CSIT (June 2017) 5(2):209–216
DOI 10.1007/s40012-016-0145-5

SPECIAL ISSUE VISVESVARAYA 2016 OF CSIT

Soft computation based spectral and temporal models of linguistically motivated Assamese telephonic conversation recognition

Mridusmita Sharma[1] · **Kandarpa Kumar Sarma**[2]

Received: 1 November 2016 / Accepted: 16 December 2016 / Published online: 26 December 2016

Abstract Speech is the natural communication means, though not the typical input means afforded by computers. The interaction between humans and machines would have become easier, if speech were an alternative effective input means supplementing the keyboard and mouse. With advancement in techniques for signal processing and model building leading to the empowerment of computing devices with expanding list of abilities, significant progress has been made in speech recognition research, and various speech based applications have been developed. In such a backdrop, telephone speech technology have been receiving more attention in many new applications of spoken language processing. From the literature it has been found that the spectro-temporal features gives a significant performance improvement for telephone speech recognition in comparison to the conventionally used features for speech/speaker identification. Speech recognition systems can be characterized by many parameters. The commonly used method to measure the performance of a speech recognition system is the recognition accuracy. For obtaining proper accuracy it is necessary to design an efficient classifier for the recognition purpose which will lead to correct recognition results.

✉ Mridusmita Sharma
mriduzb@gmail.com

Kandarpa Kumar Sarma
Kandarpaks@gmail.com

[1] Department of Electronics and Communication Engineering, Gauhati University, Guwahati, Assam 781014, India

[2] Department of Electronics and Communication Technology, Gauhati University, Guwahati, Assam 781014, India

Keywords Speech recognition · Speaker identification · Emotion recognition · Spectro-temporal features · i-Vectors · Assamese language

1 Introduction

Speech is the natural means of communication which is easy, fast to communicate and do not require elaborate frameworks for manipulation, storage, retrieval and use, unlike other forms of communication like videos, data, etc. With the increasing popularity of the speech processing technology, the Human–Machine Interaction (HMI) has become integral elements of a host of present day multimedia systems. It is much comfortable on the part of the human being to communicate directly with the machine than to use primitive interfaces such as keyboard, mouse or other pointing devices because of the fact that such devices like keyboard and pointing devices require certain amount of skills for their effective usage. In order to use the computer efficiently, apart from a certain level of literacy, the user is also expected to have a sound proficiency in English or in certain contact languages (like Chinese, Russian, French, etc.) and a proper typing skill. However, a physically challenged person finds it difficult to use the computer as well as the interface devices. In such situations, speech based HMI plays a significant role. It is especially significant for a large country like India with vast linguistic and ethnographic diversity. The present work focuses on such an aspect with a stress to develop certain speech based HMI applications in Assamese. Recognition of speech in the regional dialects permits the removal of digital divide among the computer illiterate people and the people with good computer knowledge. Dialect recognition hence permits a huge section of the

population to use the benefits of technology [1]. Research and development in the field of speaker recognition dates back to the last century and this area is still an active topic for research. The application of speaker recognition technology has been continually growing in various fields of application such as forensic applications, dynamic signatures, gait, keystroke recognition, data encryption purpose, user validation in contact centres [2].

Now-a-days, telephone speech technology is getting more acceptance in many new applications of spoken language processing such as Voice Service Centre in hotels and restaurants, voice navigation in traffic and transportation systems, call centre support in medical, banking, agriculture etc. sectors and many more. However, there are significant challenges that need solutions while designing system for real time telephone speech recognition with better accuracy with regional linguistic orientation. Greater reliability in real time telephonic speech recognition is another constraint often seen while dealing with the speech that comes through a telephone channel. Recording over the telephone lines introduces severe distortions due to the variations in the transmission channels [3]. Speech recognition over the lines has also become an integral part of the various applications of Large Vocabulary Continuous Speech Recognition (LVCSR).

Also, India being a country with vast linguistic variations, it provides a sound area of research towards language-specific speech recognition technology. However, not much work has been done in Assamese as well as other north-eastern languages of the country in telephonic speech recognition and related development of Human Computer Interface (HCI) systems. Hence, the primary objective of this work is to develop certain soft computational models for linguistically motivated telephonic speech recognition which can be effectively used as HCI segment in present day and upcoming multimedia and computing devices.

