Identification of Hindi Dialects and Emotions using Spectral and Prosodic features of Speech

K Sreenivasa Rao and Shashidhar G Koolagudi
School of Information Technology,
Indian Institute of Technology Kharagpur, Kharagpur - 721302, West Bengal, India.
E-mail: ksrao@iitkgp.ac.in, koolagudi@(ieee.org,yahoo.com)

Abstract— **In this paper, we have explored speech features to identify Hindi dialects and emotions. A dialect is any distinguishable variety of a language spoken by a group of people. Emotions provide naturalness to speech. In this work, five prominent dialects of Hindi are considered for the identification task. They are Chattisgharhi (spoken in central India), Bengali (Bengali accented Hindi spoken in Eastern region), Marathi (Marathi accented Hindi spoken in Western region), General (Hindi spoken in Northern region) and Telugu (Telugu accented Hindi spoken in Southern region). Along with dialect identification, we have also carried out emotion recognition in this work. Speech database considered for dialect identification task consists of spontaneous speech spoken by male and female speakers. Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC) is used for conducting the emotion recognition studies. The emotions considered in this study are anger, disgust, fear, happy, neutral and sad. Prosodic and spectral features extracted from speech are used for discriminating the dialects and emotions. Spectral features are represented by Mel frequency cepstral coefficients (MFCC) and prosodic features are represented by durations of syllables, pitch and energy contours. Auto-associative neural network (AANN) models and Support Vector Machines (SVM) are explored for capturing the dialect specific and emotion specific information from the above specified features. AANN models are expected to capture the nonlinear relations specific to dialects or emotions through the distributions of feature vectors. SVMs perform dialect or emotion classification based on discriminative characteristics present among the dialects or emotions. Classification systems are developed separately for dialect classification and emotion classification. Recognition performance of the dialect identification and emotion recognition systems is found to be 81% and 78% respectively.**

Keywords— **Auto-associative neural networks, Emotion recognition, Hindi dialect, Prosodic features, Spectral features, Support vector machines.**

I. Introduction

In this section, first we discuss about the importance of dialects, applications of automatic identification of dialects, speech features related to dialects, existing approaches for dialect identification and proposed methods for dialect identification. Later part of this section deals with need for emotion recognition, applications of emotion recognition, speech features for capturing emotions, literature survey on emotion recognition and proposed methods for the recognition of emotions in this work.

A. Dialect Identification

Dialects of a given language are the differences in speaking styles of a particular language, because of geographical and ethnic differences of the speakers. Number of studies have shown that the acoustic space spanned by phonemes for native speakers will shift when speakers are non-native. Other factors such as voice onset time, voiced stop release time, durations of the sound units and pitch contours are also play an important role while identifying the dialect [1]. Recent studies have considered the features extracted from spectral trajectories for dialect classification [2], [3].

Automatic dialect classification has several applications. For increasing the performance of the speech systems (such as speech recognition and speaker recognition), dialect identification at the front end will narrow down the search space and improve the performance further. For the natural human machine interface, dialect identification system will help the machine in understanding the speech spoken by the human and to synthesize the speech in the appropriate dialect of the person communicating with the machine [4], [1].

The dialect specific information is present in speech at different levels. At the segmental level, the dialect specific information can be observed in the form of unique sequence of the shapes of the vocal tract for producing the sound units. The shape of the vocal tract is characterized by the spectral envelope. In this work, spectral envelope is represented by Mel frequency cepstral coefficients (MFCC). At the suprasegmental level, the dialect specific knowledge is embedded in the duration patterns of the syllable sequences and the dynamics of the pitch and energy contours. At the subsegmental level, the dialect specific information may present in the shape of the glottal pulse and durations of open and close phases of vocal folds.

In this work, we have explored segmental and suprasegmental features for the identification of dialects of Hindi language. Usually, segmental features are extracted by analyzing the speech segments of duration 20-30 ms. Mostly, these features are extracted from the frequency spectrum of the speech segment, hence these features are known as spectral features. Suprasegmental features also known as prosodic features extracted from the speech segments of duration greater than 100 ms. Subsegmental features are extracted from the speech segments of duration less than 3 ms.

Automatic dialect identification studies were carried out for the languages of western and eastern countries such as USA (United States of America) and Japan [1], [5] . Few studies on the analysis of dialects of Indian languages are also observed. But, no systematic study is carried out on the dialects of any Indian language using the features derived from speech. Hence, we are exploring the features derived from speech for identifying the dialects of Hindi. At present, hundreds of dialects of Hindi are in use in different geographical regions of India. We have considered five prominent dialects of Hindi spoken in central, eastern, western, northern and southern regions of India. We named these five Hindi dialects using their local language names. Chattisgharhi is the local language in the central part of India. Hence, the dialect of Hindi spoken in that region is named as Chattisgharhi. Similarly, the dialects of Hindi spoken in eastern, western and southern regions of India are named as Bengali, Marathi and Telugu respectively. The local language of northern region of India is mainly, Hindi. Hence, it is named as General.

B. Emotion Recognition

Speech is the natural mode of communication among human beings. Human beings use emotions extensively to convey their intentions and feelings. For developing the speech interface to the machine, it is essential that the interface should be sophisticated enough to handle the emotions. For an effective human machine interaction, the machine should be able to adapt its interaction policies according to user's emotional status. Hence sophisticated interface to machine should ensure understanding of emotions expressed by humans, and responding back with appropriate emotions. The capability of a machine to process emotions, has several applications in day to day life. For example, in call center applications, the machine first analyzes the emotional state of the customer, and responds to the customer by itself, if he/she is in positive mood, otherwise transfers the call to human attendant. In case of story telling and E-tutoring applications, the system should automatically analyze the listeners'/ students' behavioral characteristics based on their emotions, and respond them accordingly with desired emotions. The automatic way to analyze the emotions in speech is useful for indexing and retrieving the audio files based on emotions. This is highly useful in organizing the movies and documentaries based on the emotional contents. Medical doctors may use the emotional contents of the patient's speech as a diagnosing tool for various ailments. Emotion analysis of tapped telephone conversation of criminals or terrorists would help crime investigation department to predict the activities of the extremists. Conversation with robotic pets and humanoid partners would be more realistic and enjoyable, if they are able to express and understand emotions like humans [6], [7], [8], [9], [10], [11], [12].

