ACOUSTIC-PHONETIC FEATURE BASED DIALECT IDENTIFICATION IN HINDI SPEECH

Shweta Sinha[1], Aruna Jain[2] and S. S. Agrawal[3]

[1,2]Department of Computer Science and Engineering, Birla Institute of Technology, Mesra, Ranchi, India
[3] Department of Electronics and Communications, KIIT Group of Institutions, Gurgaon, Haryana, India
Emails:[1]meshweta_7@rediffmail.com,[2]arunajain@bitmesra.ac.in,[3]ss_agrawal@hotmail.com

Submitted: Nov. 5, 2014 ***Accepted: Jan. 12, 2015*** ***Published: Mar. 1, 2015***

Abstract- Every individual has some unique speaking style and this variation influences their speech characteristics. Speakers' native dialect is one of the major factors influencing their speech characteristics that influence the performance of automatic speech recognition system (ASR). In this paper, we describe a method to identify Hindi dialects and examine the contribution of different acoustic-phonetic features for the purpose. Mel frequency cepstral coefficients (MFCC), Perceptual linear prediction coefficients (PLP) and PLP derived from Mel-scale filter bank (MF-PLP) have been extracted as spectral features from the spoken utterances. They are further used to measure the capability of Auto-associative neural networks (AANN) for capturing non-linear relation specific to information from spectral features. Prosodic features are for capturing long - range features. Based on these features efficiency of AANN is measured to model intrinsic characteristics of speech features due to dialects.

Index terms: Dialect Identification, Auto-associative neural network, Feature compression, Hindi dialects, Spectral and Prosodic features.

I. INTRODUCTION

Speech is the most common and natural means of communication among human. Advancements in technology and need for access to massive online resources have made human-machine communication essential in everyday life. Considering speech as one of the medium for communication between man and machine will always be a welcome step by the society. It will help extend the use of information technology to the population who are not well acquainted with the peripheral devices of computers. For an automatic speech recognition system of any language, with varieties of dialects the performance is highly dependent upon the variability captured during training of the system. Next to gender dialect/accent of speakers is the most influencing factor for any ASR [1]. This has motivated the researchers in the area of speech to concentrate on acoustic variation naturally present in speech. Accent is a pattern of pronunciation that can identify speaker's linguistic, cultural or social background. Apart from socio-economic background, speaker's native tongue influenced by the dialect spoken by them and in their surrounding is the major factor influencing accent of speakers. Dialect of a given language is a pattern of pronunciation or vocabulary of words used by the community of native speakers belonging to the same geographical area. The aim of dialect identification system is to identify the dialect of the speaker from the spoken utterances based on their speech characteristics. Once the dialect is identified system performance can be improved by adapting to the appropriate language and acoustic model [1].

Human speech consists of a wide range of information regarding speech features that may be guided by speaker's speaking style, their speed of speech production, age and emotional state. These characteristics are somehow controlled by speaker's spoken dialect of the language. Phonotactic, spectral and prosodic features contained within the speech sample can give sufficient information regarding the native tongue of the speaker [2].

Modeling techniques in identifying dialects of a language take advantage of different linguistic hierarchy layers. These approaches include phonotactic and acoustic models. Phonotactic models are based on phone sequence distribution, where vowel inventory, tense marking, diphthong formation, etc. are the base for the study [3,4]. Dialect recognition is analogous to language identification (LID) task. Most of the work done is motivated by LID systems. In [5] LID system is applied for recognition of 14 regional accents of British English. This system scores 96.5% recognition rate with the features used in LID task. Most of the task based on acoustic models for dialect identification uses spectral features with

Gaussian mixture models [6,7]. In [2] it has been emphasized that accent variations stretch out in both phonetic as well as prosodic characteristics of speakers. Evidence of better performance of the system based on a combination of phonetic and prosodic features is provided in [8]. In [9] MFCC, energy, and prosodic features have been used to classify regional Hindi dialects. Pitch and formant contours are promising candidate as prosodic features. They have been used with stochastic trajectory nodes to distinguish between Americans, Chinese and Turkish accents [10].

Hindi is a language spoken by huge population of the world. It has around 50 prominent dialects. Number of speakers in these dialects varies from thousands in one to millions in other. Influence of native tongue on speaker's speaking style is prominent even when standard Hindi is spoken by them. This influence makes a huge impact on the ASR performance. Identification of dialect becomes evident for better performance of ASR. Available literature highlights that energy; formants and information related to the fundamental frequency are found to be the most discriminative features for identifying possible accents. Considering these in mind, few dialect based studies have been initiated using suprasegmental features of speech samples. Supervised learning approach has been widely used in recognition task with small or medium size databases [11]. This approach has been used for dialect classification problems recently. Spectral and prosodic features have been used for classification of accents on a small database collected from people who are non-native speakers of Hindi [12]. Impact of Hindi due to their mother tongue (regional language) is studied in this paper. Research in [13] discusses dialect classification of isolated utterances using multilayer feedforward neural network as a classifier. Findings of this work highlight that prosodic features carry substantial information about the spoken dialect. Duration of syllable can be used to model spoken rhythm. Keeping this in view, the impact of several spectral and prosodic features on the performance for identification of Hindi dialects in continuous speech is explored in this work. Work in the area of Hindi speech recognition highlights the importance of MFCC and PLP as significant spectral features [14]. Integration of these features has been explored in [15], and possible improvement has been obtained by the integration. Due to their significance in speech recognition, performance of dialect recognition system has been explored for these spectral features and their combination. This work further explores their combination with prosodic features. Speech samples used here are read speech, recorded in standard Hindi, by speakers of different Hindi dialects.

