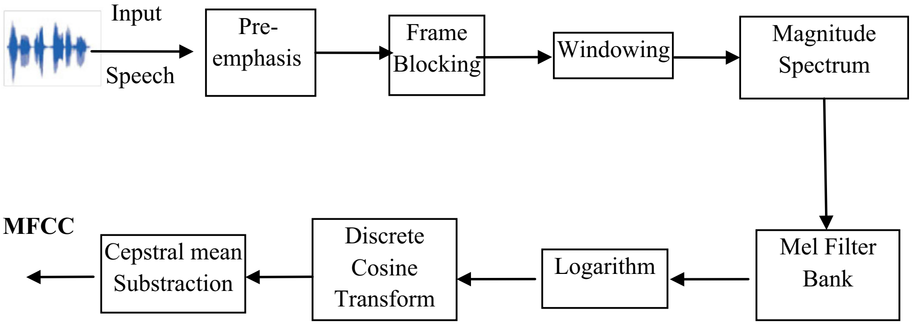

**Fig. 1** Block diagram of Mel-frequency Cepstral coefficients

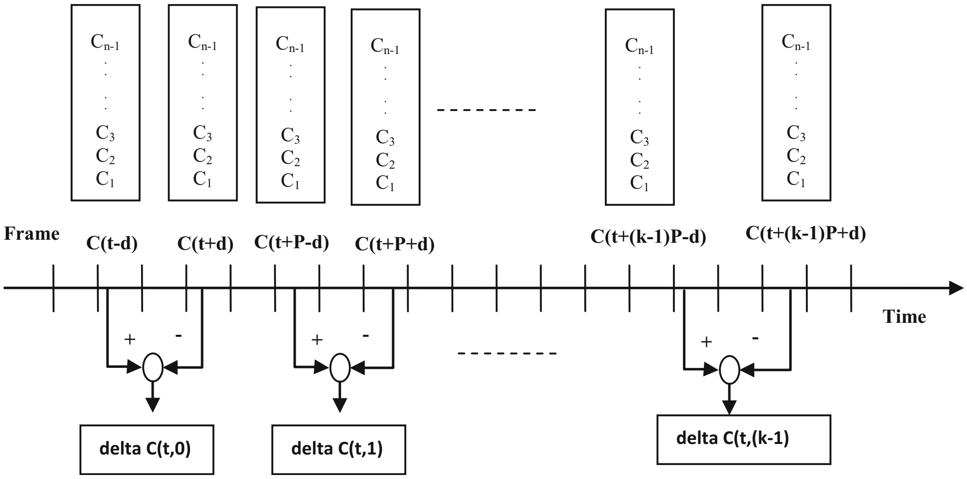

**Fig. 2** Shifted delta Cepstral coefficients

these researches. In this work, syllables were segmented using Donlabel tool [30].

Formants as prosodic features are used by many researchers for accent classification. As formants correspond to resonant frequencies with maximum amplitude, they give information regarding voiced sounds, of which most notable are vowels [31]. Researchers in [6, 22] have used formant frequencies for classification of accents. Their work uses phoneme-based model for classification. Intensity refers to the sound power per unit area. Listeners perceive it as loudness of the sound. Individual's vocal loudness; emotional state, environment and distance from the recording devices are few factors that contribute to this characteristic of speech. Due to varying influential characteristics it can not be used as a concrete measure for dialect identification. However, in [32], intensity has been considered for the classification of two British dialects. Since, it is based on the speaker who can speak both the dialects, i.e. the subjects are common along with same environmental conditions, and the use of intensity as a prosodic feature is justifiable. This paper uses pitch, energy and duration as prosodic features for classification of Hindi dialects.

