**Neural Network based Speech Recognition Model for low resourced language Sylheti**

**Gautam Chakraborty**

**Assistant Professor, Department of Computer Science, NERIM, Guwahati. emailid:gauchak2012@gmail.com**

***Abstract***:***As*** *a wide-ranging branch of computer science*, *artificial intelligence (AI) is becoming an important intelligent tool in our day-to-day activities. Speech recognition-based applications, a domain in AI, finds its wide popularity in the recent years. Many interactive speech-based applications are developed considering major languages like English, Japanese, Chinese, German, etc. involving latest technology in consecutive periods, but practical usages of these applications are still limited due to language barrier and/or socio-economic constraints. In recent times, researchers have been concentrating to design and build speech recognition model in various low resourced languages. Sylheti is one of such low resourced languages and its native speakers mostly reside in the Sylhet division of Bangladesh and partly in the southern part of Assam, and Tripura, India. This work contributes by designing speech recognition systems for the Sylheti language to recognize Sylheti isolated words by applying three variants of neural network classifiers. The performances of these recognition systems are evaluated with a newly constructed database and the observations are presented accordingly.*

*Keywords*: *Automatic Speech Recognition, Mel Frequency Cepstral Coefficient, Sylheti language, Low resourced Language, Feed-forward neural network, Recurrent Neural Network, Time Delay Neural Network.*

# **INTRODUCTION**

With the advancements of technology in last four decades, artificial intelligence stands itself an important intelligent tool in our day-to-day activities. Subsequently, speech recognition-based tools are becoming more and more popular in various practical applications in recent times. Over the past decades, an incredible amount of research has been observed on the use of machine learning in speech processing applications, especially in speech recognition.Speech, in its simple terms, acts as a key communication media among people. Each uttered word (speech) holds linguistic contents (vowel and consonant speech segments) specific to a human language. With the advances in machine learning capability, it has become more pertinent to use voice for man-machine interaction. Mostly used Isolated word-based speech recognition tools are functional in mobile telephony, interactive television, support systems for differently abled people, robotics, etc.

The fundamental theory of Automatic Speech Recognition (ASR) is that it maps a given speech signal into machine readable format and subsequently transforming it into desired outputs. In the context of pattern recognition problem, a speech recognition system compares a given test pattern with the training pattern of each of the speech classes for classification. Depending on the type of applications, ASR model can work with various patterns of recorded speech like isolated words, connected words or continuous speeches stored in small vocabulary to large vocabulary databases [1],[2],[3],[17],[18],[30],[36]. Three functional blocks work in an ASR system following a sequence [5]:

i) *Signal Pre-processing*: To extract only voiced parts by exercising an input speech signal through a series of signal analysis steps like analog-to-digital conversion, pre-emphasis filtering followed by windowing.

ii) *Feature Extraction:* To derive various features from each voiced part in the pre-processed signal. Some developed speech features in ASR systems include Linear Predictive Coding (LPC) coefficients [5],[6], Mel Frequency Cepstral Coefficients (MFCC) [4],[6],[7],[8],[10], short-time energy [6], i-vector [11], etc. Among these, MFCC features are mostly used in ASR systems because, the mel-frequency scale in MFCC coefficients is proportional to the logarithm value of the linear frequency below 1000 Hz due to which it closely emulates the human perception.

iii) *Classification:* To match the feature vector of an input speech signal into 1 out of N word classes of the considered vocabulary during testing. This process is carried out by a classification technique with the help of a developed model during learning. Some widely applied classifiers in ASR are Artificial Neural Network (ANN) [5],[10],[12],[13], Hidden Markov model (HMM) [14],[15], Dynamic Time Warping (DTW) [16],[17], Deep Neural Network (DNN) [9],[41],[45], etc. Due to its unique and specific characteristics like Non linearity, Robustness, Adaptability, ANN finds its wide popularity in designing ASR system [5],[6],[19],[20],[21],[22],[23],[24], [35],[53],[54] since 1990 even though DNN techniques witness more machine learning capability than other techniques.

Though remarkable progress of the ASR system in well-resourced human languages such as English, Russian, German, etc. has been witnessed, language barrier is found to be major factor which demands for ASR systems in "low resourced languages" [25],[29]. A low resourced or under-resourced language is one which has some challenging factors like the lack or absence of writing system, shortage of linguistic study, limited or unavailability of electronic speech resources, etc [25],[26]. A language database Ethnologue in its 25th edition presents a list of 7,151 living languages in the world [27] which includes both the well- and low- resourced languages. Researchers in recent years have investigated and reported speech recognition solutions for some of the low-resourced languages [28],[29],[51],[52].