Digital processing of speech requires extracting features that characterize the spoken utterances that can be used further processing. Feature extraction is the process that identifies

the salient properties of data to facilitate its use. It is often observed that superficial dimensionality of data is much greater than the intrinsic dimensionality of the extracted feature set. In speech recognition system, large numbers of spectral features are computed from several speech frames. Due to the high dimensionality, several millions of parameters are determined from the training data, and this increases the demand for storage while decreasing the speed for processing. To overcome these shortcomings, it is required to reduce the dimensionality of data while preserving discrimination between different phonetic sounds.

The purpose of dimensionality reduction is feature transformation for improved speech recognition along with a reduction in data size. Several algorithms exist for this purpose. Most of them are linear in nature, so can discover only those data that have linear or near linear structure in high dimensional input space. Research in the area of speech have shown that speech sounds lie on low-dimensional curved subspaces embedded in high dimensional acoustic feature space [16]. Due to this non-linearity linear methods can not discover the exact embedded structure of the data. Comparative analysis of linear and non-linear dimensionality reduction methods applied to speech processing shows that non-linear methods outperform linear methods [17]. Principal component analysis (PCA) [18] and linear discriminant analysis (LDA) [19] are few of linear dimensionality reduction methods while non-linear PCA (NLPCA) is used as non-linear methods. Neural network has been widely used with high success rate for not only speech based task but also for several other artificial intelligence tasks [20]. This non-linearity for data reduction can be achieved by auto-associative neural network (AANN) [21]. The non-linear features are generated using artificial neural network(ANN) training procedure and the mapping between original and reduced data set is done by AANN.AANN has served as good classifier in many of language and emotion recognition work[12]. They have been successfully deployed for segmentation and indexing of audio signals [22]. Its capability of being a good classifier along with a good compressor motivated us to use it for dialect identification task.

Hereafter, the paper is arranged as follows: Section 2 presents the mathematical formulation of the problem in hand. Section 3 describes speech corpus creation for this work. Acoustic phonetic feature extraction and its reduction process are explained in section 4. Proposed model for dialect identification is presented in section 5. Section 6 evaluates system performance based on several spectral and prosodic features, and the work is concluded in section7.

II. DIALECT IDENTIFICATION PROBLEM

Let, A={a_1,a_2,a_3….a_n} where a_j , 1<=j<=n represents acoustic features (Spectral and prosodic) corresponding to any of the dialects in the set D = {D_1,D_2….D_m} of m dialects. The aim is to obtain most likely dialect D^* corresponding to the input speech consisting of n acoustic features under consideration. This can be mathematically expressed as:

$$D^* = \arg \max_i P\left(D_i / A \right) \quad (1)$$

Where $P(D_i/A)$ is the posterior probability of the dialect D_i. If the input vector belongs to any one of M classes D_i, 1<=i<=m, then the main objective of this classification problem is to decide the class to which the given vector A belongs to. According to Baye's rule; the problem turns out as one of the joint probability maximization problem and can be given as:

$$P(A,D_i)=P(A|D_i)\,P(D_i) \quad (2)$$

From Eq (1), the objective is to choose the class D^* for which the posterior probability $P(D_i|A)$ is maximum for given A and can be implemented as:

$$D^* = \arg \max_i P\left(A / D_i \right) P(D_i) \quad (3)$$

Where, $P(A|D_i)$ represents the likelihood probability of A corresponding to D_i and $P(D_i)$ denotes a piori probability of dialect that is assumed to be uniform for all dialect and so this can be ignored. Eq (3) simplifies to:

$$D^* = \arg \max_i P\left(A / D_i \right) \quad (4)$$

Thus, the dialect identification task becomes estimation of the posterior probability based on (1) and likelihood estimation based on (4).

III. HINDI DIALECT SPEECH CORPUS

No standard speech database for Hindi dialect exists for speech processing research. Hindi language is spoken by majority of the population in India. Most of the speaker's of this language are from North and Central India. Due to varied geographical and lingual background of speakers huge dialectal diversity exists among Hindi-speaking regions. From around 50 dialects of Hindi four prominent dialects (spoken by considerably large population); Khariboli (KB)(spoken in Delhi and boundary area of neighbouring states), Haryanvi (HR) (Haryana and border area of Delhi), Bhojpuri (BP) (parts of Uttar Pradesh,

Bihar, and Jharkhand) and Bagheli (BG) (Madhya Pradesh and bordering area of Chhattisgarh) have been identified for this research. For initial experimental purpose, a small database is constituted based on all phonemes of the language (11 vowels and 36 consonants). The sentences are based on words from travel domain. The subjects selected for recording belongs to any of the four dialects under consideration. The recording is done as read continuous speech.

The text prompt consists of 300 continuous sentences. These sentences are based on Khariboli dialect and are written in Devanagri script. The length of sentences is not uniform. Minimum number of words in a sentence is six, and the maximum is fourteen. Maximum number of syllable in any sentence is 28. It has been observed during recording that variations exist within the dialect also. Due to this, speakers were selected from close geographical locations. Each speaker was asked to speak ten sentences themselves in their native dialect to capture the effect of dialect on the spoken utterances, before the start of actual recording of samples. This helped to acclimatize the speaker's who were not involved in deep and frequent conversation in their dialect.