### *3.1.3 Pitch as prosodic cue*

Fundamental frequency F0, dominates the overall performance of voice pitch and is perceived as pitch range. It is influenced by some common factor such as anatomy, speaker's language background and the emotional state of speakers. The temporal pitch dynamics conveys intonation related information. Tests based on human perception shows that pitch movements can be used to distinguish one language or dialect from another [33]. Each language or dialect has its unique pattern of rise and fall of tone and stress. In [34] it is shown that variations of pitch contour are dialect dependent and can be used as a cue for dialect identification. Fundamental frequency as a prosodic feature for dialect/accent recognition tasks has been mainly used for tonal languages. Most of the work done for Mandarin uses pitch as a representation of tone and have shown considerable improvements in system performance [35, 36]. Grover et al. [37] shows that speakers of French, German and English differ in their pitch contour. Work done for Hindi dialects in [38] highlights the presence of lexical tone due to native dialects, and this motivated the authors to consider it as a prosodic cue in this work.

Extraction of F0 from the speech segment requires two-step processes. In the first step, speech frames are classified as voiced and unvoiced and then fundamental frequency is computed from the voiced frames, setting the unvoiced frame value as zero. Several algorithms are available for F0 extraction. "Yet another Algorithm for Pitch Tracking" (YAAPT) [39], which works in both time and frequency domain is adapted for this work. Average pitch for male and female speakers in the database is found to be 131 and 235 Hz respectively.

### *3.1.4 Energy as prosodic cue*

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. Most of the work done for accent/dialect recognition uses frame energy as a prosodic feature in their work. Some researchers consider it as a separate stream [15], and some consider it along with the spectral feature set [18]. In both the cases performance of the system is improved. Energy of each overlapping frames of segmented speech is obtained by summing the squared amplitude of each sample.

### *3.1.5 Duration*

Biadsy [40] in his research outlines that due to speaking style of individual's length of spoken segment varies and is mainly concerned with the vowel duration. Duration of vowels has been used for British regional dialect classification in [32] 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 each dialect is different, phonetic duration differences occur among dialects. Figure 3 shows mean vowel duration for 10 Hindi vowels in four dialects under consideration. These differences are obtained by considering CV and CVC syllables of our database. The significant differences between the duration of vowels due to dialects motivate the authors to exploit it for dialect identification task.

## 4 Modeling of multi-stream feature using AANN

All acoustic features have some strength and weaknesses. Feature combination aims to fuse features obtained from multiple streams of speech signal with the aim to capture discriminative information. The fusion of multi-stream features has been widely used for speech, speaker and emotion recognition tasks. Evidence of their better performance is available in literature.

The simplest approach to combine two feature set is to concatenate them into single feature vector [41]. But, this method simply increases the dimensionality. In [42], articulatory features are combined with standard acoustic features using direct method based on linear discriminate analysis (LDA). In [43], Heteroscedastic linear discriminant analysis (HLDA) is used to combine multiple streams of acoustic features. Both LDA and HLDA assumes that feature in each class obey Gaussian distribution. Another limitation of LDA is that it stores same covariance structure across all classes, where as HLDA deals with this limitation by storing separate covariance across all classes, there by increasing the storage requirement. Principal component analysis (PCA) is also widely used independent feature extraction/reduction method. As with LDA, PCA also assumes that the structure of data is inherently linear. However, since speech features are not necessarily linear non-linear methods can be assumed to give better representation in fewer dimensions. In NLPCA [44], neural network is trained as an identity map. This type of neural network is referred as Auto-associative neural net (AANN), which is trained to minimize mean square error using target same as the input. AANNs have been shown to successfully capture the variability in the input speech