Speech is a composite signal that mainly carries information about the message to be conveyed, emotional content of the message, speaker characteristics and language information. The emotion specific characteristics of the speech can be attributed to (1) characteristics of the excitation source, (2) shape of the vocal tract system, while producing different emotions, (3) supra-segmental characteristics (prosodic parameters : energy, pitch and energy), (4) linguistic information and (5) emotional behavior of the speaker. Emotion specific characteristics of vocal tract are represented by its unique shapes while producing sound units in different emotions. Emotion specific knowledge at the supra-segmental level is characterized by unique patterns of duration, pitch and energy contours. The excitation source signal may also contain the emotion specific information in the form higher order relations among the LP residual samples, parameters of the instants of significant excitation and the characteristics of glottal pulse waveform.

Emotion specific vocal tract information is mainly represented by spectral features such as mel frequency cepstral coefficients (MFCCs), linear prediction cepstral coefficients (LPCCs) and their derivatives. Framewise parameters like pitch, duration, voice quality and energy are used as basic prosodic features. Their derivatives and sequence of prosodic features, extracted from longer speech segments are also used to categorize the emotions present in the speech [13]. McGilloway *et al.* have used the peaks and troughs in the profile of fundamental frequency and intensity, durations of pauses and bursts for identifying the four emotions namely fear, anger, sadness and joy. They have reported the classification performance of 55% using discriminant analysis [14]. Delleart *et al.* analyzed F_0 information for emotion classification and reported that minimum, maximum and median values of F_0 and slopes of F_0 contours are emotion salient features. The accuracy of recognition of four emotions is quoted as 79.5%, using K-nearest neighbor classifier [15]. Nicholas *et al.* analyzed the speech of radio artists involving 8 emotions using prosodic and phonetic features. Prosodic features used were power and intonation patterns, where as phonetic features adopted were linear prediction coefficients and their delta features. Recognition performance of 50% was achieved using neural networks [16]. Williams *et al.* and Norhaslinda *et al.* proposed log frequency power coefficients (LFPC) as feature parameters of speech to represent the energy distribution across the frequency spectrum, and a four stage ergodic hidden Markov model (HMM) was used as a classifier to classify six emotions. Performance of LFPC parameters was compared with conventional LPCC and MFCC features [17], [18]. Ververidies *et al.* have used short time supra-segmental features and their statistics for analyzing the emotions. Some of the prosodic features used by them include: pitch frequency F_0, energy, formant locations and their bandwidths, dynamics of pitch, energy and formant contours, speaking rate and transition time [9]. Iida *et al.* exploited the complex relations between pitch, duration and energy for detecting the emotions [19]. Gobl *et al.* combined vocal tract features, voice quality features and pitch dynamics to classify the emotions [20]. In addition to pitch related information, Kwon *et al.* used log energy, formants, mel based energy, MFCC velocity and acceleration coefficients, for classifying the emotions. They have achieved 96.3% success in classifying the stressed and neutral speech, and 70% for classifying the four speaking styles [21]. Wang *et al.* used 55 features (25 prosodic, 24 MFCCs and 6 formant frequencies) for recognizing six emotions. The recognition performance is observed to be 67% using Fisher's linear discriminate analysis [22].

For illustrating the presence of emotion specific information in spectral and prosodic features of speech, we have plotted the Figs 1 and 2 using spectral and prosodic parameters, respectively. Fig. 1 shows the discriminative capability of spectral features for the emotions considered in this work. The plot shows the LP spectrum of vowel /aa/ from the utterance *"maat<u>aa</u> aur pitaa kaa aadar karnaa chaahiee"* spoken in five different emotions. From the figure, it is observed that the spectral envelope of /aa/ for each of the emotions is distinct. The formant frequencies, their bandwidths, strengths and spectral roll-off are observed to be different for different emotions. In particular, higher formants and their strengths and bandwidths are highly discriminative.

Fig. 2 shows three subplots indicating the (a) duration patterns of the sequence of syllables, (b) energy contours and (c) pitch contours of an utterance *"maataa aur pitaa kaa aadar karnaa chaahie"* in five different emotions. The subplot indicating the duration patterns shows that for some emotions such as fear and happy, the durations of the initial syllables of the utterance are longer, for happy and neutral emotions middle syllables of the utterance seems to be longer, and the final syllables of the utterance seems to be longer for fear and anger (see Fig. 2(a)). From the energy plots, it is observed that the utterance with anger emotion has highest energy for the entire duration. Next to the anger emotion, fear and happy show little more energy over other two emotions. The dynamics of energy contours can be used to discriminate fear and happy (see Fig. 2(b)). Fig. 2(c) shows that anger, happy and neutral have some what higher pitch values, compared to other two emotions. Using the dynamics (changes with respect to time) of pitch contours, one can easily discriminate anger, happy and neutral emotions, even though, on an average they have higher pitch values. Thus, Figs. 1 and 2 provide the basic motivation to use spectral and prosodic features for discriminating the emotions.

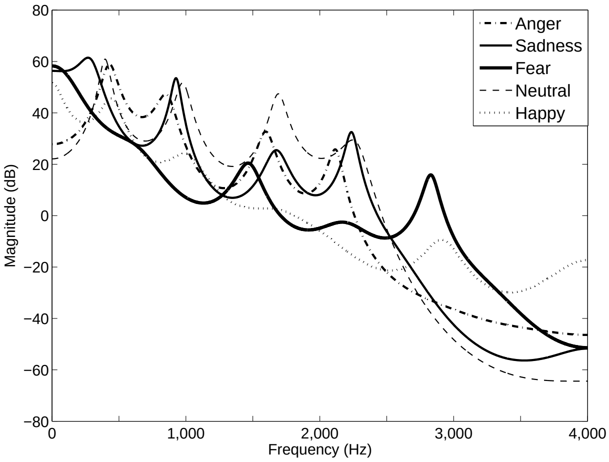

Fig. 1. LP spectrum of vowel /aa/ in five different emotions.

In this paper, Auto-associative neural networks (AANN) and Support Vector Machines are explored for capturing the dialect specific and emotion specific information from the proposed spectral and prosodic features. The reason for choosing these two models is that the classification philosophy is very different, and we want to explore which model performs better. AANN models are expected to capture the nonlinear relations specific to dialects or emotions through the distributions of feature vectors. whereas, SVMs perform dialect or emotion classification based on discriminative characteristics present among the dialects or emotions.

Rest of the paper is organized as follows: The details of the speech databases used in this study are discussed in Section 2. The details of the proposed neural network model and support vector machine for the identification of dialects and emotions are given in Section 3. Development of the dialect and emotion identification systems are discussed in Section 4. The evaluation details of the developed dialect and emotion identification systems are discussed in Section 5. Section 6, summarizes the contents of the paper, and also provides the future extensions to the present work.

II. Databases

In this section, the details of speech databases used for dialect identification and emotion recognition are discussed.