Having the challenging factors of a low-resourced language as reported in [31],[32],[33],[34],[38], the Sylheti language might be considered a low resourced language. Sylheti is a member of the Indo-Aryan language family [32] and more than ten million speakers in Sylheti are counted globally [27]. It is spoken largely in the Sylhet Division of Bangladesh. A fraction of Sylheti speakers resides in the northern part of Tripura and the Barak Valley region of Assam, India. Sylheti is defined in its unique script ‘Sylheti Nagari’ and all the 32 alphabets of Sylheti comprising of 5 vowels and 27 consonants are inscribed in this script [34]. A phonetic level research identifies 5 vowel- and 17 consonant- phonemes in Sylheti [32]. Further, scientific experiments [32] report that Sylheti witnesses some characteristics like distinctive way of pronunciation, de-aspiration and deaffrication which are found specific to the language.

As a machine learning model, ANN can process a large collection of nonlinear information in the form of artificial neurons or nodes and its structure comprises of three layers: one input layer, one or more hidden layers and one output layer [24]. An artificial neuron, say Y, can produce the output  based on an activation function according to:

 (1)

Where  (2)

Here ,  and are the respective outputs of three neurons X1, X2 and X3 through communication links having weights ,  and  respectively as shown in Figure 1.

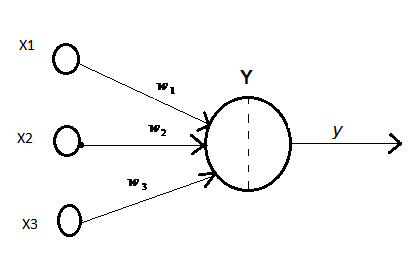
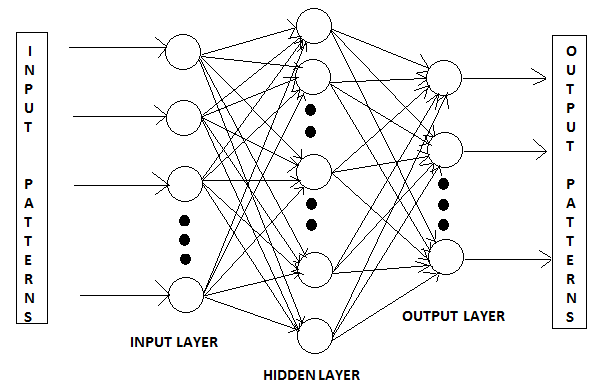
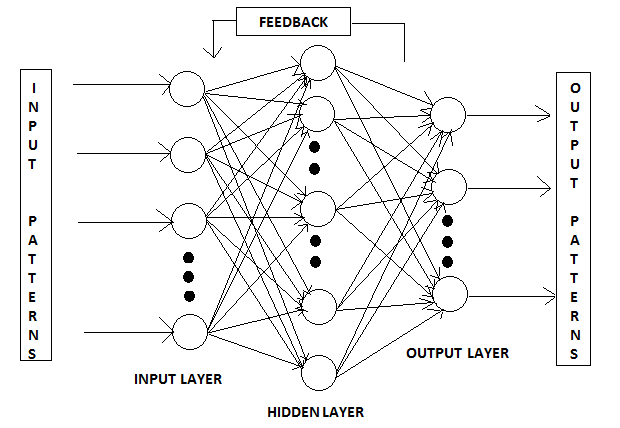


Figure 1. Model of an artificial neuron

The activation functions such as Sigmoid function, hyperbolic tangent function, Softmax function are used in neural network to decide whether a neuron activates or not. This outputmay be connected to one or more neurons in next layer. In general, based on the interconnections of neurons, two types of ANN structure namely feed-forward neural network (FFNN) and recurrent neural network (RNN) are employed in various machine learning applications. Skeletal structure of these two patterns is presented in Figure 2a and 2b showing one input -, one hidden- and one output layer. Due to having feedback concept, RNN can able to deal with time-varying dynamic inputs. It is to be noted that the stability of an ANN depends on the number of neurons in the hidden layer(s).



(a) Structure of FFNN



(b) Structure of RNN

Figure 2. Types of ANN

Time Delay Neural Network (TDNN) is another technique of ANN which is capable of capturing the relationship between the input and output data with variations in time. It is associated with a time delay in its input neuron and hidden neuron. TDNN provides a mapping between past and present values. A sketch of TDNN architecture is presented in figure 3.