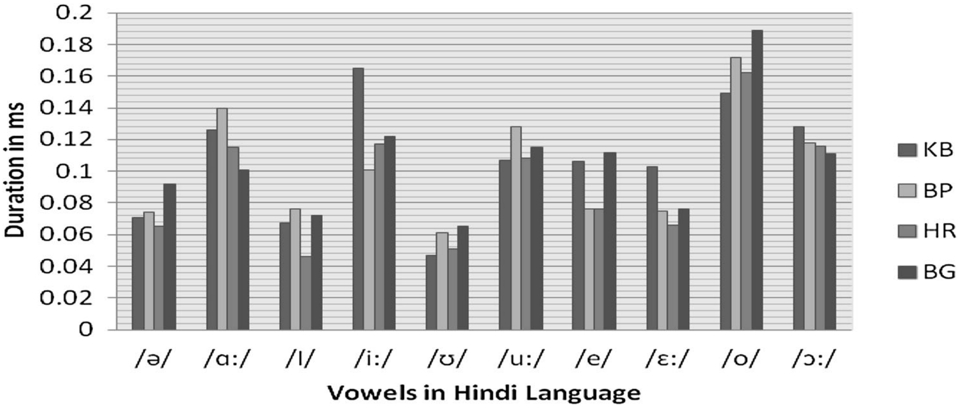

**Fig. 3** Comparative chart of average duration of Hindi vowels in four dialects

feature while representing them in small numbers [19, 45]. Feature streams used in this work are:

(i) 12 MFCC features + log energy.
(ii) 24 Delta features corresponding to obtained from SDC.
(iii) 4 Pitch values ($f_{min}$, $f_{max}$, $f_{mean}$, $\Delta f0(f_{max}-f_{min})$.
(iv) Duration of syllables.

Feature streams from (i), (ii) and (iii) are obtained at frame level and the duration feature is obtained at the syllable (multiple frames) level. These individual feature streams are combined to generate 3 streams as;

Stream A: combination of (i) and (ii), Stream B: combination of (i) and (iii) and stream C as combination of (i),(ii) and (iii). Duration is obtained at the syllable level so to study its impact as prosodic feature it is combined in the fusion logic at the final decision level for identification.

Recent research in ADI uses GMM [18, 40] or HMM [6, 10, 33] for modeling of feature set, which is then used by the classifier to identify the correct class of input utterance. Although the GMM based systems have shown good performance for automatic identification of accent but their performance is highly dependent on the number of Gaussian mixtures and the initial partition. Also, GMMs are constrained by the fact that the shape of the distribution of component is assumed to be Gaussian [19] and number of mixture needs to be fixed in advance. Speech data have complex structure and hence cannot be adequately represented using GMM that uses first and second order statistics. HMMs on the other hand support acoustic and temporal modeling but again, it makes number of suboptimal modeling assumptions. It is based on the assumption that successive acoustic vectors follow Gaussian distribution and are uncorrelated. They give best results in case of context dependent models. But these models have large number of parameters and handling them requires complex processing. ANNs have often been favorable choice for researchers. They are universal approximators and can model any continuous function with a simple structure. Classification, reduction, clustering all can be done using ANNs.

In [19]; AANNs have been proposed as the alternative to GMM. They relax the assumption of feature vectors to be normal locally and capture higher order moment. In few cases it has been observed that GMM slightly performs better than AANN, but the number of parameters used by AANNs is much—much fewer than that required by GMM. The efficiency of AANN is utilized in this work.

AANN is a feed forward neural network (FFNN) with one input, one output layer and one or more hidden layers. The number of units at each hidden layer and the number of hidden layer is dependent upon the problem in hand. It has been shown that a three layer AANN clusters the input data in linear subspace. When the constraints on number of hidden layer are relaxed, the network is able to cluster the input data in the non-linear subspace [46].

For each of the four dialects 3 AANNs corresponding to three output streams are created. The number of neurons at the input and output layer is dependent upon the number of features from each input stream. Figure 4 represents schematic diagram for multi stream fusion.

The first and the last hidden layers capture local information contained among the feature vectors. The number of nodes in these layers are derived experimentally. Second hidden layer is known as bottleneck of the network and compresses the input vector producing reduced dimensional feature. This compression layer is also responsible for capturing global information from the input feature.

The network structure used for the generation of stream A is 37L-83N-19N-83N-37L, for stream B the structure used is 17L-44N-9N-44N-17L and for stream C it is 41L-80N-22N-80N-41L. Here, L represents linear units and N represents non-linear units. Output function at nonlinear unit is *tanh*($x$), where x is the activation parameter. Network learns by minimizing the mean squared error for each frame using back propagation learning algorithm to adjust network weights. Variation of parameter such as, epoch is not very critical, as it does not influence the performance of the system [19].