Table 1: Statistical description of the Hindi dialect corpus

Database Description	
Tool used for Recording	GoldWave
Sampling Frequency	16kHz
Number of Sentences	300 sentences
Number of Speakers	20 Male, 10 Female
Age of Speakers	Between 21 years to 50 years
Number of Dialects	4 Hindi Dialects
Utterances/dialect	9000 utterances

IV. FEATURE EXTRACTION AND DIMENSIONALITY REDUCTION FOR DIALECT IDENTIFICATION

Abundance of information is embedded in any speech signal. All these acoustic and linguistic information stored within makes the signal unique. A suitable parametric representation is required for extracting statistically relevant information and in turn, reducing them in number. These parametric representations, which are further used for digital processing is termed as

speech features. Several research studies have investigated accent/dialect sensitive acoustic speech characteristics. Many features have been investigated to be used as dialect sensitive traits at both high and low levels of acoustic knowledge.

The acoustic information stored within the signal can be categorized as spectral and prosodic features. Different spectral and prosodic features used for this dialect identification task have been defined in this section.

a. Spectral Features and Prosodic Features:

Mel Frequency Cepstral Coefficients (MFCC): Mel Frequency Cepstral Coefficient (MFCC) is used in state of the art speech processing systems [17-20] and is proven to be one of the most successful spectral feature representatives in speech related tasks. Speech analysis assumes that signal properties change slowly with time [23]. This motivates short time window based processing of the speech signal to extract its parameters. Every 10 ms, a Hamming window is applied to pre-emphasized 20 ms long speech segment. Fast Fourier Transform (FFT) is used to compute short-term spectrum. 20 overlapping Mel scale triangular filters are applied to this short-term spectrum. The output of each filter is the sum of the weighted spectral magnitude. Discrete Cosine Transform is obtained from the logarithm of the filter output to obtain cepstrum coefficients. Figure 1 represents steps in MFCC computation process.

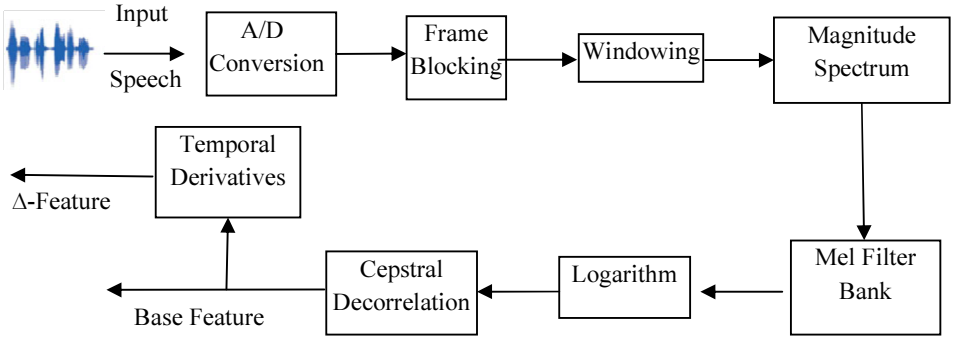

Figure 1. Block diagram of Mel frequency cepstral coefficients

Perceptual linear prediction coefficients (PLP): Motivation behind PLP feature extraction is similar to that of MFCC method. As in MFCC, every 10 ms Hamming window is applied to the 20 ms speech segments. FFT is applied to obtain the short-term spectrum. Further, 20

equally spaced overlapping Bark scale trapezoid filter is applied to the power spectrum. Equal loudness pre-emphasis is then applied to the filter bank output. It is followed by the application of intensity loudness law. To obtain the cepstrum coefficients, Inverse Discrete Fourier Transform (IDFT) is applied to calculate the autocorrelation coefficients. Levinson-Durbin recursion is used to obtain the LP coefficients from these autocorrelation coefficients that are converted to Cepstral coefficients. Figure 2 represents PLP coefficient extraction process.

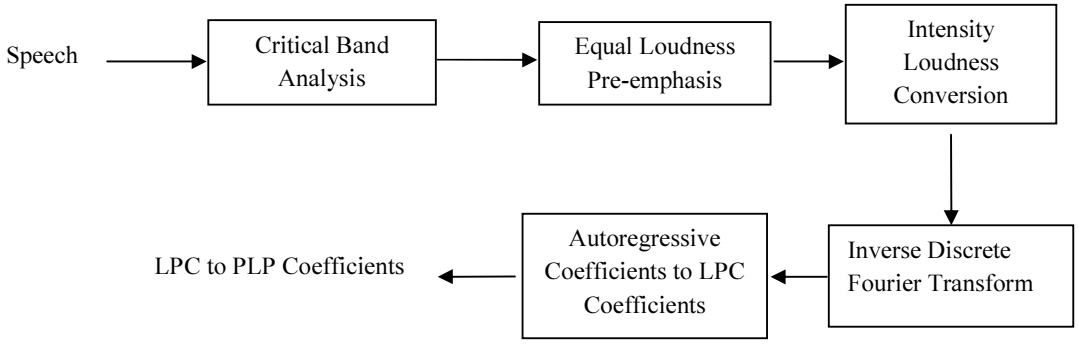

Figure 2. Block diagram of Perceptual linear prediction coefficients (PLP)

PLP derived from Mel-scale filter bank (MF-PLP): Research [24] shows that MFCC and PLP features are complementary in nature. To exploit their combined capability, MF-PLP features are extracted by merging MFCC and PLP techniques into one algorithm. In the first step, output of filters is computed using Mel scale triangular filter bank based on MFCC algorithm. These filters here are applied on the power spectrum in MFCC. For generating cepstrum coefficients, steps of PLP algorithms are followed. The intensity loudness law modifies Filter bank outputs are modified by, and cepstrum coefficients are calculated from this output using all-poles approximation.