A. Hindi dialect speech corpus

Speech data is collected from five different geographical regions (central, eastern, western, northern and southern) of India, representing five dialects of Hindi. For each dialect, speech data is collected using five male and five female speakers. Speech data is collected from the speaker, by posing the questions arbitrarily such as to describe one's childhood, about the history of the home town, about the details of the career, views on habits and so on. From each speaker, 5-10 mins of speech is collected from the spontaneous response to the above questions. Altogether, for each dialect the duration of the speech collected, is about 1-1.5 hrs. Instead of reading some study material or uttering the small fixed text sentences, responses to the general questions usually contain the natural accent of the language. With this reason, we have used the spontaneous response to the questions as the speech material for the identification of dialects.

B. Indian Institute of Technology Kharagpur - Simulated Emotion Hindi Speech Corpus (IITKGP:SEHSC)

This speech corpus is recorded using 10 (5 male and 5 female) professional artists from All India Radio (AIR) Varanasi, India. The artists have sufficient experience in expressing the desired emotions from the neutral sentences. All the artists are in the age group of 25-40 years, and have professional experience of 8-12 years. The eight emotions considered for recording this database are anger, disgust, fear, happy, neutral, sadness, sarcastic and surprise. Fifteen emotionally neutral, Hindi sentences are chosen as text prompts for the database. Each of the artists has to speak 15 sentences in 8 given emotions in one session. The number of sessions recorded for preparing the database is 10. The total number of utterances in the database is 12000 (15 sentences X 8 emotions X 10 artists X 10 sessions). Each emotion has 1500 utterances. The number of words and syllables in the sentences vary from 3-6 and 11-18 respectively. The total duration of the database is around 7 hours. The speech samples are recorded using SHURE dynamic cardioid microphone C606N. The speech signal is sampled at 16 kHz, and each sample is represented as 16 bit number. The sessions were recorded on alternate days to capture the variability of human speech production system. Recording was done in such a way that each artist had to speak all sentences at a stretch in a particular emotion. This provides coherence among sentences for a specific emotion category. Entire speech database was recorded using single microphone at the same location, in a quiet room, without any obstacle in the recording path. The proposed speech database is the first one developed in an Indian language for analyzing the basic emotions. This database is sufficiently large to analyze the emotions in view of speaker, gender, text and session variability [7].

III. Classification models

In this work AANN and SVM models are explored to capture the emotion specific and dialect specific information from spectral and prosodic features. The details of AANN and SVM models are briefly provided in the following subsections.

A. Auto-associative neural network

Auto-associative neural network models are feed forward neural networks performing an identity mapping of the input space, and are used to capture the distribution of the input data [23], [24], [25]. In this work, a five layer AANN model as shown in Fig. 3 is used to capture the distribution of the feature vectors. The input and output (first and fifth) layers have same number of units. The second and fourth layers of the network have more units than the input layer. The third layer has fewer units than the input or output layers. The second, third and fourth layers are hidden layers.

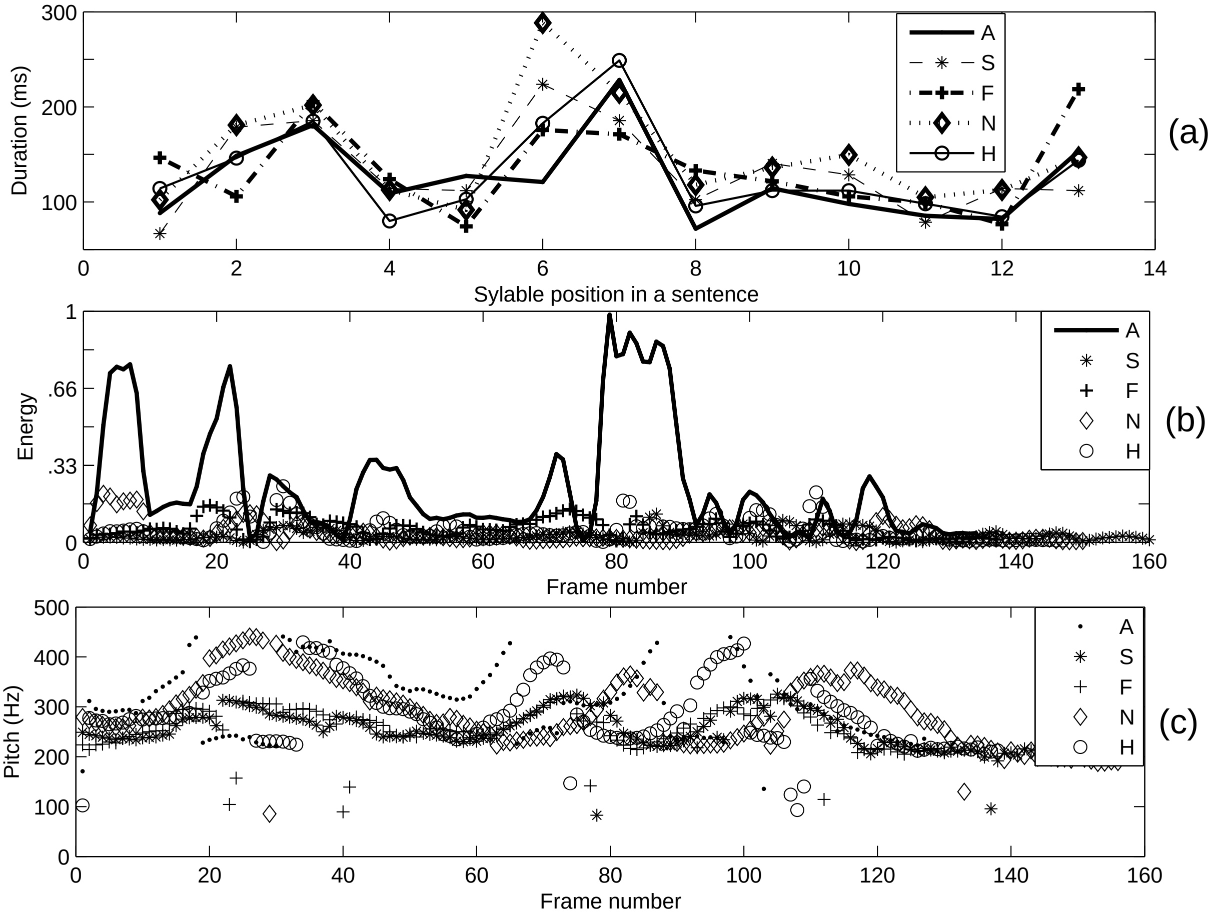

Fig. 2. (a) Duration patterns for the sequence of syllables, (b) Energy contours and (c) Pitch contours in different emotions for the utterance *"maataa aur pitaa kaa aadar karnaa chaahiee"*.