## 5 Classification model for Hindi dialects

Support vector machines are the most favorable choice of classifier for most of the recent research in language or dialect/accent recognition. In [47], SVM is used with SDC feature vector for language recognition by machine. Generalized linear discriminant sequence kernels have been designed for this task and the shows that performance of SVM classifiers is comparable to that of GMM based systems. Advantage of SVM is that it can cope with hard classification problems and gives solution with maximum margin, but the optimality of solution depends upon the kernel used. Also, special effort is required to deal with variable length speech segments. ANNs have always been

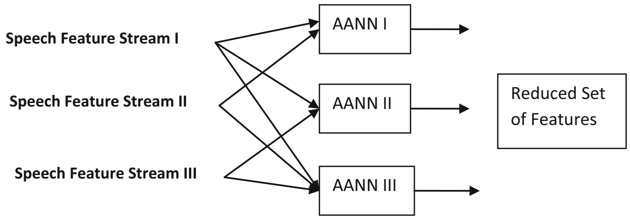

**Fig. 4** Schematic representation for fusion of input features at front-end

a preferable choice for pattern classification by researchers. AANNs have been used as classifier in many of the pattern classification task. In [45], Mixture of AANNs has been used for speaker verification tasks. In [48], AANNs are trained with residual phase features and MFCCs for speaker recognition. The data reduction and classification capability motivated us to use AANNs as dialect classifiers for Hindi.

AANN model for each dialect is created separately. Each classifier consists of 4 AANN model representing one dialect each. Again for each dialect the model consists of 4 AANNs; 3 for the reduced feature stream A,B and C and fourth for the duration of syllable. The structure of AANN for each dialect is same. Figure 5 represents AANN based dialect classifier. The fourth classifier of each dialect is represented as second level classifier. Classification performance of the system is evaluated by fusing the confidence score from each of the combined feature stream for the dialects with the confidence score obtained for the AANN model used for modeling syllable duration of that dialect.

The classifier for all the dialects is trained using dataset from that particular dialect. The model captures the distribution from the feature vector given as input to the AANN model. The difference between the observed output and the input is used to compute squared error ($\mathbf{e_{jk}}$) for each feature stream k in jth frame and is obtained $\mathbf{e_{jk} = \| y_{jk} - o_{jk} \|^2}$, where $y_{jk}$ is the kth feature vector value given as input from the jth frame and $o_{jk}$ is the observed output from the model for kth feature vector of jth frame. Mean frame error is computed as, $E_j = \frac{1}{T}\sum_{k=1}^{T} e_{jk}$ where T is the total number of feature from each frame. The error $\mathbf{E_j}$ is converted to obtain confidence score from each frame using, $\mathbf{C_j} = \boldsymbol{exp}(\mathbf{-E_j})$. The total confidence value for the test utterance is computed as, $C = \frac{1}{N}\sum_{j=1}^{N} c_j$ where N is the total number of frames. This is obtained from all AANN models, representing one dialect each. Total confidence score is obtained as the weighted sum of the confidence score from fused feature stream and duration ($C_d$) based classifier obtained as $C_t = w \times C + (1-w) \times C_d$, w represents the weight of the stream and vary from 0 to 1. For each syllable of the utterance the confidence score from dialect based classifier is combined and based on predefined threshold value final classification decision is made (Fig. 6).