1.1 Background

One of the most common approaches in speech recognition involves frame based approach where development of front end feature measurement coefficients or spectral analysis is done. This involves feature detection in the frequency space as in Fast Fourier Transform (FFT) or in time space as in Linear Predictive Coding (LPC) or in cepstral space as in Mel Frequency Cepstral Coefficient (MFCC) [4]. Pattern Recognition Technique is purely statistical for which the analog pattern must be digitized by sampling. Speech being highly dynamic in time, short time stationary signal is obtained by forming small comparable frames and windowing it to remove the discontinuity at the end. There are many techniques to compare two signals in the frequency domain. Some popular techniques are auto-correlation, Discrete Fourier Transform (DFT), Fast Fourier transform (FFT), Linear Predictive Coding (LPC) and Mel Frequency Cepstral Coefficient (MFCC). Since speech signal is non-stationary so its FFT is not possible. Short time Fourier transform may be calculated for a small frame that is duly windowed. A detailed study of the literature has drawn a clear picture of the various works reported in the area of speech recognition across the globe. The foremost requirement of a speech recognition technology is the design of a robust speech recognition system so that it eases the communication between human and machines [5]. Feature extraction plays a vital role in the design of an efficient speech recognition system. Different experiments carried out by various researchers all over the world have come to the conclusion that the use of spectral and temporal features is effective in the recognition of the speech sound. The emotion recognition system also yields better results with the use of combined features like formants and spectral features. A recent result from the physiological and psychoacoustic studies reveals the spectrally and temporally localized time–frequency envelope patterns a relevant basis of auditory perception. This however motivates the invention of novel feature extraction approaches for ASR that utilizes two dimensional spectro-temporal filters. The i-vector based paradigm has also been widely accepted by the speaker recognition as well as the Language Identification research groups. It has also been popularly used for the speech recognition purposes. The i-vector framework have greatly contributed towards the reduction of a series of acoustic feature vectors of a speech utterance to a low dimensional vector representation and hence enhance the performance of the speaker recognition technology and language identification system with lower computation requirements [6].

ASR is one of the important areas of research since it helps people to interact with the machine in an effective manner. Speech recognition systems can be characterized by many parameters. The commonly used method to measure the performance of a speech recognition system is the recognition accuracy. And for measuring the accuracy it is necessary to design an efficient classifier for the recognition purpose which will lead to correct recognition results.

2 Basic considerations

In order to get a proper understanding of the speech recognition problems, it is necessary to study the basic theoretical considerations of speech recognition and other such related topics. This section provides a brief overview of the basic theoretical considerations related to speech recognition, various feature extraction methods and soft computing techniques.

2.1 Feature extraction techniques

The objective behind feature extraction is to transform the input signal space to an output signal in a feature space with the help of prior knowledge in order to achieve some desired criteria. Feature extraction in speech processing is necessary to reduce the complexity of the problem [7].

There are various techniques for feature but mostly used technique is MFCC.

2.1.1 Mel frequency cepstral coefficients (MFCC)

There are various methods of parametric representation for the acoustic signals. Among them MFCC is the most widely used technique. The computation of MFCC vector is based on short term analysis and thus from each frame MFCC vector is computed. To minimize the discontinuities of the input speech sample, Hamming window is applied to the sample to extract the coefficients. It also decreases spectral distortions created by overlapping. For phonetically important characteristics, speech signal is expressed in Mel Frequency Scale [8]. The general formula for calculating MFCC is

$$Mel(f) = 2595 * log10(1 + f/700) \quad (1)$$

Figure 1 below shows the steps involved in MFCC feature extraction.

2.1.2 i-Vectors

The concept of i-vector or identity vector was originally proposed by Dehak et al. in the year 2010 [9]. The i-vector representation is used as a data driven approach for feature extraction which consist of mapping a number of frames of a speech utterances into dimensionally reduced speech representation super vector based on high dimensional linear GMM and traditional low dimensional MFCC features. In i-vector extraction the channel variability is included in the total variability subspace which includes many standard channel compensation techniques to attenuate channel variability in the i-vector extraction. The variations in channel may include mismatch between training and test utterances, arising from the differences in microphones, acoustic environments, transmission channels (such as telephone) etc. A speaker and channel dependent GMM super vector, S_{gmm}, can be represented as follows:

$$S_{gmm} = S_{ubm} + Tw \quad (2)$$

where, S_{ubm}, is the speaker and the channel independent universal Background Model (UBM) super vector, T is the total variability subspace which is a low rank matrix that represents the primary directions for variations across a large collection of data to be developed. Using training samples, T is optimized in such a way that the within-class variability is minimized and the inter-class variability is maximized as described in [10]. w is normally distributed with parameters N(0;1), and is the i-vector representation of the utterance. Once the matrix T is calculated, the i-vector, w, can be found out for speech utterances from Eq. 2. Figure 2 shows the basic steps for i-vector extraction.

2.2 Various classifiers for speech recognition

Classification is another important part of a speech recognition system since the patterns are classified into different classes during this stage. The decisions of a classifier are based upon the similarity measures from the training patterns then they are tested using the unknown patterns. The main objective of the classifier is to produce a model from the training data which predicts the target values of the test data.

Some of the classifiers used for speech recognition is discussed below

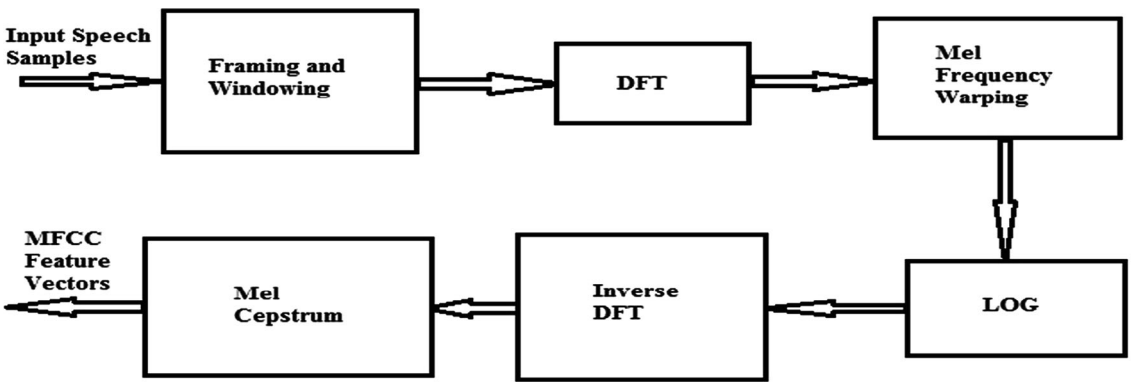


Fig. 1 Steps involved in MFCC extraction

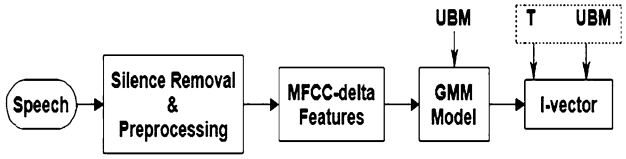


Fig. 2 Basic block diagram of i-vector extraction

2.2.1 *Multi-layer feed forward neural network (MLFNN)*

An MLFNN is the simplest form of neural network where information moves only in one direction. Here, information enters into the input layer and moves towards the output layer of neurons through one or more stages of hidden layers. A schematic diagram of MLFNN is shown in the Fig. 3.

2.2.2 *Recurrent neural network (RNN)*

In an RNN the output of the network not only depends on the weights, bias and inputs, but also the sequence of previous inputs the network has encountered during the operation. The schematic diagram of a RNN is shown in Fig. 4.

2.2.3 *Time delay neural network (TDNN)*

TDNNs are structurally similar to MLFNNs except the input layer. In the input layer, each neuron has some delayed input connections from D_0 to D_n [11]. Figure 5 shows a TDNN unit.

2.2.4 *Support vector machine (SVM)*

SVM uses linear and non-linear hyper-planes for classification of the data. The demerit of SVM approach is its ability to classify only fixed length data for which it cannot be used for the classification of data having variable length [12].