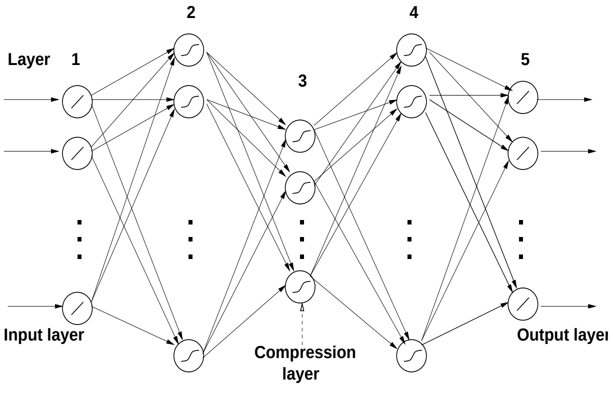

Fig. 3. Five layer Auto-associative neural network (AANN)

The second and fourth layer (first and third hidden layers) of the network has more units than the input layer, and it can be interpreted as capturing some local features in the input space. The third layer (second hidden layer) has fewer units than the first layer, and can be interpreted as capturing some global features. The activation functions of second, third and fourth layers are nonlinear, whereas first and fifth (input and output) layers are linear. The nonlinear units use $tanh(s)$ as the activation function, where s is the activation value of that unit. Generalization by the network is influenced by three factors: the size of the training set, the architecture of the neural network, and the complexity of the problem. We have no control over the first and last factors. In this study, network structures at different levels are arrived empirically. All the input and output features are normalized to the range [-1, +1] before presenting them to the neural network. The standard back propagation learning algorithm is used to adjust the weights of the network to minimize the mean square error for each feature vector.

B. Support Vector Machine

SVMs are designed for two-class pattern classification. Multiclass (n-class) pattern classification problems can be solved using a combination of binary (2-class) support vector machines. One-against-the-rest approach is used for decomposition of n-class pattern classification problem into $'n'$ two-class classification problems. The set of training examples $\left\{\{(x_i, k)\}_{i=1}^{N_k}\right\}_{k=1}^{n}$ consists of N_k number of examples belonging to the k^{th} class, where the class label $k \in \{1, 2, 3, ..., n\}$. All the training examples are used to construct the SVM for a class. The SVM for the class k is constructed using the set of training examples and their desired outputs, $\left\{\{(x_i, y_i)\}_{i=1}^{N_k}\right\}_{k=1}^{n}$ The desired output y_i for the training example x_i is defined as follows:

$$y_i = \begin{cases} +1 & \text{if } x_i \in k^{th} \text{ class} \\ -1 & \text{otherwise} \end{cases}$$

The examples with $y_i = +1$ are called positive examples, and those with $y_i = -1$ are negative ones. An optimal hyperplane is constructed to separate positive examples from negative ones. The separating hyperplane (margin) is chosen in such a way as to maximize its distance from the closest training examples of different classes. Fig. 4 illustrates the geometric construction of hyperplane for two dimensional input space. The support vectors are those data points that lie closest to the decision surface, and therefore the most difficult to classify. They have a direct bearing on the optimum location of the decision surface [26]. For a given test pattern x, the evidence $D_k(x)$ is obtained from each of the SVMs. In the decision logic, the class label k associated with SVM, which gives maximum evidence is hypothesized as the class (C) of the test pattern, that is

$$C(x) = argmax_k(D_k(x))$$

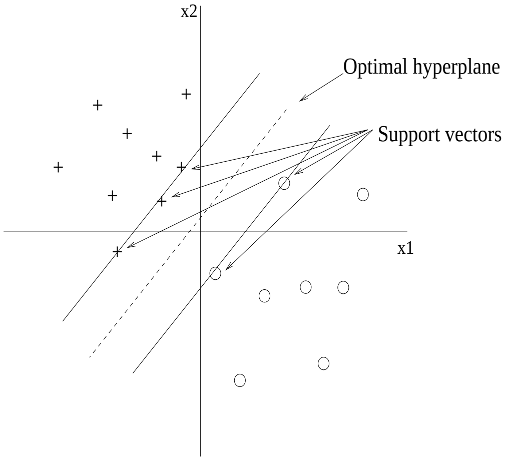

Fig. 4. Classification mechanism of support vector machine.

IV. Development of Dialect and Emotion Identification systems

For each dialect, four AANN models are developed using spectral, duration, pitch and energy features. MFCCs are used for representing the spectral features [27], [28]. MFCCs are extracted from a speech frame of 20 ms, with a frame shift of 10 ms. In this study, we use 13 dimensional MFCC feature vector to represent the speech frame. The derived MFCC feature vectors of a particular dialect are given as input and output of the AANN model. The reason for giving the feature vectors to input and output is to capture the distribution of the feature vectors. The number of epochs needed for training depends on the behavior of the training error. It is found that 100 epochs are adequate for the AANN models used in this work. For developing the AANN models using prosodic parameters, the size of all feature vectors should remain same. For deriving the prosodic feature vectors, speech data is segmented into phrases using the knowledge of longer pauses. The average phrase duration is observed to be about 2.5 secs, maximum and minimum phrase durations are observed to be 4.2 and 0.9 secs respectively. Maximum number of syllables in a phrase is found to be 23. Therefore, the size of the duration vector is fixed to 23 dimensions indicating 23 duration values. If the number of syllables in a phrase is less than 23, then the tail portion of the duration vector is appended with zeros to maintain the size of the duration vector to be 23. Syllable durations are determined automatically, using vowel onset points [29], [30]. The sequence of fundamental frequency values constitutes pitch contour. In this work, pitch contours are extracted from speech using the autocorrelation of the Hilbert envelope of the linear prediction residual signal [31]. Energy contour of a speech signal is derived from the sequence of frame energies. Frame energies are computed by summing the squared sample amplitudes. In this study, we have chosen the frame size of 20 ms and a frame shift of 10 ms. The size of pitch and energy contours of the phrases are proportion to the length of the utterance. To obtain the fixed dimensional vector, we have used resampling technique. The dimension of pitch and energy contours is chosen to be 23. Here, the dimension 23 for pitch and energy contours is not crucial. The reduced size of pitch and energy contours has to be chosen such that the dynamics of the original contours have to be retained in the resampled versions. The basic reasons for reducing the dimensionality of the original pitch and energy contours are (1) Need for the fixed dimensional input feature vector for developing the AANN models and (2) The number of feature vectors used for the training is proportion to the size of the feature vector. The structures of the AANN models used in this work are 13L 26N 6N 26N 13L and 23L 40N 10N 40N 23L for capturing the distributions of spectral and prosodic feature vectors, respectively. For training the AANN models speech data of 7 (3 female and 4 male) speakers is used. Speech data from the other 3 (2 female and 1 male) speakers is used for evaluating the models.

