IDENTIFICATION OF HINDI DIALECTS USING SPEECH

K Sreenivasa Rao, Sourav Nandy and Shashidhar G Koolagudi

School of Information Technology
Indian Institute of Technology Kharagpur
Kharagpur - 721302, West Bengal, India.
E-mail: ksrao@iitkgp.ac.in, sourav.joy@gmail.com, koolagudi@yahoo.com

ABSTRACT

In this paper, we have explored speech features to identify Hindi dialects. A dialect is any distinguishable variety of a language spoken by a group of people. 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). Speech database considered for this study consists of spontaneous speech spoken by male and female speakers. Prosodic and spectral features extracted from speech are used for discriminating the dialects. Spectral features are represented by Mel frequency cepstral coefficients (MFCC) and prosodic features are represented by durations of syllables, pitch and energy contours. Autoassociative neural network (AANN) models are used to capture the dialect specific information from the distributions of the feature vectors. The recognition performance of the developed models is observed to be 62%, 73% and 79% using the spectral, prosodic and spectral plus prosodic features, respectively.

1. INTRODUCTION

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, 2]. Recent studies have considered the features extracted from spectral trajectories for dialect classification [3, 4].

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 [5, 2].

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 [2, 6] . Few studies on the analysis of dialects of Indian languages are 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, there are 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. The local language of northern region of India is mainly, Hindi. Hence, it is named as General.

In this paper, Autoassociative neural networks (AANN) are explored for capturing the dialect specific information from the proposed spectral and prosodic features. The rest of the paper is organized as follows: The details of the speech database used in this study are discussed in Section 2. The details of the proposed neural network model for the identification of dialects are given in Section 3. Development of the dialect identification systems is discussed in Section 4. The evaluation details of the developed dialect 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.

2. DATABASE

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 describe the childhood, history of the home town, details of the career, 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 speech will be 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.

3. AUTOASSOCIATIVE NEURAL NETWORK (AANN)

Autoassociative neural network models are feedforward neural networks performing an identity mapping of the input space, and are used to capture the distribution of the input data [7, 8, 9]. In this work, a five layer AANN model as shown in Fig. 1 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 activation functions of second, third and fourth layers are

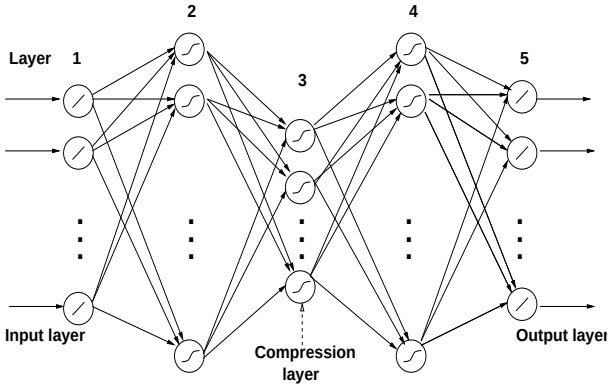

Fig. 1. Five layer Autoassociative neural network (AANN)

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. The standard backpropagation learning algorithm is used to adjust the weights of the network to minimize the mean square error for each feature vector.

4. DEVELOPMENT OF DIALECT 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 [10, 11]. 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 emotion 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 [12, 13]. 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 [14]. 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.

5. EVALUATION OF 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. 3. 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.

Performance of the DI system using spectral features is given in Table 1. The average identification performance

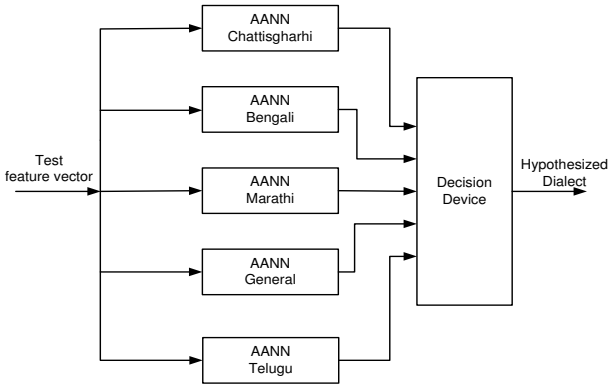

Fig. 2. Dialect identification system (DIS) using AANN models

is observed to be 62%. The diagonal entries of the table indicates the correct classification, and the rest indicates the misclassification. Performance of the DI system using

Table 1. Performance of the 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

prosodic features is given in Table 2. 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. **??**. 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

Table 2. Performance of the 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

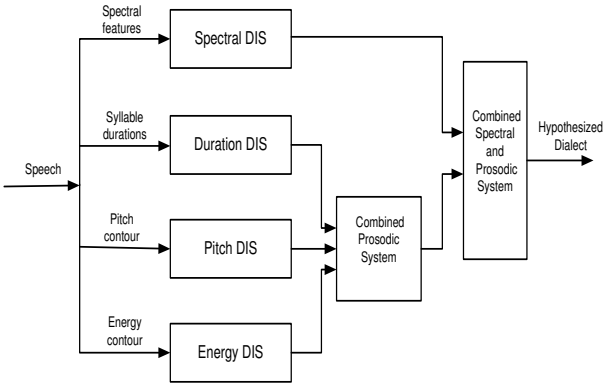

Fig. 3. Combined Dialect identification system using the evidences of spectral and prosodic DI systems.

Table 3. Performance of the 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

be 78% (see Table 3). The reason for the improved performance of the combined system may be due to the complementary nature of features.

6. SUMMARY AND CONCLUSIONS

In this paper, spectral and prosodic features extracted from speech were explored for the identification of Hindi dialects. Autoassociative neural network models were used to capture the dialect 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 in this study was collected from the spontaneous response of the speakers, when they were asked some general questions. The average performance of the dialect identification system was found to be 62%, 69% and 78% using spectral, prosodic and combined spectral and prosodic features respectively. The increased performance of the combined system can be attributed to the complementary evidences given by the spectral and prosodic features. Excitation source features can be explored for identification of dialects. The evidences from the source features can be combined with the evidences obtained using spectral and prosodic features for developing the robust dialect identification system.

7. REFERENCES

[1] L. Arslan and J. Hansen, "Language accent classification in american english," *Speech Communication*, vol. 18, pp. 353–367, July 1996.

[2] L. 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 1997.

[3] P. Angkititrakul and J. L. Hansen, "Stochastic trajectory model analysis for accent classification," in *ICSLP:*, (Denver, CO, USA), pp. 493–496, Sept. 2002.

[4] P. Angkititrakul and J. L. Hansen, "Use of trajectory models for automatic accent classification," in *Proc.*

Eurospeech, (Geneva, Switzerland), pp. 1353–1356, Sept. 2003.

[5] C. S. Blackburn, J. P. Vonwiller, and R. W. King, "Automatic accent classification using artificial neural networks," in *Proc. Eurospeech*, vol. 2, pp. 1241–1244, Sept. 1993.

[6] S. Itahashi and K. Tanaka, "A method of classification among japanese dialects," in *Proc. Eurospeech*, vol. 1, pp. 639–642, Sept. 1993.

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

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

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

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

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

[12] 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.

[13] 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.

[14] 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.