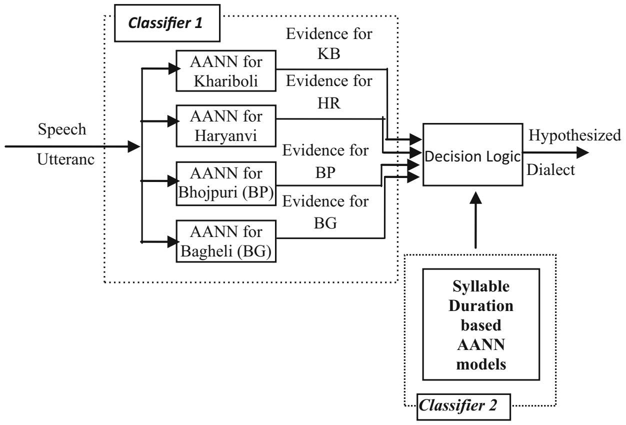

**Fig. 5** Block diagram of dialect identification system based on evidences from each dialect

## 6 Evaluation of dialect identification model

System performance is evaluated using reduced feature set obtained during fusion. Evaluation of system is done for feature level fusion performed by AANN and also for score level fusion. The score level fusion is obtained by varying weights from 0 to 1 to combine the confidence score of fused feature classifier with duration based classifier.

The MFCC features are obtained by dividing speech segment into successive overlapping frames of 20 ms with an overlap rate of 10 ms. 12 MFCC features with along with energy coefficient is computed from each frame. For this 13 coefficient shifted delta coefficients are obtained. The parameter for SDC in this work is set as (12-1-2-2). 24 coefficient constitute one SDC. The classifier structure for training with Stream A is 19L-36N-11N-36N-19L, for stream B the classifier structure used is 9L-21N-4N-21N-9L and C stream based classifier is designed as 22L-50N-14N-50N-22L. These values have been obtained by testing the system for different number of nodes at hidden layer and the one giving best performance is finally used. Minimum number of syllable in any sentence of the text corpus used here is 10 and the maximum is 28. 28 input and output layer neurons are used in each AANN for duration based classification. The structure of AANN model for duration in each dialect is 28L-48N-13N-48N-28L. For sentences that have lesser than 28 syllables the tailed portion of the input is appended with zeros to make it 28 in number (Table 1).

Average performance of the classifier for stream A is obtained as 66 %, with stream B the classifier average performance is noted as 72 % and with stream C the classifier performance is observed as 86 %. From the findings it is clear that prosodic feature such as Pitch provides significant improvement to the dialect classifier. Also, the delta features from SDC is able to capture temporal dynamics of speech signal. Inclusion of delta features captured over sequence of frame is significant for system performance.

For the score level fusion, the classifier output obtained with the combined feature set is added as weighted sum to the output of syllable duration based classifier. The system