Figure 3. TDNN architecture

As mentioned above, a remarkable amount of research investigates the use of variants of ANN algorithms to propose ASR systems in well- as well as low- resourced languages. B. P. Das and his associate proposed a neural network based ASR approach to recognize isolated English digits by feeding MFCC, LPC and short-time energy features for testing [6]. In [8], N. Seman et al. designed a learning model for isolated Malay words where MFCC features were trained and tested, deriving a recognition rate of 84.73%. In a work while proposing ASR system for Russian language, I. Kipyatkova et al. employed RNN technique which derives a high learning accuracy [44]. M. Oprea and his associate proposed an ASR system for Romanian language by using neural network classifier [12]. In a well-resourced language Japanese, S. Furui introduced a speaker-independent ASR system to learn Japanese cities [40]. For Turkish language, three ASR systems were designed by applying three variations of neural networks. In this work, MFCC features were investigated which resulted learning accuracy rates in the range 98.12% to 100%. In another work, authors addressed three ASR systems to recognize isolated digits in Assamese by using LPC features [5]. These systems employed FFNN, RNN and Cooperative Heterogenous ANN (CHANN) techniques for classification. For Bangla isolated words to recognize, an ASR model was proposed in [7] by investigating MFCC features of the preprocessed speech. In this experiment, authors used a semantic time delay neural network. From these recent works, it is observed that researchers are still applying various ANN techniques to design ASR systems in diverse well- and low- resourced languages [6],[42],[43],[44] due to their attractive characteristics.

This chapter mainly concentrates on the experimental work for designing ASR systems for the Sylheti language. In order to carry out the experiments, a speech dataset for the Sylheti language is constructed by considering commonly used isolated words in Sylheti. Experimental results are presented accordingly followed by discussions.

# **Data set Preparation**

A speech database is a collection of utterances for a human language, and it is an integral and vital resource for building a speech recognizer. Each utterance is having the information about phonemes which describe the basic units of speech sound [46]. Phonemic status of a sound varies across languages. Moreover, the count of phonemes in one language may differ from another language. As addressed in [32], Sylheti contains some unique phonemes which are not found in Bangla - or in globally used English language. This scientific work [32] also proves the nature of de-aspiration, spirantization and deaffrication features in Sylheti phonemes. Sylheti language contains a total of 22 phonemes as shown in Table 1, out of which 5 act as vowel and 17 as consonant [32]. On the other hand, Bangla contains 37 phonemes (7 are vowel and 30 are consonant phonemes) [46]. Five vowel phonemes in Sylheti (/i/, /e/, /a/, /u/ and /ǝ/) are similar to Bangla. The two other vowel phonemes of Bangla, /o/ and /æ/, are merged with the vowel phonemes /u/ and /e/ respectively in Sylheti due to restructuring in articulation [32]. Further, out of the 17 consonant phonemes in Sylheti, 13 phonemes (/b/, /t/, /ɡ/, /m/, /n/, /ʃ/, /s/, /h/, /r/, /l/, /ŋ/, /t̪/, /d̪/) also exist in Bangla [46]. The 4 consonant phonemes /z/, /x/, /ɖ/ and /Φ/ are unique in Sylheti language.

Table 1. Sylheti Phonemes

|  |  |
| --- | --- |
| Vowel phonemes | Consonant phonemes |
| /i/, /e/, /a/, /u/, /ɔ/ | /b/, /t/, /ɡ/, /m/, /n/, /ʃ/, /s/, /h/, /r/, /l/, /ŋ/, /t̪/, /d̪/, /z/, /x/, /ɖ/, /Φ/ |

On other hand, English language has a total of 12 vowel phonemes [46] out of which the 5 vowel phonemes in Sylheti are common to English. The remaining 7 vowel phonemes in English (/ə/, /æ/, /I/, /ɒ/, /ɜ/, /ʌ/, /ʊ/) are unique to the language. Further, 12 consonant phonemes /b/, /t/, /ɡ/, /m/, /n/, /ʃ/, /s/, /h/, /r/, /l/, /ŋ/, /z/ in Sylheti [32] also exist in English [46] and the other 5 consonants phonemes in Sylheti are specific for the language. From this perspective, it can be stated that there is enough scope to study the Sylheti language from the linguistic point of view [32],[38]. Further, it is stated that the technological study can also be carried out on Sylheti to explore and investigate more features.

In constructing the speech database in Sylheti, 30 mono syllabic phonetically rich Sylheti words are considered. Of these, 10 are the utterances of the digits 0-9 in Sylheti and remaining 20 are other Sylheti words. Table 2 lists these selected isolated words in Sylheti with the meaning in English of each word and the phonemes present (in bold letters) in the word.

Table 2: Isolated Sylheti words in the proposed database along with phonemes present in the words

|  |  |  |  |
| --- | --- | --- | --- |
| *Sylheti word* | *Meaning in English* | *Sylheti word* | *Meaning in English* |
| [s**ui**njɔ] | Zero | [**e**x] | One |
| [d̪**u**i] | Two | [t̪i**n**] | Three |
| [s**a**ir] | Four | [**Ф**as**]** | Five |
| [s**ɔ**y] | Six | [**s**at̪] | Seven |
| [a**ʈ**] | Eight | [nɔy] | Nine |
| [**d̪**an] | Donate | [**d̪**an] | Paddy |
| [**p**ua] | Boy | [**p**uri] | Girl |
| [**d̪**ud̪] | Milk | [**b**ari] | Home |
| [**p**ul] | Flower | [**b**ari] | Heavy |
| [**b**ala] | Good | [**b**ala] | Bracelet |
| [j**a**mai] | Husband | [**b**ɔu] | Wife |
| [ba**t̪**] | Boiled Rice | [ba**t̪**] | Arthritis |
| [ma**t̪**a] | Head | [mu**x**] | Face |
| [**ɡ**a] | Body | [**ɡ**a] | Wound |
| [**ɡ**ai] | Stroke | [**ɡ**ai] | Cow |