In general, support vector machines provide the classification based on discrimination principle. For a given feature vector, SVM decides whether it belongs to the class of interest or not. Class discrimination will be poor at low level features such as spectral features, better discrimination will be observed at the utterance level. Therefore, while using SVMs, features are expected to be derived from utterance level. But, spectral features are meaningful at the segmental level (i.e., frame of 20 ms). For developing the SVM model using spectral features, we need to derive single feature vector correspond to each utterance. In this work, first a Gaussian mixture model (GMM) is developed using the spectral features of utterance. The parameters of the mixture components (mean, variance, mixture weight) are concatenated to form a fixed dimensional spectral feature vector representing the utterance. These utterance level spectral feature vectors are used for developing the SVM. While developing the SVM, feature vectors representing the desired class are viewed as positive examples and the feature vectors belongs to all other classes (other than the desired) are viewed as negative examples. In case of prosodic features, SVMs will be developed directly. There is no need for using GMMs at the first stage, because prosodic features are extracted from the utterance only.

While developing emotion recognition system, the emotions models are developed in a similar way as of dialect specific models. In this study, we are using the emotional speech of 8 speakers (4 male and 4 female) for developing (training) the models, and the remaining 2 speakers (1 male and 1 female) speech is used for evaluating the performance of the developed models.

V. Evaluation of Dialect and Emotion Identification systems

A. Dialect Identification Systems

In this work, we have developed four Dialect Identification (DI) systems using (1) spectral features, (2) durations of syllables in the utterance, (3) pitch contour (sequence of pitch (F_0) values) and (4) energy contour. Each DI system consists of 5 AANN models representing the five dialects: Chattisgharhi (C), Bengali (B), Marathi (M), General (G) and Telugu (T). The block diagram of the basic DI system using AANN models is shown in Fig. 5. For evaluating the performance of the DI system, the feature vectors derived from the test speech utterances are given as input to five AANN models. The output of the each model is compared with the input to compute the normalized squared error. The normalized squared error (e) for the feature vector y is given by, $e = \frac{||y-o||^2}{||y||^2}$, where o is the output vector given by the model. The error e is transformed into a confidence score (c) using $c = exp(-e)$. The average confidence score is calculated for each model. The identity of the dialect is decided based on the highest confidence score. In this work, first we analyzed the performance of the four DI systems separately, and then they are combined using score level fusion.

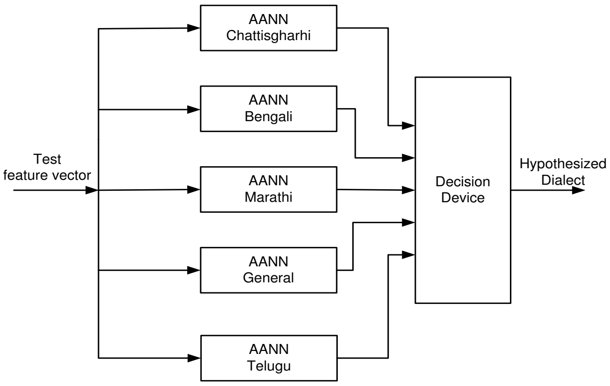

Fig. 5. Dialect identification system (DIS) using AANN models

Performance of the DI system using spectral features is given in Table I. The average identification performance is observed to be 62%. The diagonal entries of the table indicates the correct classification, and the rest indicate the misclassification. Performance of the DI system using prosodic features is given in Table II. Columns 2-6, 7-11 and 12-16 show the performance of the DI systems developed by duration, pitch and energy features respectively. The average performance is observed to be 55%, 61% and 48% for the DI systems developed using duration, pitch and energy features respectively. Columns 17-21 show the performance of the DI system by combining the confidence scores of the individual prosodic systems. The average performance of the combined prosodic system is observed to be much better (69%) compared to the individual prosodic systems. Finally, the evidences of spectral based DI system is combined with the evidences of the combined prosodic system. The block diagram of the combined DI system (spectral+prosodic) is shown in Fig. 6. The results of this combination has shown the drastic improvement with respect to their individual performances. The average performance of the combined system (i.e., spectral + prosodic) is found to be 78% (see Table III). The reason for the improved performance of the combined system may be due to the complementary nature of features.

TABLE I
Performance of the AANN based dialect identification system developed using spectral features. The entries in the table indicate the percentage of recognition. (C: Chattisgharhi, B: Bengali, M: Marathi, G: General and T: Telugu)

	Identification performance (%)				
	C	B	M	G	T
C	58	8	4	19	11
B	6	68	4	13	9
M	12	6	52	17	13
G	6	10	6	70	8
T	8	6	10	15	61

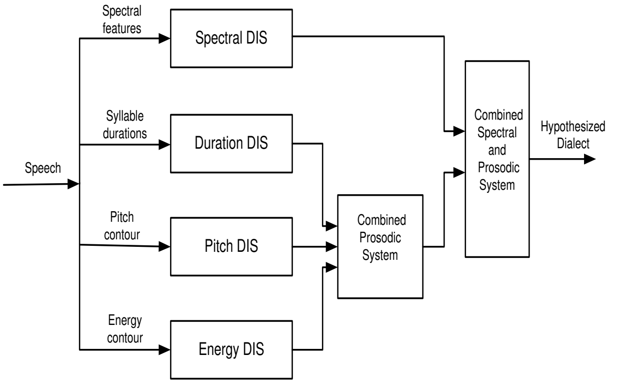

Fig. 6. Combined Dialect identification system using the evidences of spectral and prosodic DI systems.