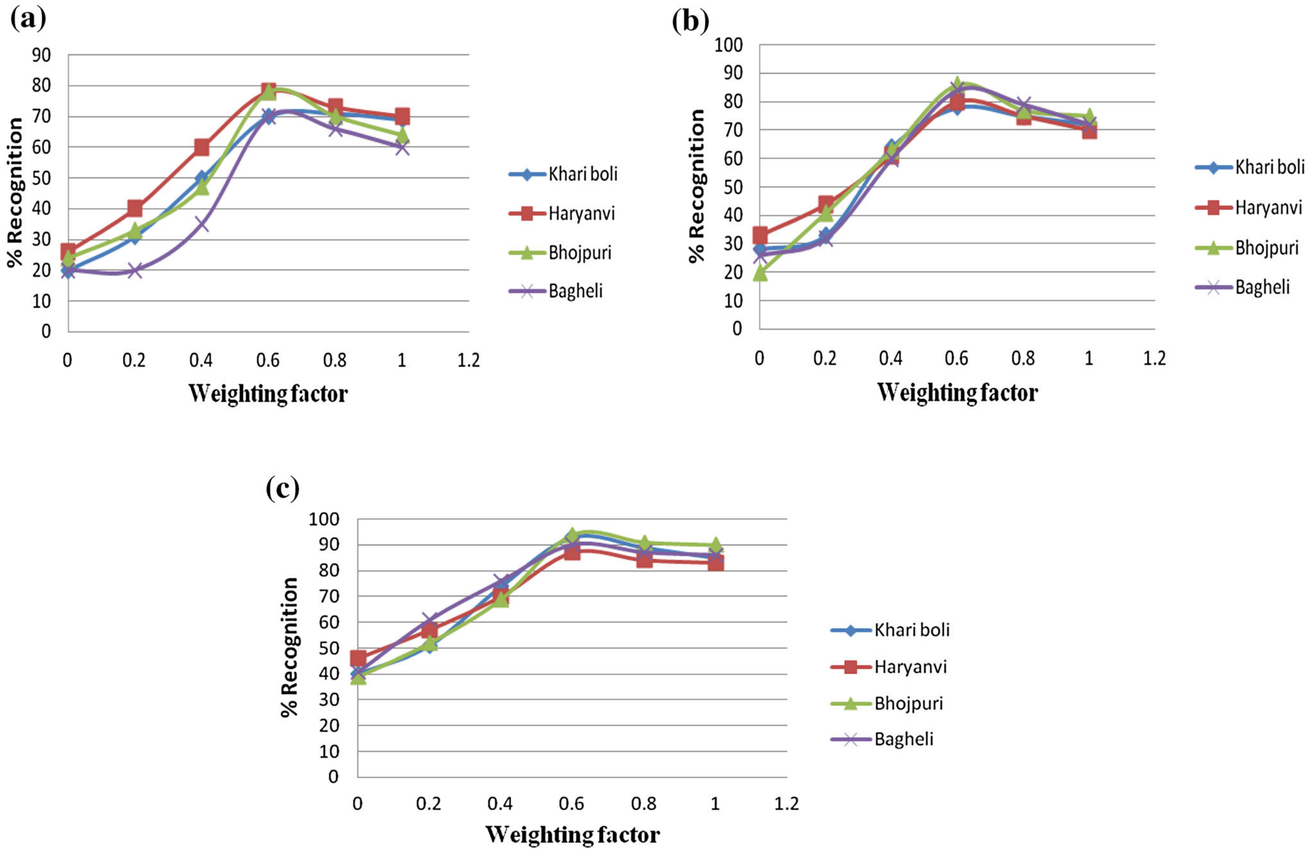

**Fig. 6** System performance with score level fusion of duration based classifier with **a** Stream A, **b** Stream B and **c** Stream C

**Table 1** System performance for feature level fusion

| Hindi dialects | Recognition performance (%) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Stream A | | | | Stream B | | | | Stream C | | | |
| | KB | HR | BP | BG | KB | HR | BP | BG | KB | HR | BP | BG |
| KB | 69 | 15 | 8 | 8 | 72 | 11 | 7 | 10 | 85 | 9 | 1 | 5 |
| HR | 11 | 70 | 10 | 9 | 15 | 70 | 10 | 5 | 6 | 83 | 6 | 5 |
| BP | 6 | 7 | 64 | 23 | 0 | 12 | 75 | 13 | 2 | 4 | 90 | 4 |
| BG | 14 | 8 | 18 | 60 | 13 | 7 | 8 | 72 | 7 | 4 | 3 | 86 |

performance is evaluated by varying the weights of the two confidence score. It is observed that for all the feature stream the score level fusion performes well. Undoubtedly, the results obtained highlights the significance of duration in Hindi dialect classification, but importance of spectral feature is also highlighted for different degrees of weights assigned to feature streams. The average performance of system with stream A is 74 %, with stream B it is 84 % and with stream C it is 91 %. For both the scheme it is obtained that the system performs best for the Bhojpuri dialect. One of the reason behind this may be the geographical proximity of speakers selected for this work.

## 7 Summary and conclusion

In this paper, fusion of speech features are explored for Hindi dialect identification. We have outlined the capability of auto-associative neural network for its use for feature combination and also for classification of speech features by capturing dialect specific information from underlying distribution of feature vectors. This system is based on Four Hindi dialects and can be extended to more dialects of the language for identification. Different feature streams are fused at the front end and the model is trained with the fused feature set. Fusion is also performed at the score level. As the feature stream combined at the feature level are all obtained from frame based processing, the duration feature, calculated at syllable level is fused at score level. It is obtained that the fusion of feature gives better results and duration is one of the important prosodic feature for Hindi dialect classification. The results obtained in this work are promising and demonstrates the potential of AANN as a candidate for dialect classification using speech. In future we would like to extend this work on more Hindi dialect.