Before recording the utterances, the following hardware and software setup are used:

* iBall Rocky unidirectional microphone (frequency range from 20Hz to 20KHz)
* Intel Core i3 processor and 6 GB RAM
* Windows 10 Operating system
* PRAAT (version praat5367\_win64) as voice recording software

Further the following parameters are fixed during acquiring of speech:

* 16000 Hz Sampling rate
* Mono Channel mode
* Noise-free closed room environment
* WAV format with 16-bit PCM encoding
* Distance of microphone from speaker's mouth: 10-12 cm

The ages of the 10 (8 male and 2 female) contributing speakers are kept in the range from 25 to 70 years in order to obtain variations in speech features [49]. Six speakers in the range 25 to 45 years are graduates wherein the remaining 4 are in the age group 46 to 70 years having undergraduate degree. In addition to Sylheti, speakers can also speak Bengali, English and Hindi. The samples are recorded and stored according to:

***speakernumber\_age\_gender\_utteredword\_utterancenumber.wav***

Two processes namely speech acquisition and labeling [47] are involved in the construction of a speech database. In acquisition stage, either read out speech or from spontaneous speech [47],[48] is considered. In this case, the read-out speech is chosen here. Further, all 10 Speakers from Sylheti speaking areas like Karimganj, Silchar and Guwahati of India are directed to read out each Sylheti words 10 times as shown in Table 2. Accordingly, the utterances are stored. Thus, this exercise derives a speech database of having 3000 speech samples of isolated Sylheti words (30 X 10 X 10). Total duration of recording for this speech database is approximately 5 hours. This is accomplished by using PRAAT software.

# PROPOSED ASR SYSTEM FOR SYLHETI LANGUAGE

Most of the ASR systems [4],[6],[7],[8],[39],[43] developed in recent times for "well-resourced" as well as "low-resourced" languages apply MFCC features of speech signal to train and test by using neural network classifiers. In the same pipeline, an ASR model for Sylheti is proposed to develop which employs MFCC features and ANN classifier as depicted in Figure 4. In this work, we consider to employ three variants of ANN classifiers in three individual experiments to derive three ASR systems for recognizing isolated words in Sylheti.

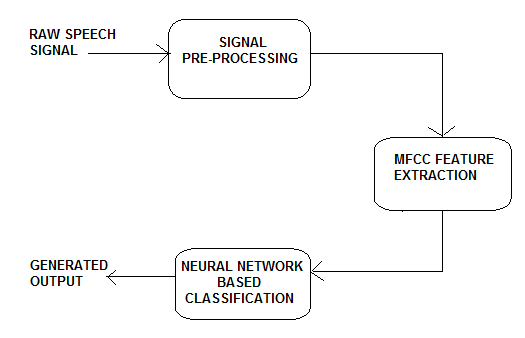


Figure 4. Architecture for ASR system employing MFCC features and ANN classifiers

Each block of the proposed ASR system as shown in Figure 4 is well-defined in the following:

**A. Signal Pre-processing**

In this signal pre-processing stage, a series of investigation like analog-to-digital (A/D) conversion, end point detection, pre-emphasis filtering and windowing is carried out. In A/D process, the acquired speech signal is sampled at 8 KHz and quantized with 16 bits/sample which produces a digital signal. Then, the voiced part is extracted from the digital signal by locating the beginning and end points in the speech. One popular method the zero-crossing rate, which is measured when the speech signal crosses the zero amplitude line by changeover from a positive to negative value or vice versa, is applied here. In pre-emphasis filter stage, a first-order high-pass finite impulse response (FIR) filter is employed to spectrally flatten the signal. Author considers the FIR filter for pre-emphasis in this work [5].

Thereafter, voiced part is segmented into *frame*s in the size usually from 5 ms to 100 ms [5] due to its time varying nature where frames are considered stationary and hence, speech analysis is carried out on the frames. In this proposed system, a frame duration of 32 ms with an overlap of 10 ms is considered. Finally, the windowing operation minimizes the spectral discontinuities at the end parts of each frame. The windowing operation is executed as [5]:

 ;  (3)

where, N is the number of frames in a speech sample, is a frame, is the window function and  is the windowed version of . This experiment employs the Hamming window in the voiced part as people usually apply Hamming window in speech analysis [4],[5],[35]. The coefficients of the Hamming window are computed according to:

; (4)

**B. MFCC feature extraction**

Feature refers to a set of representative numeric values of a speech sample that uniquely characterize the sample. Here, windowed version of each frame in a speech sample is considered independently to figure out a feature set for the frame. The feature sets of all the frames of the sample are then aggregated to figure out the features for the speech signal.