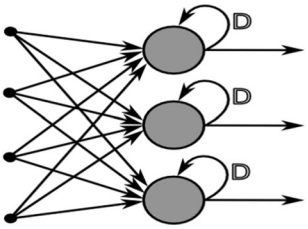


Fig. 4 A simple RNN

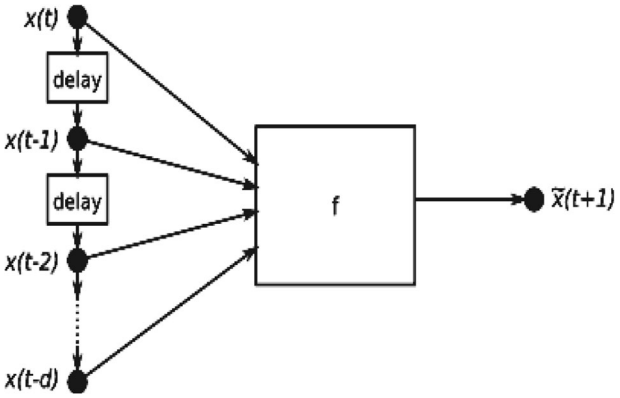
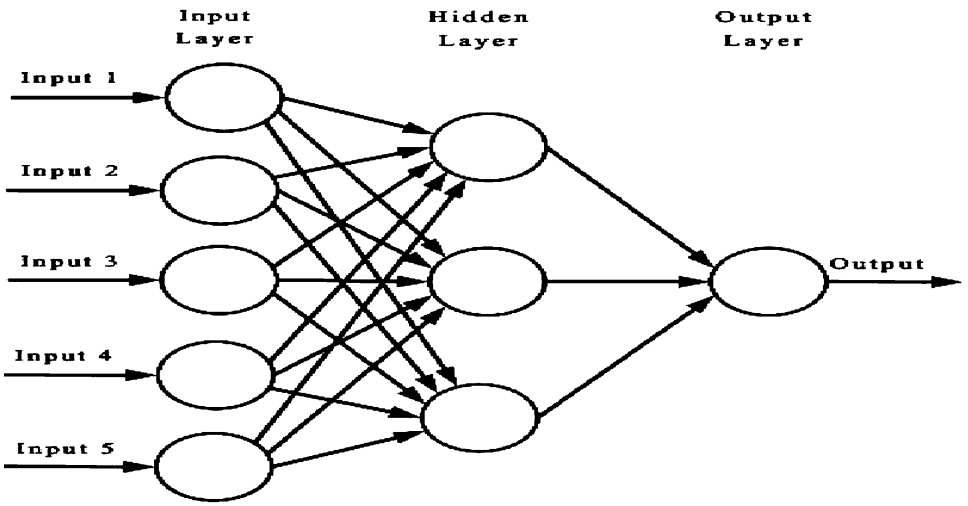


Fig. 5 A TDNN unit

2.3 Emotion recognition basics

Emotion recognition is the process of extracting information about the emotional state of the individual. Linguistics have defined some emotions that are commonly encountered in our life. Many researchers have proposed some small sets of emotions called primary emotions [13]. The block diagram of a typical emotion recognition system from speech signal is shown in Fig. 6.

Fig. 3 A Schematic diagram of FFNN



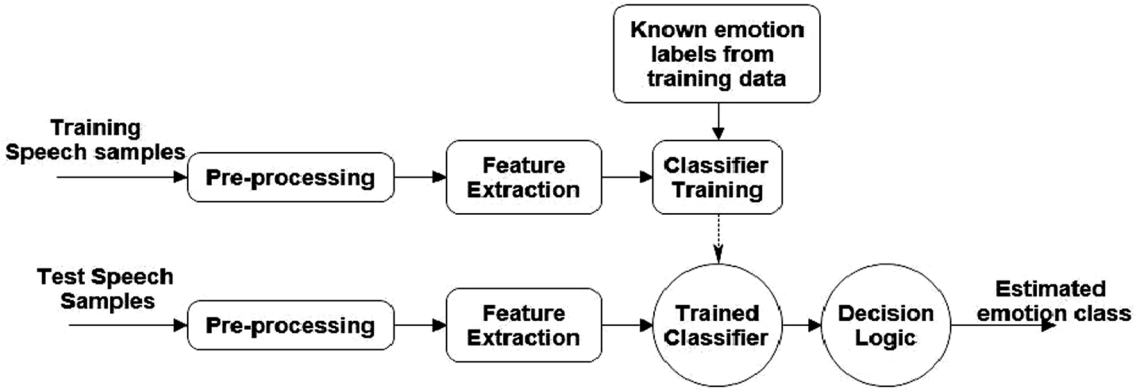


Fig. 6 Block diagram of emotion recognition system from speech

2.4 Speaker recognition basics

Speaker recognition is the process of recognition of the person speaking on the basis of the uttered speech. The application of speaker recognition technology has been continually growing in various fields of application such as forensic applications, dynamic signatures, gait, keystroke recognition, data encryption purpose, user validation in contact centers, etc. [2].