## References

1. Zissman MA, Gleason TP, Rekart DM, Losiewicz BL (1996) Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech. In: Proceedings on IEEE Conference acoustics, speech, signal processing, vol. 2, pp 777–780
2. Kumpf K, King RW (1996) Automatic accent classification of foreign accented Australian English speech. Proc Fourth IEEE Int Conf Spok Lang 3:1740–1743
3. Torres-Carrasquillo PA, Sturim D, Reynolds D, McCree A (2008) Eigen-channel compensation and discriminatively trained gaussian mixture models for dialect and accent recognition. In Interspeech, Brisbane
4. Huang R, Hansen JHL, Angkititrakul P (2007) Dialect/accent classification using unrestricted audio. IEEE Trans Audio Speech Lang Process 15(2):453–464
5. Wells JC (1982) Accent of English, vol 2. Cambridge University Press, London
6. Arslan LM, Hansen JHL (1996) Language accent classification in american English. Speech Commun 18:353–367
7. Rouas J-L, Farinas J, Pellegrino F, Andre-Obrecht R (2003) Modeling prosody for language identification on read and spontaneous speech. In: Proceedings on IEEE international conference acoustical, speech, signal process, Hong Kong, vol 1, pp 40–43
8. Yan Q, Vaseghi S, Rentzos D, Ho C-H, Turajlic E (2003) Analysis of acoustic correlates of British, Australian and American accents. In IEEE workshop on automatic speech recognition and understanding, pp 345–350
9. Ma B, Zhu D, Tong R (2006) Chinese dialect identification using tone features based on pitch flux. In: Proceedings of ICASP'06, pp 901–904
10. Alorfi FS (2008) Automatic identification of Arabic dialects using hidden markov models. PhD Dissertation, University of Pittsburgh
11. Peters J, Gilles P, Auer P, Selting M (2002) Identification of regional varieties by intonational cues. An experimental study on Hamburg and Berlin German. Lang Speech 45(2):115–139
12. Barkat M, Ohala J, Pellegrino F (1999). Prosody as a distinctive feature for the discrimination of Arabic dialects. In: Proceedings of Eurospeech'99, p 1
13. Hamdi R, Barkat-Defradas M, Ferragne E, Pellegrino F (2004) Speech timing and rhythmic structure in Arabic dialects: a comparison of two approaches. In: Proceedings of interspeech'04
14. Blackburn CS, Vonwiller JP, King RW (1993) Automatic accent classification using artificial neural networks. In: Proceedings of Eurospeech '93, Vol 2, pp 1241–1244
15. Rao KS, Koolagudi SG (2011) Identification of Hindi dialects and emotions using spectral and prosodic features of speech. IJSCI 9(4):24–33
16. Biadsy F, Hirschberg J, Ellis DPW (2011) Dialect and accent recognition using phonetic-segmentation supervectors. In: Interspeech, Florence
17. Hou J, Liu Y, Zheng TF, Olsen J, Tian J (2010) Multi-layered features with SVM for Chinese accent identification. In: IEEE international conference on audio language and image processing (ICALIP), pp 25–30
18. Lazaridis A, Khoury E, Goldman JP, Avanzi M, Marcel S, Garner PN (2014) Swiss French regional accent identification. In: Proceedings of Odyssey, Joensuu
19. Yegnanarayana B, Kishore SP (2002) AANN: an alternative to GMM for pattern recognition. IEEE Trans Neural Netw 15:459–469
20. Sinha S, Jain A, Agrawal SS (2014) Speech processing for Hindi dialect recognition. Adv Signal Process Intell Recognit Syst 264:161–169
21. Rubio Ayuso AJ, Lopez Soler JM (1995) Speech recognition and coding new advances and trends. Springer, New York
22. Ghorshi S, Vaseghi S, Yan Q (2008) Cross-entropic comparison of formants of British, Australian and American English accents. Speech Commun 50:564–579
23. Davis S, Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process 28(4):357–366
24. Rong T (2006) Automatic speaker and language identification. A First Year Report Submitted to the School of Computer Engineering of the Nanyang Technological University, Nanyang
25. Behravan H (2012) Dialect and accent recognition. Dissertation, University of Eastern Finland
26. Torres-Carrasquillo PA, Singer E, Kohler MA, Greene RJ, Ryenolds DA, Deller JR, Jr. (2002) Approaches to language identification using Gaussian mixture models and shifted delta Cepstral features. In: Proceedings on ICSLP, Denver, pp 89–92
27. Calvo J, Fernndez R, Hernndez G (2007) Channel/handset mismatch evaluation in a biometric speaker verification using shifted delta Cepstral features. In: Proceedings of CIARP 2007, LNCS 4756, pp 96–105
28. Esther G, Brechtje P, Francis N, Kimberley F (2000) Pitch accent realization in four varieties of British English. J Phon 28:161–185
29. Ganapathiraju A et al (2001) Syllable-based large vocabulary continuous speech recognition. IEEE Trans Speech Audio Process 9(4):358–366
30. Deivapalan PG, Jha M, Guttikonda R, Murthy HA (2008) Donlabel: an automatic labeling tool for Indian languages. In: National conference on communications, Bombay
31. Ladefoged P (1996) Elements of acoustic phonetics, 2nd edn. The University of Chicago Press, Chicago
32. Zheng DC et al (2012) A new approach to acoustic analysis of two British regional accents—Birmingham and Liverpool accents. Int J Speech Technol 15(2):77–85
33. Kumpf K, King RW (1997) Foreign speaker accent classification using phoneme-dependent accent discrimination models and comparisons with human perception benchmarks. In: Proceedings on Eurospeech, pp 2323–2326
34. Mehrabani M, Boril H, Hansen JH (2010) Dialect distance assessment method based on comparison of pitch pattern statistical models. In: Proceedings on IEEE international conference on acoustics speech and signal processing (ICASSP), pp 785–797
35. Chang, E et al. (2000) Large vocabulary Mandarin speech recognition with different approaches in modeling tones. In: Proceedings on Interspeech
36. Chen CJ et al. (1997) New methods in continuous Mandarin speech recognition. In: Proceedings on Eurospeech, Spain
37. Grover C, Jamieson DG, Dobrovolsky MB (1987) Intonation in English, French and German: perception and production. Lang Speech 30:277–296