Prosody deals with auditory qualities of sound. These features have been proven to be the key feature in human perception of speech. In [9] it is shown that combination of spectral and prosodic features can improve the system performance. Few prosodic features have been investigated here in combination with spectral features for their capability to give dialect specific information. The prosodic features are also referred as supra-segmental features as they are extracted from units bigger than phonemes. These features give information about utterance as well as the speaker. Pitch, energy, duration are well-proven prosodic features.

Fundamental Frequency: Fundamental frequency (F0) represents perceived pitch in human speech and is inherent in any periodic signal. The temporal dynamics of pitch across a signal of speech conveys intonation related information. Cues for perception of rhythmic

characteristics of speech are assumed to lie in regular recurrence of salient speech interval. Different types of speech intervals have been considered to be acoustic correlates of speech rhythm, fundamental frequency related envelop are one of them [25]. Several studies exist to show the importance of prosodic features such as word accent and the phrase intonation in human speech processing, but very few use F0 in combination with other acoustic features [27]. F0 has been mostly found suitable for tonal languages, but the presence of prosodic tone due to accent influenced by native dialect of speakers [26, 29,30] motivated us to investigate it for Hindi dialect recognition.

F0 requires a prior decision to be made regarding voiced/unvoiced status of each frame. Due to its differing nature in voiced and unvoiced speech it is difficult to estimate them correctly[28]. Several algorithm for F0 estimation has been proposed in literature and can be broadly categorized based on their feature's domain i.e. time domain, frequency domain, hybrid time and frequency domain and event detection methods [29-31]. In this work, YAAPT ("Yet Another Algorithm for Pitch Tracking")[30], a noise robust and fast algorithm has been used. This algorithm works in combination of time domain and frequency domain processing and produces F0 value for each frame. The algorithm is adapted for the problem in hand based on characteristics of speech database. Table 2 represents the mean and the standard deviation of the fundamental frequency of male and female speakers of each dialect in the corpus.

Table 2: Mean and standard deviation (STD) of the fundamental frequency for recorded male and female speakers of Hindi dialects

Dialects / Statistical Variations	Khariboli (KB)		Haryanvi (HR)		Bhojpuri (BP)		Bagheli (BG)	
	Male	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**
Mean	122.76	241.75	140.63	221.32	132.63	203.16	164.39	233.37
Standard Deviation	12.22	30.12	11.64	16.09	20.02	26.71	9.59	18.62

Frame Energy: Level of energy helps in identifying the voiced/unvoiced portion of speech. Together with pitch and duration it represents stress pattern of speakers. Energy of each overlapping frames of segmented speech is obtained by summing the squared amplitude of each sample.

Duration: Due to the dialectal influence on one's speaking style length of spoken segment vary, which is mainly concerned with the vowel duration used in the segment. Also, duration represents the rhythm that is guided by speaker's native dialect. Hindi is a syllable-timed language. Syllables are assumed to be better representative than the phonemes in Hindi. Due to its durational stability the variations observed are more systematic at this level. Figure 3 shows that mean vowel duration for 10 Hindi vowels significantly varies in each dialect. This can be used as an important cue to the classifier decision-making. The segmentation of speech into syllables is done using Donlabel tool [32]. This tool uses a combination of group delay function, envelope of the LP residual and energy of prominent spectral peaks to reach a final decision regarding syllable boundary. Naming convention used for syllables is such as to represent its position in the utterance.

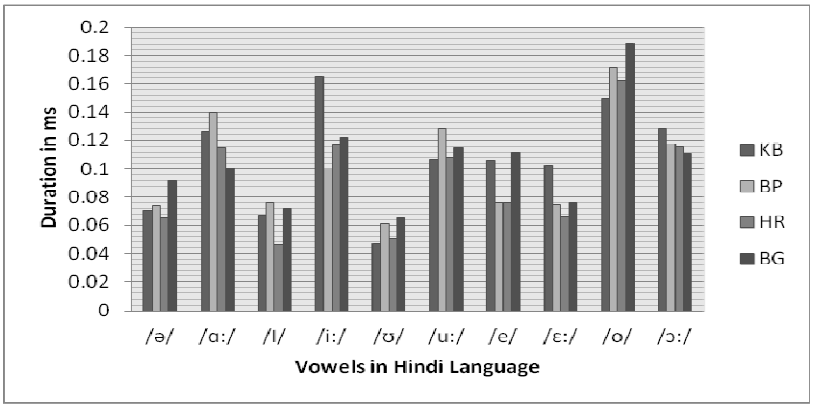

Figure 3. Comparative chart of average duration of Hindi vowels in four dialects

b. Non-Linear Feature Compression

In this work, AANN model is explored for feature compression as well as the classification of dialect by capturing acoustic-phonetic features specific to the dialects. The detail of AANN model based on its capability for the above-mentioned tasks is briefly provided here.