TABLE III
Performance of AANN based dialect identification system by combining the evidences from the DI systems developed using spectral and prosodic features. The entries in the table indicate the percentage of recognition. (C: Chattisgharhi, B: Bengali, M: Marathi, G: General and T: Telugu)

	Identification performance (%)				
	C	B	M	G	T
C	80	6	4	7	3
B	5	77	8	4	6
M	6	9	74	5	6
G	2	2	4	86	6
T	4	9	6	8	73

We have also carried out the performance evaluation of dialect identification (DI) systems developed using SVMs. The process of development and evaluation of DI systems using SVMs is similar to the systems developed using AANNs. Table IV shows the performance of dialect identification systems developed using spectral and prosodic features. The numbers in the table indicate the percentage of identification. Column 1, indicates different dialects considered in this

TABLE II

Performance of AANN based dialect identification systems developed using duration, pitch and energy features. The entries in the table indicate the percentage of recognition. (C: Chattisgharhi, B: Bengali, M: Marathi, G: General and T: Telugu)

	Duration (D)					Pitch (P)					Energy (E)					Combined (D+P+E)				
	C	B	M	G	T	C	B	M	G	T	C	B	M	G	T	C	B	M	G	T
C	52	10	18	12	8	63	7	9	11	10	43	11	9	17	20	69	10	5	7	9
B	9	62	7	10	12	10	59	8	12	11	12	50	9	15	14	6	69	6	9	10
M	10	8	54	12	16	10	8	62	7	13	8	17	42	10	23	6	10	66	10	8
G	4	10	8	60	18	6	8	6	73	7	9	11	10	54	16	7	4	6	79	4
T	8	12	10	17	53	8	15	9	19	49	14	16	8	12	50	7	10	8	12	63

study. Column 2 indicates the identification performance using spectral features. Columns 3-5, indicate the identification performance using duration, pitch and energy features respectively. Column 6, indicates the identification accuracy after combining the evidence from duration, pitch and energy features. Finally, column 8 indicates the performance of DI system by combining the evidence from prosodic and spectral based systems. The performance of SVM based DI system developed using spectral features is about 64%. The average identification accuracy of SVM based DI systems developed using duration, pitch and energy is found to be 58%, 64% and 49%, respectively. Whereas the performance of DI system with their combination is observed to be about 73%. The combination of evidence from spectral and prosodic features have further raised the identification performance to 81%.

From the results (see Tables I,II,III, and IV), it is observed that prosodic features contain more dialect specific information compared to spectral features. However, due to presence of complimentary evidence in spectral and prosodic features, combining the evidence of DI systems developed using these features will enhance the identification performance. The combined (spectral plus prosodic) system also seems to be more robust against to the degradations.

TABLE IV

Performance of the SVM based dialect identification systems developed using different features (SP: spectral, D: duration, P: pitch, E: energy and PR: prosodic (D+P+E)) and their combinations. The entries in the table indicate the percentage of recognition. (C: Chattisgharhi, B: Bengali, M: Marathi, G: General and T: Telugu)

	Identification performance (%)					
	SP	D	P	E	D+P+E	SP + PR
C	61	54	61	41	72	79
B	69	61	64	52	74	81
M	56	58	66	45	72	78
G	68	62	75	57	79	85
T	65	56	52	52	70	80

B. Emotion Recognition Systems

In this work, emotion recognition is carried out using six basic emotions present in IITKGP-SEHSC database. The six emotions considered are anger (A), disgust (D), fear (F), happy (H), neutral (N) and sad (S). In this study also, both AANN and SVM models are used for developing emotion specific models. Six AANN and six SVM models are developed using each of the features. That is a total of 24 AANN and 24 SVM models are developed using spectral, duration, pitch and energy features. The block diagram of basic emotion recognition (ER) system using SVM models is shown in Fig. 7. The block diagram of ER system based on AANN models is also similar to Fig. 7.

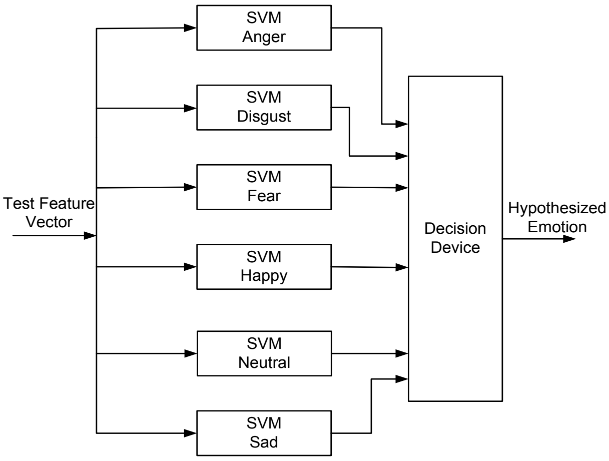

Fig. 7. Emotion recognition system (ERS) using SVM models

Performance of the SVM based ER system using spectral features is given in Table V. The average recognition performance is observed to be 73%.

TABLE V

Performance of the SVM based emotion recognition system developed using spectral features. (A: Anger, D: Disgust F: Fear, H: Happy, N:Neutral and S: Sad)

	Emotion recognition performance (%)					
	A	D	F	H	N	S
A	72	5	6	4	7	6
D	4	70	8	6	5	7
F	3	9	75	6	7	0
H	6	7	5	69	6	7
N	5	4	7	0	76	8
S	0	6	7	9	4	74

The performance of the ER systems developed using individual prosodic features is given in Table VI. Columns 2-7, 8-13 and 14-19 show the performance of the ER systems developed by extracted features from duration contour, pitch contour and energy contour, respectively. The diagonal en-

TABLE VI

Performance of SVM based emotion recognition systems developed using the features representing (1) Duration contour, (2) Pitch contour and (3) Energy contour. The entries in the table indicate the percentage of recognition. (A: Anger, D: Disgust F: Fear, H: Happy, N:Neutral and S: Sad)

	Duration contour						Pitch contour						Energy contour					
	A	D	F	H	N	S	A	D	F	H	N	S	A	D	F	H	N	S
A	77	0	7	13	3	0	54	0	13	20	13	0	40	20	13	20	7	0
D	10	50	23	7	10	0	20	60	20	0	0	0	13	34	20	13	0	20
F	3	33	47	10	7	0	13	7	67	0	0	13	0	7	80	0	0	13
H	27	6	10	57	0	0	17	0	0	83	0	0	7	26	0	54	13	0
N	0	0	7	3	73	17	7	13	0	0	67	13	0	13	0	7	54	26
S	0	0	3	7	10	80	7	0	13	0	26	54	0	0	13	0	33	54

tries of the corresponding ER systems indicate the correct recognition performance of the emotions considered in this study. The other entries indicate the percentage of misclassification. The average recognition performance of the ER systems developed using duration, pitch and energy contours is observed to be 64%, 67% and 53% respectively. From the classification results, it is observed that anger, disgust, fear and happy form a group (see rows 1-4 of Table VI), and other two emotions neutral and sad form the other group (see rows 5 and 6 of Table VI). Hence, the classification and misclassification is observed within the respective groups. This phenomenon is also observed in the performance of ER systems developed by feature and score level fusion methods (see Table VII).

For enhancing the performance of the individual ER systems, fusion techniques are tried out at feature and score levels. In this work, feature level fusion is performed by concatenating the individual prosodic features, and the ER system is developed using the concatenated feature vectors. The performance of the ER system developed using feature level fusion is given in columns 2-7 of Table VII. The average recognition performance using feature level fusion is observed to be 69%.