ASR systems reported for different languages usually derive MFCC coefficients as features due to their high similarity with human hearing system [4],[5],[6],[8]. Considering this, the proposed ASR systems for Sylheti language also use a set of MFCC coefficients as the features for the speech sample. MFCC features are extracted from the windowed version of speech frame passing through four stages as presented in Figure 5.

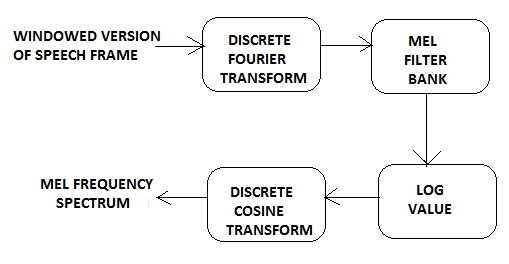


Figure 5. Computation of MFCC coefficients

In the first block, the discrete Fourier transform (DFT) coefficients is computed by employing the fast Fourier transform (FFT) to derive the amplitude spectrum. The mel filter bank in the second phase is performed to convert the frequency scale to the mel-scale according to:

 (5)

where  is the mel frequency corresponding to the linear frequency . Finally, log value is computed from the  and discrete cosine transform (DCT) is applied to it to obtain the magnitudes of the resulting spectrum [4]. As the first 12 to 13 MFCC coefficients contain maximum information of a speech signal [37], we apply the first 13 MFCC coefficients of a frame as features to represent the frame in this work.

**C. Classification by three Neural network models**

Due to its inherent characteristics as described above, neural network classifiers are still in use in designing speech recognition tools. The present work proposes to employ three variants of ANN classifiers: FFNN, RNN and TDNN techniques separately for classification in designing ASR systems for Sylheti language. The objective of ANN classifier is to segregate an input speech and put in a class by measuring its similarity with a pattern derived in training. Each of the neural networks is designed with one input layer, one hidden layer and one output layer. Count in input layer neurons is fixed as 13 as the 13 coefficients of MFCC feature set are used to represent an utterance. The output layer uses 30 neurons to represent 30 different classes. The selection of the exact number of neurons in the hidden layer is a challenging issue. Too few neurons in hidden layer may result underfitting wherein many numbers of hidden neurons cause overfitting [24],[50]. There are three rule-of-thumbs to select the exact number of hidden neurons [50]:

* It should be in between the input and output layer sizes.
* It is smaller than double of the input layer size.
* It may be the sum of the output layer size and 2/3 of the input layer size.

Still these rules may not work properly to find optimum hidden layer neurons. Hence, trial and error approach is generally adopted to find the optimum network architecture [50]. The present work decides to fix 46 number of neurons in hidden layer empirically as 46 neurons in the hidden layer demonstrate the optimal performance for all the network models.

Moreover, the non-linear activation (transfer) functions logsigmoid and tansigmoid are used in this work for the output layer and hidden layer respectively. The basic reasons of using sigmoid function are its smoothness, continuity and positive derivation. The logsigmoid function in the output layer produces the network outputs in the interval [0,1] i.e. output of one class is closer to 1 if the word is detected and 0 otherwise. Again, in tansigmoid function, the output is zero centrical in between -1 to 1 and hence optimization is easier. Further, the scaled conjugate gradient back-propagation procedure is used to train the networks due to its better learning speed. Most of the works applied this training algorithm due to this.

The following section presents the experimental setup and results of the proposed ASR systems.

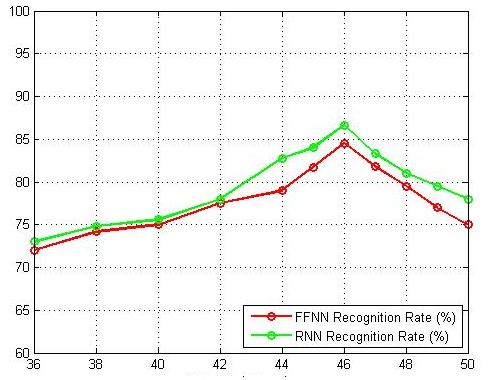
# **EXPERIMENTAL RESULT AND ANALYSIS**

This work performs three sets of experiments relating to the above-said three ASR systems for isolated Sylheti words. In the first experiment, the FFNN based ASR model is trained and tested. Accordingly, second set and third set deal with the RNN based and TDNN based techniques respectively. The following parameters are considered during experimentations:

1. Features: The set of 13 MFCC-based features for each utterance as presented.
2. Classifiers: FFNN, RNN and TDNN types individually.
3. Activation functions: tansigmoid for hidden layer and logsigmoid for output layer.
4. Training and testing datasets: Out of the 3000 utterances as stored in the Sylheti speech database, 1500 utterances comprising of 50 utterances of each word are considered for training the networks and other 1500 utterances are used for testing.
5. Convergence: Mean-squared error (MSE) value of 0.001 is fixed during training.
6. Performance measure: The performances of ASR systems are studied in terms of Percentage recognition rate (%RR), which is computed according to:

 (6)

The performances of the proposed ASR systems change when the number of neurons in the hidden layer is varied. As described above, to achieve optimal performances, the trial and error approach is adopted here by considering hidden layer neurons in the range 36 to 50. From three individual experiments carried out by FFNN, RNN and TDNN network models respectively, the hidden layer with 46 neurons derives the best performance for each case. Figure 6, for instance, presents plots of the observed performances of the ASR systems using the FFNN and RNN networks.



**%RR**

**Nodes in hidden layer**

Figure 6. Observed Performance plots with different number of neurons in the hidden layer

Either one of two conditions: a) the maximum number of epochs is reached, or b) performance is converged to the goal, are to be satisfied to stop the training of a neural network model. In these experiments, the first condition is satisfied. A convergence plot is often generated in the training phase to show the closeness of the network outputs to the target values. Figure 7 presents convergence plots for the proposed ASR systems in terms of MSE values. It is observed from this that the convergence of the TDNN based ASR system proves best than that of the FFNN and RNN based system. This is due to the time delay nature in TDNN, which tries to adjust the errors of outputs of the neurons during training.



Figure 7. Convergence plots for the proposed ASR systems

Observed performances of developed ASR systems are presented in Table 3.

Table 3. Performances of the ASR systems

|  |  |
| --- | --- |
| **ASR system using** | %RR |
| FFNN | 84.5 |
| RNN | 86.6 |
| TDNN | 91 |

Further, experiments are carried out by considering different training and testing Sylheti datasets (by grouping) for the proposed ASR systems. It is also observed here that the proposed systems perform more or less consistently within the range of 85-90%RR. This implies good robustness of the proposed systems to variations in datasets. From the recognition performances it can be summarized that the observed performances of the ASR systems for Sylheti are comparable to the performances of similar systems available for other languages [6],[7],[8],[37],[44] and hence are considered to be satisfactory.

# **CONCLUSION**

Speech Recognition using neural network techniques has been an area of research interest for last 4 to 5 decades, and consequently, many ASR systems developed for different languages around the globe are in practical use. By considering the "low-resourced" Sylheti language, this paper presents ASR models for this language. As no speech database for Sylheti in electronic form is available, a new speech database of Sylheti language has been proposed which can be used by researchers working in the domains of speech processing in Sylheti. This paper has presented three ASR systems for the Sylheti language to recognize isolated Sylheti words by applying three variants of neural network classifiers, FFNN, RNN and TDNN. It is observed that the overall performance of ASR system using the TDNN network (recognition rate:91%) shows best than that of the FFNN and RNN based ASR systems. To obtain 100% learning accuracy, more experiments are to be carried out by employing advanced machine learning techniques for speech processing in Sylheti language.

**REFERENCES**

[1] C. Kurian, "A Survey on Speech Recognition in Indian Languages", *International Journal of Computer Science and Information Technologies,* vol. 5, no. 5, 2014, pp. 6169-6175*.*

[2] R. Matarneh, S. Maksymova, V. V. Lyashenko and N. V. Belova, "Speech recognition systems: A comparative Review", *IOSR Journal of Computer Engineering*, vol. 19, no. 5, 2017, pp. 71-79.

[3] S. K.Gaikwad, B.W.Gawali and P. Yannawar, "A Review on Speech Recognition Technique", *International Journal of Computer Applications,* vol. 10, no.3, Nov. 2010, pp. 16-24.

[4] G. Dede and M. H. Sazli, "Speech recognition with artificial neural networks", *Elsevier journal of Digital Signal Processing*, vol.20, no. 3, May, 2010, pp.763-768.

[5] M. Sarma, K. Dutta, and K. K. Sarma, "Assamese Numeral Corpus for speech recognition using Cooperative ANN architecture", *International Journal of Computer, Electrical, Automation, Control and Information Engineering* vol.3, no.4,2009.

[6] B. P. Das and R. Parekh, "Recognition of isolated Words using Features based on LPC, MFCC, ZCR and STE, with Neural Network Classifiers", *International Journal of Modern Engineering Research (IJMER),* vol. 2, no. 3, May-June 2012, pp. 854-858*.*

[7] Y.A. Khan, S. M. Mostaq Hossain and M. M. Hoque, "Isolated Bangla word recognition and Speaker detection by Semantic Modular Time Delay Neural Network (MTDNN)", *18th International conference on Computer and Information Technology,* Dhaka, Bangladesh , 21-23 Dec. 2015.