The speaker recognition system makes a claim to identify the speaker based on the trained model and matching the characteristics of the given speech. Speaker recognition comprises of two fundamental divisions namely Speaker verification and Speaker identification.

2.5 Assamese language: certain relevant considerations

Assamese, which is the anglicized form of Asamiya, is a major language of North East India. Assamese is the eastern most Indo-Aryan language which is found in the entire north-eastern part of India. The language has developed from Magadhi Prakrit which is the eastern branch of Apabhramas that followed Prakrit. The Assamese language is the eastern-most member if the Indo-European family tree which can be traced back to the history of very early times [14, 15].

Assamese being rich in ethnographic contents is also dialectically diverse with four major dialects as described by renowned linguists of the modern days. The four major dialects are Goalparia, Kamrupi, Central and Easter n dialects. Recently some works have reported design of speech based systems in Assamese [16].

3 Soft computation based spectral and temporal models

The work done so far is summarized into the following sections.

3.1 i-Vector based emotion recognition in Assamese speech

Here, we report the design of an emotion recognition system in Assamese language exploiting the ability of the Recurrent Neural Network (RNN) to track the temporal variations in the speech sample. We have designed i-vector and Mel Frequency Cepstral Coefficients delta (MFCC-delta) features and report the comparative performance, for the recognition of various moods, derived from the RNN and Distributed Time Delay Neural Network (DTDNN) classifiers.

3.2 Speaker recognition with normal and telephonic Assamese speech using i-vector and learning based classifiers

Here, we report the design of a speaker recognition system using normal and telephonic Assamese speech for our case study. In our work, we have implemented i-vectors as features to generate an optimal feature set and have used the Feed Forward Neural Network for the recognition purpose which gives a fairly high recognition rate.

3.3 Spectral features and learning aided mood and dialect recognition using telephonic Assamese speech

Dialects and moods recognition has become an important topic in HCI as computers have become integral part of our lives and there is a need for more natural communication between human and machines. In our work, we have reported the design of a mood and dialect recognition system for Assamese telephonic speech using Fourier parameters using RNN, DTDNN and SVM as classifiers.

4 Results

In this section we have included the results obtained from our experimental works as mentioned in the previous section.

4.1 Results obtained for i-vector based emotion recognition in Assamese speech

Two approaches of the conventional ANN techniques namely- RNN and TDNN are used for the recognition purpose. We have collected 1440 samples for both training and testing.

For the recognition of the Assamese speech using a composite features set of i-vectors and MFCC-delta features, the RNN gives a higher recognition rate of 86.55% and the DTDNN gives a recognition rate of 80% using the SCG algorithm but the computational time in case of the DTDNN is marginally higher.

In comparison to the composite feature set, the MFCC-delta feature set gives an accuracy of 83.65 and 74% for RNN and DTDNN using SCG algorithm respectively. For MFCC-delta also, the computational time for DTDNN is higher than that of RNN using the SCG algorithm.

Similarly, the recognition accuracy of RNN and DTDNN using the RP algorithm for the composite feature set turns out to be 78.1 and 77.35% respectively and that for MFCC-delta features are 83.65 and 72.8% respectively. Here, the computational time for RNN in case of the composite feature set is higher than that of the DTDNN. For MFCC-delta features, the computational time for RNN is higher than that of DTDNN using the RP training algorithm. From the experimental results, it is seen that the RNN gives a better recognition accuracy of 86.55% than the DTDNN in case of the composite feature set. The recognition accuracy for MFCC-delta features is also found to be higher in case of RNN. However, the computational time taken by the RNN algorithm is marginally higher than that taken by DTDNN.