Auto-associative neural network is feed-forward network (FFNN) with identical input and output vectors. This network tries to map the input vector onto itself and hence is named Auto-associative [34,35]. AANN consists of an input layer, an output layer, and one or more hidden layers. The ability of neural network to fit arbitrary non-linear functions depends on the presence of non-linear hidden layer [21]. Available literature shows that 5-layer AANN with 3-hidden layers has been successfully used in many speech based tasks [12,22]. The network structure is represented by XL-YN-ZN-YN-XL, where X, Y and Z refer to the number of processing units at each layer, L represents linear units, and N is for non-linear

units. The number of nodes in input and output layer is equal to the dimension of features used in the problem in hand. In general, number of neurons in first and last hidden layers, also known as mapping and de-mapping layer [33] are greater than that at input and output layer. These layers are responsible for capturing local information contained in the feature vectors. There is no definitive method for deciding in advance the number of nodes in this layer and are derived experimentally. The number of nodes in the middle layer consists of the lesser number of neurons than that at the input/output and other hidden layers [21]. This layer compresses the input vector producing reduced dimensional feature [31,34]. For a large number of training data samples, this reduced feature output can be suitably used as input that will reduce the cost while achieving classification accuracy. This compression layer is responsible for capturing global information from the input feature.

The activation function at the bottleneck layer can be either linear or non-linear, but the activation function at the mapping and de-mapping layer has to be essentially non-linear. This non-linearity provides the capability for modeling arbitrary vector function.

V. MODEL DEVELOPMENT FOR DIALECT IDENTIFICATION

a. Frame Blocking and Feature Compression

To capture the spectral features, energy, and pitch of the spoken utterances the recorded speech signal is divided into overlapping frames of 20ms with an overlap rate of 10ms. Since the FFNN requires fixed size of input vector in all the iterations and speech utterances are not of equal duration; the system is trained using feature set obtained from each frame and the final decision is based on the cumulative sum of output from each frame. From all the obtained frames, the silence frames are removed based on amplitude threshold obtained from the available samples. For spectral features, 13 MFCC coefficients along with 13 velocity and 13 acceleration coefficients, 13 PLP, their corresponding higher order coefficients and 13 MF-PLP along with their high-order delta and delta-delta coefficients are extracted from each frame. All the coefficients are normalized in the range [-1 +1] before feeding as input to the network. The velocity and acceleration coefficients are used to capture spectral trajectories over the spoken duration. Since, the overall classification is based on spectral feature along with prosodic feature from each frame; number of input turns out to be very large and may increase computation time. Some reduction technique is required. AANN's bottleneck aspect

helps in reducing the input feature set and hence the characteristic of AANN is exploited further.

The architecture of AANN model used for compression is 39L-65N-16N-65N-39L. The reduced number of 16-dimensional feature vectors from each frame is used with other prosodic features to train the classifier. Number of nodes at the mapping, de-mapping and compression layer has been experimentally derived. The activation function used in this work is tangent hyperbolic function given as *tanh(k)*; where k is net input value for that unit. The network is trained using conjugate gradient backpropagation learning algorithm for its better speed and convergence property. In order to minimize the mean square error for each feature vector, the weights of the network were adjusted.

b. Classification Model

Figure 4 represents the flow of the input speech processing for classification of dialect. Two separate dialect classifiers one for frame-based feature and other based on sub-word unit have been used in the identification process. Each classifier consists of 4 AANN model representing one dialect each and is trained with spectral and prosodic features of the corresponding dialect. For the prosodic feature F0, ΔF0, frame energy, and syllable duration have been used in this work. First three, as well as the spectral features, are obtained from frame-based analysis of speech sample; whereas, duration computation is done after segmentation of speech into syllables. Different spectral features are combined one at a time with the prosodic features. The first classifier is trained with spectral and prosodic features extracted from the analysis of speech frames. F0 for unvoiced frames are set to zero and ΔF0 from each frame is obtained as the difference between the maximum and minimum F0 value of each frame. Thus, the input size in each cycle remains same. The three prosodic features obtained from each frame are combined with the uncompressed/compressed set of spectral features to form the input to the first classifier. 42(using uncompressed)/19 features from each frame is given as input and output to the dialect-specific AANN. This helps the network to capture the distribution of feature vectors. With uncompressed spectral features structure of AANN in the first classifier, is 42L-65N-24N-65N-42L and with compressed feature set it is 19L-30N-8N-30N-19L for all four dialects. *tanh(k)* is used as non-linear activation function and gradient descent Back-propagation is used as learning algorithm by all AANN models.

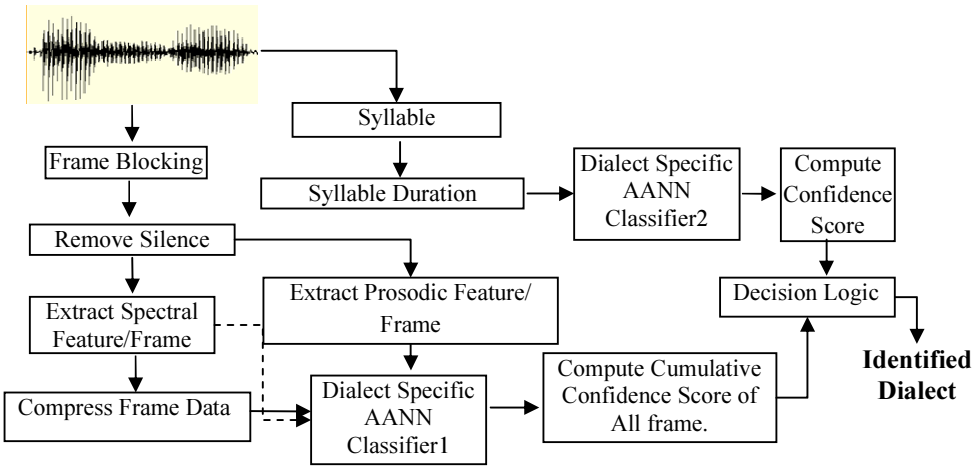

Figure 4. Flow graph of AANN based dialect identification system

Syllable lengthening gives the knowledge about the rhythm in the spoken utterance. Duration of syllables has been used as another prosodic feature and training of 4 AANN models of the second classifier is done by these values. Analysis of the text corpus used in this work shows that the minimum number of syllable in any sentence is 10, and maximum number is 28. Thus, 28 inputs and output layer neurons are used in each AANN of this classifier. The structure of all four AANN is 28L-48N-13N-48N-28L. For sentences with a number of syllables, less than 28 the tailed portion of the input is appended with zeros to make it 28 in number. Since, the differences in syllable duration for different dialects are significant the output of the second classifier is reinforced by the decision logic for the final decision regarding the utterance class.