[8] N. Seman, Z. A. Bakar and N. A. Bakar, "Measuring the performance of Isolated Spoken Malay Speech Recognition using Multi-layer Neural Network", *International Conference on Science and social Research(CSSR 2010),* Kualalumpur, Malaysia*,* December,2010.

[9] A. Mohammed, G. E. Dahl, and G. Hinton, "Acoustic Modeling using Deep Belief Networks", *IEEE transactions on Audio, Speech and Language Processing,* vol.20, no.1, January 2012, pp. 14-22*.*

[10] M. K. Luka, I. A. Frank and G. Onwodi, "Neural Network Based Hausa Language Speech Recognition", *International Journal of Advanced Research in Artificial Intelligence*, vol. 1, no. 2, 2012, pp. 39-44.*.*

[11] A. Kanagasundaram, "Speaker Verification using I-vector Features", *a PhD thesis of Queensland University of Technology,* 2014*.*

# [12] M. Oprea and D. Schiopu, "An artificial neural network-based isolated word speech recognition system for the Romanian language", *16th International* *conference on* *System Theory, Control and Computing (ICSTCC),* 12-14 Oct.*,* Sinaia, Romania, 2012.

[13] K. R. Ghule and R. R. Deshmukh, "Automatic Speech Recognition of Marathi isolated words using Neural Network", *International Journal of Computer Science and Information Technologies,* vol.6(5)*,* 2015, pp. 4296-4298.

[14] M. K. Sarma, A. Gajurel, A. Pokhrel and B. Joshi, "HMM based isolated word Nepali speech recognition", *Proceedings of International conference on Machine learning and Cybernetics,* Ningbo, China*,*2017*.*

[15] S. S. Bharali and S. K. Kalita, "A comparative study of different features for isolated spoken word recognition using HMM with reference to Assamese language", *International Journal of Speech Technology, Springer ,*vol. 18, no. 4, 2015, pp. 673–684*.*

[16] S. Xihao and Y. Miyanaga, "Dynamic time warping for speech recognition with training part to reduce the computation", *International Symposium on Signals, Circuits and Systems ISSCS2013,* 11-12 July, 2013*.*

[17] B.W.Gawali, S. Gaikwad, P. Yannawar and S.C. Mehrotra, "Marathi isolated word recognition system using MFCC and DTW features", *ACEEE International Journal of Information Technology,* vol. 01, no. 01, Mar 2011*,* pp. 21-24.

[18] C. Madhu, A. George and  L. Mary, "Automatic language identification for seven Indian languages using higher level features",*IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, Kollam, India, 2017, pp. 1-6.

# [19] P. Swietojanski, "Learning Representations for Speech Recognition using Artificial Neural Network", *a doctoral thesis*, 2016.

[20] M. Borsky, "Robust recognition of strongly distorted speech", *a doctoral thesis*, 2016.

[21] S. G. Surampudi and Ritu Pal, "Speech Signal processing using Neural Networks", *IEEE International Advance Computing Conference (IACC 2015)*, Bangalore, India, 12-13 June 2015.

[22] A. Zaatri, N. Azzizi and F. L. Rahmani, "Voice Recognition Technology using Neural Networks", *Journal of New Technology and Materials,* vol. 5, no. 1, 2015, pp. 26-30.

[23] O.I. Abiodun, A Jantan, A.E.Omolara, K.V. Dada, N.A. Mohamed and H. Arshad, "State-of-the-art in artificial neural network applications: A survey", *Heliyon, an Elsevier Journal,* vol. 4, no. 11, 2018.

[24] L. Fausett, "Fundamentals of Neural Networks: Architecture, Algorithms and Applications", Prentice-Hall, Inc., New Jersey 1994.

[25] L. Besacier, E. Barnard, A. Karpov and T. Schultz, "Automatic Speech Recognition for Under-Resourced Languages: A Survey", *Speech Communication,* vol. 56, January, 2014, pp. 85-100.

[26] V. Berment, "Methods to computerise "little equipped" languages and group of languages", PhD. Thesis, J. Fourier University-Grenoble I, May 2004.

[27]"Ethnologue Languages of the World" https://www.ethnologue.com/statistics/status.

[28] H.B.Sailor, M.V.S. Krishna, D. Chhabra, A. T. Patil, M.R. Kamble and H.A. Patil, "DA-IICT/IIITV system for low resource speech recognition challenge 2018", *Interspeech 2018,*  2-6 September 2018, Hyderabad, pp. 3187-3191.

[29] M.A. Hasegawa-Johnson, P. Jyothi, D. McCloy, M. Mirbagheri, G. M. di Liberto, Amit Das, B. Ekin, C. Liu, V. Manohar, H. Tang, E. C. Lalor, N. F. Chen, P. Hager, T. Kekona, R. Sloan and A. K. C. Lee, "ASR for Under-Resourced Languages From Probabilistic Transcription", *IEEE/ACM transactions on Audio, Speech, and Language processing,* vol. 25, no. 1, January 2017.