In this work, score level fusion is performed by summing the weighted confidence scores (evidences) derived from the ER systems developed using individual prosodic features. The weighting rule for combining the confidence scores of individual modalities is as follows: $c^m = \frac{1}{m}\sum_{i=1}^{m} w_i c_i$, where c^m is the multimodal confidence score, w_i and c_i are weighting factor and confidence score of the i^{th} modality, and m indicates number of modalities used for combining the scores. In this work, we have combined three modalities: (1) Model developed using durational features, (2) Model developed using sequence of pitch values and (3) Model developed using sequence of frame energies. In our study, one of the weights (w_i) is varied in steps of 0.1 from 0 to 1, and the other weights are determined using the formula: $w_j = \frac{1-w_i}{m-1}$, where j =1 to m and $j \neq i$, $i = 1$ to m. In this study, weighting factors associated to each system is varied from 0 to 1, with the steps of 0.1. With this we get a total of 33 combinations (11 combinations with respect to each weighting factor) of weighting factors. The recognition performance of the combined system for various combinations of the weighting factors is shown in Fig. 8. It is observed that the best recognition performance is about 74% for the weighting factors 0.2, 0.6 and 0.2 for the confidence scores of duration, pitch and energy contour based ER systems respectively. The details of the performance recognition is shown in the columns 8-13 of Table VII.

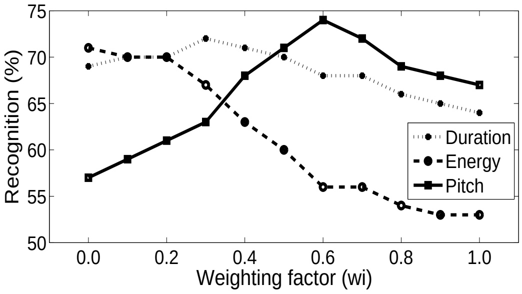

Fig. 8. Recognition performance for different combinations of weighting factors.

TABLE VII

Performance of SVM based emotion recognition systems developed by using (1) Feature level fusion and (2) Score level fusion using prosodic features. The entries in the table indicate the percentage of recognition. (A: Anger, D: Disgust F: Fear, H: Happy, N:Neutral and S: Sad)

	Feature level fusion						Score level fusion					
	A	D	F	H	N	S	A	D	F	H	N	S
A	80	0	3	17	0	0	77	0	3	20	0	0
D	0	57	23	13	7	0	10	63	17	3	7	0
F	0	20	63	7	10	0	7	17	70	3	0	0
H	23	0	3	67	7	0	23	0	0	70	7	0
N	0	7	0	0	73	20	0	0	0	7	83	10
S	0	7	7	0	13	73	0	0	0	3	17	80

From the analysis of DI systems, it is observed that evidence from spectral and prosodic features is complementary, and hence combining these evidence will further improve the classification performance. Therefore, in emotion recognition task also, we have explored the combination of evidence due to spectral and prosodic features. The performance of the combined ERS developed using the evidence from spectral and prosodic feature based ERS is given in Table VIII. Recognition performance of the combined system has found to be about 78%, and it has been observed to be improved compared to the individual ERS. AANN based ERS are developed and evaluated similar to SVM based ERS. The recognition performance of various AANN based ERS is summarized in Table IX. The average recognition performance of AANN based ER systems developed using spectral, duration, pitch, energy, prosodic and spectral plus prosodic features is observed to be 69%, 60%, 61%, 49%, 70% and 74% respectively. From the evaluation results of

TABLE VIII
Performance of the SVM based emotion recognition system developed by combining the evidence from ERS developed using spectral and prosodic features. (A: Anger, D: Disgust F: Fear, H: Happy, N:Neutral and S: Sad)

	Emotion recognition performance (%)					
	A	D	F	H	N	S
A	76	2	4	0	10	8
D	4	75	6	5	4	6
F	3	6	81	5	5	0
H	3	4	2	78	6	7
N	3	4	7	0	80	6
S	2	6	5	7	4	76

TABLE IX
Performance of the AANN based emotion recognition systems developed using different features (SP: spectral, D: duration, P: pitch, E: energy and PR: prosodic (D+P+E)) and their combinations. The entries in the table indicate the percentage of recognition.

	Identification performance (%)					
	SP	D	P	E	D+P+E	SP + PR
Anger	68	71	52	38	72	71
Disgust	67	52	55	30	60	69
Fear	70	50	62	72	66	80
Happy	67	51	77	52	64	72
Neutral	71	66	65	51	81	78
Sad	69	72	55	50	79	74

ER systems, it is observed that spectral features contains more emotion specific information compared to individual prosodic features. However, combination of prosodic features competes with spectral features for discriminating the emotions. Combining the evidence of ERS developed using spectral and prosodic features has improved the recognition performance indicating that spectral and prosodic features capture complementary emotion specific information.

VI. Summary and Conclusions

In this paper, spectral and prosodic features extracted from speech were explored for the identification of Hindi dialects and emotions. In this work, dialect identification and emotion recognition tasks were performed separately. Autoassociative neural network models and support vector machine models were used to capture the dialect specific information and emotion specific information from spectral and prosodic features. Five dialects of Hindi considered in this study are Chattisgharhi (spoken in central India), Bengali (Bengali accented Hindi spoken in Eastern region), Marathi (Marathi accented Hindi spoken in Western region), General (Hindi spoken in Northern region) and Telugu (Telugu accented Hindi spoken in Southern region). Speech corpus used for dialect identification, was collected from the spontaneous response of the speakers, when they were asked some general questions. Emotion recognition studies were conducted using IITKGP-SEHSC. The emotions considered in this study are anger (A), disgust (D), fear (F), happy (H), neutral (N) and sad (S). DI and ER systems were developed using individual (spectral, duration, pitch and energy) features and with their combination. Recognition systems were also developed separately using AANN and SVM models for analysing the capturing ability of the models. The performance of various DI and ER systems developed using spectral and prosodic features is summarized in Table X.

TABLE X
Performance of DI and ER systems developed using different features (SP: spectral, D: duration, P: pitch, E: energy and PR: prosodic (D+P+E)) and their combinations. The entries in the table indicate the percentage of recognition.