VI. EVALUATION OF DIALECT IDENTIFICATION MODEL

System performance is evaluated using both uncompressed and compressed feature set. System is trained with 12 male and six female data from each dialect, and the rest are used for testing purpose. Features under consideration are extracted from the test utterances. In the baseline system 39 spectral features obtained as each of MFCC, PLP and MF-PLP can be directly used one by one in combination with 3 prosodic features or their reduced set of 16 dimensional features can be used with 3 prosodic features. These 42/19 features from each frame are given as input to every AANN models of the first classifier. 133 epochs were

required to train the system properly using uncompressed feature set, and 91 epochs were required with the compressed feature set.

To make the classification decision confidence score of the input utterance is obtained from all four AANN models. To do this, firstly, the squared error ($\mathbf{e_{ik}}$) for each feature k in each frame i is obtained as $\mathbf{e_{ik}}=\| \mathbf{Y_{ik}}-\mathbf{O_{ik}}\|^2$, where Y_{ik} is the k^{th} feature vector input value given to the i^{th} frame and O_{ik} is the observed output from the model for k^{th} feature vector of i^{th} frame. Mean frame error is computed as, $E_i = \frac{1}{T}\sum_{k=1}^{T} e_{ik}$, where T is the total number of feature from each frame, for this work it is 42/19. This error $\mathbf{E_i}$ is used to obtain frame confidence score using, $\mathbf{C_i} = exp(-\mathbf{E_i})$. The total confidence value for the test utterance is computed as, $C = \frac{1}{N}\sum_{i=1}^{N} C_i$, where N is the total number of frames. This is obtained from all AANN models, representing one dialect each. Based on the confidence scores from four dialects and considering the predefined threshold logic is applied to decide the class of input. Performance of the system based on first classifier only is given in Table 3(uncompressed spectral feature) and Table 4(compressed spectral feature). The average performance of the system is 71% using MFCC, 68% with PLP and 72% with MF-PLP as uncompressed spectral feature combined with prosodic features. When compressed set of spectral features is used the performance in each case slightly increased. The recognition score was 73%, 70%, and 73% respectively with MFCC, PLP, and MF-PLP spectral features.

Table 3: System performance based on uncompressed spectral features along with F0, ΔF0 and Energy as a prosodic feature

Hindi Dialects	**Recognition Performance (%)**											
	KB			**HR**			**BP**			**BG**		
	MFCC	PLP	MF-PLP	MFCC	PLP	MF-PLP	MFCC	PLP	MF-PLP	MFCC	PLP	MF-PLP
KB	72	68	74	15	16	13	2	6	4	11	10	9
HR	14	12	11	69	72	74	7	12	14	10	4	1
BP	5	13	10	11	9	10	74	62	65	10	16	15
BG	9	7	5	5	3	3	17	20	17	69	70	75

Table 4: System performance based on compressed spectral features along with F0, ΔF0 and Energy as a prosodic feature

Hindi Dialects	Recognition Performance (%)											
	KB			HR			BP			BG		
	MFCC	PLP	MF-PLP	MFCC	PLP	MF-PLP	MFCC	PLP	MF-PLP	MFCC	PLP	MF-PLP
KB	75	68	71	13	12	15	2	8	2	10	12	12
HR	13	8	12	71	76	70	7	11	9	9	5	9
BP	5	16	9	10	9	9	75	64	77	10	11	5
BG	7	8	8	6	3	6	16	17	12	71	72	74

In the modified system, performance is further evaluated by combining the confidence score obtained from AANN models of the second classifier trained on syllable duration of input utterances with that of the output of the first classifier. Figure 5 represents the block diagram for AANN models.

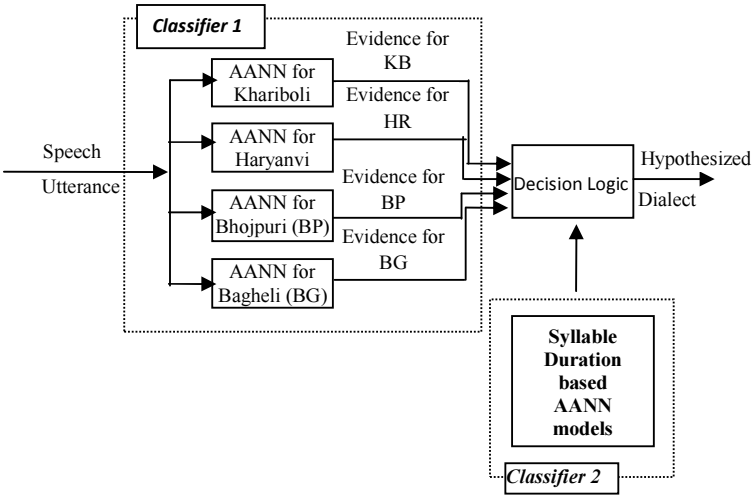

Figure 5. Block diagram of dialect identification system based on evidences from each dialect