38. Kulshreshtha M, Mathur R (2012) Dialect accent feature for establishing speaker identity: a case study. In: Neustein A (ed) Springer briefs in electrical and computer engineering. Springer, New York
39. Zahorian SA, Hu H (2008) A spectral/temporal method for robust fundamental frequency tracking. J Acoust Soc Am 123(6):4559–4571
40. Biadsy F (2011) Automatic dialect and accent recognition and its application to speech recognition. Dissertation, Columbia University
41. Gonzalez DR, Calvo de Lara JR (2009) Speaker verification with shifted delta Cepstral features: its pseudo-prosodic behaviour. In: Proceedings on I Iberian SLTech
42. Zolney A, Kocharov D, Schluter R, Ney H (2007) Using multiple acoustic feature sets for speech recognition. Speech Commun 49:514–525
43. Aggarwal RK, Dave M (2013) Performance evaluation of sequentially combined heterogeneous feature streams for Hindi speech recognition system. Telecommun Syst 52(3):1457–1466
44. Kramer MA (1991) Non linear principal component analysis using auto associative neural networks. AIChE J 37:233–243
45. Sivaram GSVS, Thomas S, Hermansky H (2011) Mixture of auto-associative neural networks for speaker verification. In: Proceedings on INTERSPEECH
46. Bianchini M, Frasconi P, Gori M (1995) Learning in multilayered networks used as autoassociators. IEEE Trans Neural Netw 6:512–515
47. Campbell WM et al (2006) Support vector machines for speaker and language recognition. Comput Speech Lang 20(2):210–229
48. Murty KSR, Yegnanarayana B (2006) Combining evidence from residual phase and MFCC features for speaker recognition. IEEE Signal Process Lett 13(1):52–55