4.2 Results obtained for speaker recognition with normal and telephonic Assamese speech using i-vector and learning based classifier

The raw speech database used for our experimental findings is collected from the native speakers of the state of Assam. The speech signal consists of a fixed length long Assamese sentence. We have considered 15 speakers of both male and female gender for collecting the samples for our proposed text-dependent speaker recognition model. In our work, we have concentrated on use of i-vectors as our feature parameters which is calculated from the total variability subspace. The experimental findings of our proposed text-dependent speaker recognition system involve trained system. We have used MLFNN classifier for the recognition purpose because of its robustness to noise and non-linear characteristics. The proposed model gives a maximum recognition rate of 98.8% with a Total Variability matrix of dimension 50 and 256 Gaussian Mixture components and 30.84 s computational time. The MLFNN is trained using the Scaled Conjugate Gradient (SCG) training algorithm and low dimensional i-vector. The MLFNN consist of 3 hidden layers, one input layer and one output layer. The input layer consists of 14 neurons, the 3 hidden layers consists of 100, 100 and 50 neurons respectively and the output layer consists of 15 neurons which is equal to the number of speakers.

4.3 Results obtained for spectral features and learning aided mood and dialect recognition using telephonic Assamese speech

The experimental work involves trained systems. Two techniques of ANN namely RNN and DTDNN and a multi-class SVM is used for the classification purpose. The training parameters has been set as per required. The recognition of moods and dialects using Assamese telephonic speech involves the use of Fourier features and MFCC feature vectors for training and testing the ANN models and multiclass SVM.

4.3.1 Recognition of moods

The experimental database consists of three moods viz. normal, loud and angry. The overall recognition rate of moods using Fourier features gives a performance rate of 81.7% using RNN as classifier and that of 78.3% using DTDNN as classifier. The algorithm used for the ANN models is Scaled Conjugate Gradient (SCG). The recognition rate obtained for mood recognition using a multi-class SVM is 68.33%. On the other hand, the MFCC features used for comparison of the performance of mood recognition gives an overall recognition rate of 92.5 and 86.7% for RNN and DTDNN respectively and that for multi-class SVM is 82.5%.

4.3.2 Recognition of dialects

The overall recognition rate for dialect recognition obtained by using Fourier parameters gives a performance rate of 95% for RNN and 92.5% for DTDNN and that obtained by using a multi-class SVM is 87.5%. The algorithm used for the ANN techniques is same as that used for the mood recognition purpose. For MFCC features the recognition rate obtained are 83.3, 82 and 80.4% for RNN, DTDNN and SVM respectively. The computational time is

basically dependent on the training. More variations in the training samples will lead to higher computational time. However, the testing is an almost instantaneous process.

5 Conclusion and future direction

The main idea behind speech recognition is to convert speech signal into a sequence of word by computer program which a machine can understand. As the main mode of communication, the goal of speech recognition is to facilitate the communication between human and machine more naturally and effectively. From the literature it has been found that the use of various spectral and temporal features provides significant performance improvements when compared to the common features. It has also been mentioned in some of the works that the combination of these features provides even better performance. The i-vector based framework also has become one of the state-of-art approaches for many speech processing applications. The i-vector approach provides an elegant way to reduce rough input utterances to a corresponding low dimensional vector representation while retaining the most important information embedded in the original input utterances. Over the recent years, i-vectors have been popularly used in various applications of speech processing. In a number of reported works related to telephone speech recognition, various soft-computational tools have been implemented for the recognition purpose.

Here, we have reported the design of a i-vector-MFCC-delta composite feature set based emotion recognition system by taking Assamese speech. Classifiers used are RNN and DTDNN which are configured for the purpose. A database has been created which contains three different mood variations recorded using male and female speakers. Despite of certain limitations, the accuracy and computational time together makes the proposed approach a suitable one for a dialectically rich language like Assamese.

Next, we have also designed a text-dependent speaker recognition model using Assamese speech. The work considers speakers with native orientation and the features used for our work is based on the i-vector framework which is considered to be one of the state-of-the-art techniques. We have used MLFNN which is non-linear and robust to noise. From the experimental results we have found that the recognition rate of the system is very high even though the speaker database contains session variations among the speakers. The proposed system gives a satisfactory recognition rate of 98.8% with a computational time of 30.84 s with Total variability matrix dimension of 50 and Gaussian components of 256, despite of certain limitations.

We have also proposed a learning based dialect and mood recognition system using Assamese telephonic speech. RNN, DTDNN and multi-class classifiers are configured for the experimental work. The database created is collected from the native speakers of the state including both male and female speakers.