During testing, the input utterance is segmented into syllable units, and the network is trained using their duration. Normalized mean square error **(E_u)** is computed for the complete utterance using; $E_u = \frac{1}{S}\sum_{s=1}^{S} ||x_s - o_s||^2$, where S is the total number of syllable in an utterance and x_s is the input syllable duration of s^{th} syllable, whereas, o_s is the observed output for the same syllable. With this error, the confidence score corresponding to each dialect is computed. Table 5 represents the influence of duration information on the performance of the

system, using compressed feature set. The performance of the combined system is drastically improved, and average recognition score is reached 81% for MFCC,78% for PLP and 82% with MF-PLP as spectral features. This significant improvement highlights the importance of speaker's tonal characteristic due to native dialect.

Table 5: System performance based on compressed spectral features along with F0, ΔF0, Energy and Syllable Duration

Hindi Dialects	**Recognition Performance (%)**											
	KB			**HR**			**BP**			**BG**		
	MFCC	PLP	MF-PLP	MFCC	PLP	MF-PLP	MFCC	PLP	MF-PLP	MFCC	PLP	MF-PLP
KB	84	81	82	10	8	8	0	3	3	6	8	7
HR	10	9	7	77	72	81	6	9	7	7	10	5
BP	0	2	4	10	14	11	83	81	81	7	3	4
BG	6	8	7	3	6	0	11	7	9	80	79	84

VII. SUMMARY AND CONCLUSION

In this paper, spectral and prosodic features are explored for dialect identification of spoken utterances. We have outlined the capability of auto-associative neural network for its use for dimension compression of speech features as well as for capturing dialect specific information from underlying distribution of feature vectors. Four Hindi dialects (Khariboli, Haryanvi, Bhojpuri, and Bagheli) have been considered in this work. To evaluate this model we have used speech samples collected from male and female speakers from these dialects. Two separate classifiers, each consisting of 4 AANN model, one for each dialect have been used in the identification process. In the baseline system, only one classifier is used. This classifier is trained with spectral features (MFCC, PLP, MF-PLP) along with F0, ΔF0 and Energy obtained from each speech frame. Decision regarding the class of the input utterance is based on the confidence score obtained from each frame. Model is evaluated for both the uncompressed and compressed set of spectral features. MF-PLP based spectral features give the best result in combination with prosodic features. But the improvement using this feature is negligible as compared to its complexity. MFCC features outperform PLP features significantly. Recognition performance of the system improved with the compressed feature

set. Also, the number of iterations in training drastically decreased. Syllable duration is included as an additional prosodic feature in the modified model and is used in the training of the second classifier. AANN model for all four dialects is trained using duration of all syllables in the input utterance. In the next level of execution, the confidence score obtained from the second classifier is reinforced by the decision logic based on results of the first classifier. The system performance increased in all cases. This increase in recognition performance shows that tonal characteristic influenced by native dialect of speakers is an important prosodic feature and syllable duration captures it properly. It also highlights that AANN is capable of capturing intrinsic feature characteristics specific to dialects.

The results obtained in this work are promising and demonstrates the potential of AANN as a candidate for dialect classification using spectral and prosodic features. Also, even if MF-PLP gives the better result than other spectral features; due to the simplicity of MFCC and its potentials it can be a promising candidate as spectral feature for further tasks. With this result, we are encouraged to increase our database and work for more dialects of Hindi. Identification of more spectral feature to capture details significant to Hindi dialects is also one of our future goals.

REFERENCES

[1] R. Huang, J. H. L. Hansen and P. Angkititrakul, “Dialect/Accent Classification using Unrestricted Audio”, IEEE Transaction on Audio, Speech and Language Processing, 15(2), pp. 453-464, 2007.

[2] J. C. Wells, “Accent of English”, 1982, VOL. 2; Cambridge University Press, Landon.

[3] S. Sinha, S. S. Agrawal and A. Jain, “Dialectal influences on acoustic duration of Hindi phonemes”, Proceeding of International Conference of The International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (OCOCOSDA), November 25-27, 2013. pp. 1-5.

[4] D. Mishra and K. Bali,“A comparative phonological study of the dialects of Hindi”, in Proceedings of International Congress of Phonetic Sciences XVII , August 17-21, 2011, pp. 1390-1393.

[5] A. H. M. Russell and M. Carey, “Human and computer recognition of regional accents and ethnic groups from British English speech”, Computer Speech and Language, 27(1), pp. 59-74, 2013.

[6] E. L. Goh, "Gender and accent identification for Malaysian English using MFCC and Gaussian mixture model", Doctoral dissertation, Faculty of Computing, Universiti Teknologi, Malaysia, 2013.

[7] P. Dhanalakshmi,S. Palanivel and V. Ramalingam, "Classification of audio signals using AANN and GMM", Applied Soft Computing, 11(1), pp.716-723, 2011.

[8] A. Waibel, "Prosody and speech recognition". Morgan Kaufmann, 1988.

[9] K. Sreenivasa Rao, "Role of neural network models for developing speech systems", Sadhana, 36(5), pp. 783-836, 2011.

[10] S. Gray and J.H.L. Hansen, **"**An integrated approach to the detection and classification of accents/dialects for a spoken document retrieval system". IEEE Workshop on Automatic Speech Recognition and Understanding, November 27- December 1, 2005, pp. 35-40.