Rec sys	Rec model	Recognition performance (%)					
		SP	D	P	E	D+P+E	SP + PR
DI	AANN	62	55	61	48	69	78
	SVM	64	58	64	49	73	81
ER	AANN	69	60	61	49	70	74
	SVM	73	64	67	53	74	78

From the studies made in this paper, we can conclude that prosodic features contains more dialect specific information compared to spectral features. In the context of speech emotions, spectral features contain more emotion specific information over individual prosodic features. Individual prosodic features contain some non-overlapping emotion specific information, hence their combination has demonstrated the improvement in performance. From these studies, the common observations are (i) Spectral and prosodic features provide complementary evidence in view of identification of dialects as well as recognition of emotions, (ii) Individual prosodic features (duration, pitch and energy) capture some non-overlapping dialect and emotion specific information, hence their combination in both the systems has enhanced the identification accuracy, and (iii) Performance of SVM based dialect identification and emotion recognition systems is slightly better compared to AANN based systems.

In this work, we have explored only spectral (vocal tract) and prosodic aspects of speech, another contributory component of speech (i.e., excitation source) has not considered. Therefore, one can explore excitation source features for identification of dialects and recognition of emotions. The evidences from the source features can be combined with the evidences obtained using spectral and prosodic features for developing the robust dialect and emotion identification systems.

References

[1] L. M. Arslan and J. Hansen, "A study of temporal features and frequency characteristics in american english foreign accent," *Journal of Acoustic Society of America*, vol. 102, pp. 28–40, July 1996.

[2] P. Angkititrakul and J. H. Hansen

[3] J. Hansen, U.Yapanel, R. Huang, and A. Ikeno, "Dialect analysis and modeling for automatic classification," in *Interspeech-2004/ICSLP-2004: Inter. Conf. Spoken Language Processing*, (Jeju Island, South Korea), pp. WeC2302p.5(1–4), October 2004.

[4] C. Blackburn, J. Vonwiller, and R. King, "Automatic accent classification using artificial neural networks," in *In Proc. of the European Conf. on Speech Comm. and Tech.*, (Berlin, Germany), pp. 1241–1244, 1993.

[5] S. Itahashi and K. Tanaka, "A method of classification among japanese dialects," in *In Proceedzngs of Eurospeech 93*, pp. Vloume–1,639–642, September 1993.

[6] C. M. Lee and S. S. Narayanan, “Toward detecting emotions in spoken dialogs,” *IEEE Trans. Speech and Audio Processing*, vol. 13, pp. 293–303, March 2005.

[7] I. R. Murray, J. L. Arnott, and E. A. Rohwer, “Emotional stress in synthetic speech: Progress and future directions,” *Speech Communication*, vol. 20, pp. 85–91, Nov. 1996.

[8] R. Nakatsu, J. Nicholson, and N. Tosa, “Emotion recognition and its application to computer agents with spontaneous interactive capabilities,” *Knowledge based systems*, vol. 13, pp. 497–504, 2000.

[9] D. Ververidis and C. Kotropoulos, “A state of the art review on emotional speech databases,” in *Eleventh Australasian International Conference on Speech Science and Technology*, (Auckland, New Zealand), Dec. 2006.

[10] T.V.Sagar, “Characterisation and synthesis of emotionsin speech using prosodic features,” Master’s thesis, Dept. of Electronics and communications Engineering, Indian Institute of Technology Guwahati, May. 2007.

[11] R. Cowie and R. R. Cornelius, “Describing the emotional states that are expressed in speech,” *Speech Communication*, vol. 40, pp. 5–32, Apr. 2003.

[12] D. Ververidis and C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods,” *SPC*, vol. 48, p. 11621181, 2006.

[13] T. L. Nwe, S. W. Foo, and L. C. D. Silva, “Speech emotion recognition using hidden Markov models,” *Speech Communication*, vol. 41, pp. 603–623, Nov. 2003.

[14] S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk, and S. Stroeve, “Approaching automatic recognition of emotion from voice: A rough benchmark,” (Belfast), 2000.

[15] F. Dellert, T. Polzin, and A. Waibel, “Recognizing emotion in speech,” (Philadelphia, PA, USA), pp. 1970–1973, 4th International Conference on Spoken Language Processing, October 3-6 1996.

[16] N. J., T. K., and N. R., “Emotion recognition in speech using neural networks,” in *6th International Conference on Neural Information Processing*, pp. 495–501, ICONIP-99, 1999.

[17] C. E. Williams and K. N. Stevens, “Vocal correlates of emotional states,” *Speech Evaluation in Psychiatry*, p. 189220., 1981. Grune and Stratton Inc.

[18] N. Kamaruddin and A. Wahab, “Features extraction for speech emotion,” *Journal of Computational Methods in Science and Engineering*, vol. 9, no. 9, pp. 1–12, 2009. ISSN:1472-7978 (Print) 1875-8983 (Online).

[19] A. Iida, N. Campbell, F. Higuchi, and M. Yasumura, “A corpus-based speech synthesis system with emotion,” *Speech Communication*, vol. 40, pp. 161–187, Apr. 2003.

[20] C. Gobl and A. Chasaide, “The role of voice quality in communicating emotion, mood and attitude,” *SPC*, vol. 40, pp. 189–212, 2003.

[21] O. Kwon, K. Chan, J. Hao, and T. Lee, “Emotion recognition by speech signals,” (Geneva), pp. 125–128, Eurospeech, 2003.

[22] Y.Wang and L.Guan, “An investigation of speech-based human emotion recognition,” pp. 15–18, IEEE 6th Workshop on Multimedia Signal Processing, 2004.

[23] B. Yegnanarayana and S. P. Kishore, “AANN an alternative to GMM for pattern recognition,” *Neural Networks*, vol. 15, pp. 459–469, Apr. 2002.

[24] S. Haykin, *Neural Networks: A Comprehensive Foundation.* New Delhi, India: Pearson Education Aisa, Inc., 1999.

[25] B. Yegnanarayana, *Artificial Neural Networks.* New Delhi, India: Prentice-Hall, 1999.

[26] K. S. Rao, *Acquisition and incorporation prosody knowledge for speech systems in Indian languages.* PhD thesis, Dept. of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, India, May 2005.

[27] J. Benesty, M. M. Sondhi, and Y. Huang, eds., *Springer Handbook on Speech Processing.* Springer Publishers, 2008.

[28] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition.* Englewood Cliffs, New Jersy: Prentice-Hall, 1993.

[29] S. R. M. Prasanna and J. M. Zachariah, “Detection of vowel onset point in speech,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Orlando, Florida, USA), May 2002.

[30] S. R. M. Prasanna, *Event-Based Analysis of Speech.* PhD thesis, Dept. of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, India, Mar. 2004.

[31] S. R. M. Prasanna and B. Yegnanarayana, “Extraction of pitch in adverse conditions,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Montreal, Canada), May 2004.