From the experimental results obtained, we have found that the RNN gives better recognition accuracy although the computational time taken by the RNN is marginally higher for the i-vector based emotion recognition task. We have also found that the linguistic diversity and phonetic richness of Assamese language is best captured by the i-vector MLFNN combination. The i-vector MLFNN combination gives a satisfactory performance in case of the text dependent speaker recognition task. In case of the Fourier features based mood and dialect recognition, the RNN out performs the other classifiers but the computational time taken by SVM is much lesser than that taken by other classifiers.

Despite of certain limitations and limited dataset, the proposed systems prove to be suitable for the recognition of emotions, speaker, moods and dialect of an ethnographically rich language like Assamese which will help in lessening the digital divide among the people of the region.

In our future work we are planning to design a Convolutional Neural Network (CNN) with feedback mechanism and use it as classifier for speech based applications. We have also thought of finding the Temporal attributes in Telephonic Conversation Recognition using Assamese speech.

Acknowledgement Funding was provided by Ministry of Electronics and Information Technology, Visvesvaraya PhD Scheme.

References

1. Kurain C (2014) A Survey on Speech Recognition in Indian Languages. Proc Int J Comput Sci Inf Technol (IJCSIT) 5(5):6169–6175
2. Beigi H (2011) Fundamentals of speaker recognition. Springer, New York
3. Zuo G, Liu W, Ruan X (2003) Telephone Speech Recognition Using Simulated Data from Clean Database. In: Proceedings of IEEE international conference on robotics, intelligent systems and signal processing, vol. 1, pp 49–53, Changsha, China
4. Venkateshwarlu RLK, Raviteja R, Rajeev R (2012) The performance evaluation of speech recognition by comparative approach. In: Karahoca A (ed) Advances in data mining knowledge discovery and applications. InTech. doi:10.5772/50640. Available from: http://www.intechopen.com/books/advances-in-data-mining-knowledge-discovery-and-applications/the-performance-evaluation-of-speech-recognition-by-comparative-approach
5. Awasthy N, Saini JP, Chauhan DS (2008) Spectral analysis of speech: a new technique. World Acad Sci Eng Technol 2(7):946–955

6. Sarma M, Sarma KK (2014) Phoneme-based speech segmentation using hybrid soft computing framework, vol 550. Springer, New York
7. Anusuya MA, Katti SK (2009) Speech recognition by machine: a review. Proc Int J Comput Sci Inf Secur (IJCSIS) 6(3):181–205
8. Ganapathy S, Thomas S, Hermansky H (2010) Robust spectro-temporal features based on autoregressive models of Hilbert Envelopes. In: Proceedings of the 2010 IEEE international conference on acoustics, speech and signal processing (ICASSP), Dallas, TX, pp 4286–4289, ISBN 978-1-4244-4296-6. doi:10.1109/ICASSP.2010.5495668
9. Boersma P (1993) Accurate Short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Proc Inst Phon Sci 17(1193):97–110
10. Dehak N, Kenny P, Dehak R, Dumouchel P, Oullet P (2011) Front-end Factor Analysis for Speaker Verification. Proc IEEE Trans Audio Speech Lang Process 19(4):788–798
11. Mozer MC (1989) A focused back-propagation algorithm for temporal pattern recognition. Complex Syst 3(4):349–381
12. Yadav KS, Mukhedkar MM (2013) Review on speech recognition. Proc Int J Sci Eng (IJSE) 1(2):61–70
13. Gunes H, Pantic M (2010) Automatic, dimensional and continuous emotion recognition. Int J Synth Emot 1(1):68–99
14. Goswami GC (1982) Structure of Assamese, 1st edn. Department of Publication, Gauhati University, Gauhati
15. Sharma M, Sarma KK (2015) Soft-computational techniques and spectrotemporal features for telephonic speech recognition: an overview and review of current state of the art. In: Bhattacharyya S, Banerjee P, Majumdar D, Dutta P (eds) Handbook pf research on advanced hybrid intelligent techniques and applications, chapter-006, Hersey, PA: Information Science Reference, pp 161–189
16. Sarma M, Sarma KK (2013) An ANN based approach to recognize initial phonemes of spoken words of Assamese language. J Appl Soft Comput 5(13):2281–2291