[11] A.S. Ghotkar and G. K. Kharate, "Study of vision based hand gesture recognition using Indian sign language", International Journal on Smart Sensing and Intelligent Systems, 7(1), pp. 96-115, March 2014.

[12] K. S. Rao, and S. G. Koolagudi, "Identification of Hindi dialects and emotions using spectral and prosodic features of speech". International Journal of Systemics, Cybernetics and Informatics, *9*(4), pp. 24-33, 2011.

[13] S. Sinha, A. Jain and S. S. Agrawal, "Speech Processing for Hindi Dialect Recognition". Advances in Signal Processing and Intelligent Recognition Systems, Vol 264, pp. 161-169, 2014.

[14] R.K. Aggarwal and M. Dave, "Integration of multiple acoustic and language models for improved Hindi speech recognition system", International Journal of Speech Technology, 15(2), pp. 165-180, 2012.

[15] R.K. Aggarwal and M. Dave, "Performance evaluation of sequentially combined heterogeneous feature streams for Hindi speech recognition system", Telecommunication Systems, 52(3), pp. 1457-1466, 2013.

[16] A. Jansen and P. Niyogi, "A geometric perspective on speech sounds", Tech. Rep. TR-2004-06, University of Chicago, June 2005.

[17] A. Errity and J. McKenna, "A comparision of linear and nonlinear dimensionality reduction methods applied to synthetic speech", Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH), Brighton, September 6-10, 2009, pp. 1095-1098.

[18] Ma Zongming, "Sparse principal component analysis and iterative thresholding", The Annals of Statistics, 41(2), pp. 772-801, 2013.

[19] A. Zolnay et al., “Using multiple acoustic feature sets for speech recognition”. Speech Communication, 49(6), pp. 514-525, 2007.

[20] A. Che Soh, K.K.Chow, U. K. Mohammad Yusuf, A. J. Ishak, M. K. Hassan, S.Khamis, “Development of neural network-based electronic nose for herbs recognition”, International Journal on Smart Sensing and Intelligent Systems,7(2), pp. 584-609, June 2014.

[21] M. A. Kramer, “Nonlinear principal component analysis using autoassociative neural networks”. AIChE journal, Wiley online, 37(2), pp. 233-243, 1991.

[22] K. Sreenivasa Rao, D. Nandi and S. G. Koolagudi. "Film segmentation and indexing using autoassociative neural networks." International Journal of Speech Technology, 17(1), pp. 65-74, 2014.

[23] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”. IEEE Transactions on Acoustics, Speech and Signal Processing, 28(4), pp. 357-366, 1980.

[24] A. N. Mishra, M. Chandra, A. Biswas and S. N. Sharan, “Robust features for connected Hindi digits recognition”. International Journal of Signal Processing, Image Processing and Pattern Recognition, 4(2), 79-90, 2011.

[25] Marie-José Kolly and Volker Dellwo, "Cues to linguistic origin: The contribution of speech temporal information to foreign accent recognition", Journal of Phonetics, Vol. 42, pp. 12-23, 2014.

[26] A. Gaddam, G. Sen Gupta and S.C. Mukhopadhyay, “Sensors for Smart Home”, Chapter -7, of the book Human Behavior Recognition Technologies: Intelligent Applications for Monitoring and Security, edited by Hans Guesgen and Stephen Marsland, IGI Global, ISBN 978-1-4666-3683-5, page 130-156, 2013.

[27] M. Kulshreshtha and R. Mathur, “Dialect Accent Feature for Establishing Speaker Identity: A case study”, Springer Briefs in Electrical and Computer Engineering, 2012.

[28] Anindya Nag and Subhas Mukhopadhyay, Smart Home: Recognition of activities of elderly for 24/7; Coverage issues, Proceedings of the 2014 International Conference on Sensing Technology, Liverpool, UK, Sep. 2 to 4, 2014, pp. 480-489, ISSN 1178-5608, http://s2is.org/icst-2014/program.asp.

[29] M. Sigmund, “Statistical Analysis of Fundamental Frequency Based Features in Speech under Stress”. Information Technology and Control Journal, 42(3), pp. 286-291, 2013.

[30] S. A. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking". The Journal of the Acoustical Society of America, 123(6), pp. 4559-4571, 2008.

[31] Y.X. Lai, Y.M. Huang and S.C.Mukhopadhyay, Interconnecting Communication for Recognition and Automation services on Home Grid, Proceedings of IEEE I2MTC 2012 conference, IEEE Catalog number CFP12MT-CDR, ISBN 978-1-4577-1771-0, May 13-16, 2012, Graz, Austria, pp. 2346-2350.

[32] P. G. Deivapalan, M. Jha, R. Guttikonda and H. A. Murthy, "Donlabel: an automatic labeling tool for Indian languages." Proceedings of Fourteenth National Conference on Communication (NCC), February 1-3, 2008, pp. 263-268.

[33] T. Quazi, S.C. Mukhopadhyay, N. Suryadevara and Y. M. Huang, Towards the Smart Sensors Based Human Emotion Recognition, Proceedings of IEEE I2MTC 2012 conference, IEEE Catalog number CFP12MT-CDR, ISBN 978-1-4577-1771-0, May 13-16, 2012, Graz, Austria, pp. 2365-2370.

[34] B. Yegnanarayana, "Artificial Neural Networks". Prentice-Hall, New Delhi,2004,

[35] B. Yegnanarayana and S. P. Kishore, "AANN: an alternative to GMM for pattern recognition", Neural Networks, 15(3), 459-469, 2002.