[30] K. Kumar, R.K. Aggarwal and A. Jain, "A Hindi speech recognition system for connected words using HTK", *International Journal of Computational Systems Engineering,*  vol. 1, no. 1, 2012*,* pp. 25-32.

[31] A. Gope and S. Mahanta, "Lexical Tones in Sylheti", *4th International Symposium on Tonal Aspects of Languages,Nijmegen,Netherlands,*May 13-16,2014

[32] A.Gope and S. Mahanta, "An Acoustic Analysis of Sylheti Phonemes", *Proceedings of the 18th International Congress of Phonetic Sciences.* Glassgow, UK, 2015*.*

[33] A.Gope and S. Mahanta, "Perception of Lexical Tones in Sylheti", *Tonal Aspects of Languages 2016,* 24-27 May 2016, Newyork.

[34] D. M. Kane, "Puthi-Pora:'Melodic Reading' and its use in the Islamisation of Bengal", *Doctoral Dissertation*, University of London, 2008.

[35] G.Chakraborty, M.Sharma, N. Saikia and K. K.Sarma, "Recurrent Neural Network Based Approach To Recognise Isolated Digits In Sylheti Language Using MFCC Features", *Proceedings of International conference on telecommunication, power analysis and computing techniques(ICTPACT-2017),*Chennai, India*,* 6-8 April,2017.

[36] X. Lei, A. W. Senior, A. Gruenstein and J. Sorensen, "Accurate and Compact Large vocabulary speech recognition on mobile devices", *INTERSPEECH 2013*, Lyon, France, 25-29 August 2013, pp. 662-665.

[37] P. Sharma and A. Garg, "Feature Extraction and Recognition of Hindi Spoken Words using Neural Networks", *International Journal of Computer Applications,* vol. 142, no.7, May 2016*.,* pp. 12-17.

[38] A. Goswami, "Simplification of CC sequence of Loan words in Sylheti Bangla", *Language in India,* vol 13, no. 6 June,2013.

[39] M. MoneyKumar, E. Sherly, and W. M. Varghese, "Isolated Word Recognition system for Malayalam Using Machine Learning", *Proc. of the 12th Intl. Conference on Natural Language Processing*, Trivandrum, India. December 2015, pp. 158–165.

[40] S. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum", *IEEE transactions on Acoustic,Speech and Signal processing,* vol. 34 , no.1, 1986, pp. 52-59.

[41] A. Das, P. Jyothi, and M. H. Johnson, "Automatic Speech Recognition using Probabilistic transcriptions in Swahili, Amharic and Dinka", *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2016,* San Francisco, USA, 8-12 September, 2016, pp. 3524-3528.

[42] N. Zerari, S. Abdelhamid, H. Bouzgou, and C. Raymond, "Bidirectional deep architecture for Arabic speech recognition", *Open Comput. Sci., DE GRUYTER*, 2019, pp. 92–102.

[43] C. Xu, X. Wang and S. Wang, "Research on Chinese Digit Speech Recognition Based on Multi-weighted Neural Network", *IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application,* 2008, pp. 400-403.

[44] I. Kipyatkova and A. Karpov, "Recurrent Neural Network- based Language modeling for an Automatic Russian Speech Recognition System", *Proceeding of the Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT*)*,* 9-14 Nov. 2015, St. Petersburg, Russia.

[45] D. T. Toledano, M. P. Fernandez-Gallego and A. Lozano-Diez, "Multi-resolution speech analysis for automatic speech recognition using deep neural networks: Experiments on TIMIT", *PLoS ONE*, vol. 13, no. 10, October 10, 2018.

[46] B. Barman, "A contrastive analysis of English and Bangla phonemics", *The* *Dhaka University Journal of Linguistics,* vol. 2, no.4, August2009.

[47] W. Hong and P. Jin'gui, "An undergraduate Mandarin Speech Database for Speaker Recognition Research, Oriental COCOSDA International Conference on Speech Database and Assessments", Urumqi, China, 10-12 August, 2009.

[48] C. Kurian, "A Review on Speech Corpus Development for Automatic Speech Recognition in Indian Languages", *International Journal of Advanced Networking and Applications,* vol.6, no.6, 2015, pp. 2556-2558.

[49] B. Das, S. Mandal, P. Mitra and A. Basu, "Effect of aging on speech features and phoneme recognition: a study on Bengali voicing vowels", *International Journal of Speech Technology, Springer,*vol. 16, no. 1, March 2013*,* pp. 19-31.

[50] F. S. Panchal and M. Panchal, "Review on methods of selecting number of Hidden nodes in Artificial Neural Network", *International Journal of Computer Science and Mobile computing,* vol. 3, no. 11, Nov.2014*,* pp. 455 – 464.

[51] B. Deka, J. Chakraborty, A. Dey, S. Nath, P. Sarmah, S. R. Nirmala, and S. Vijaya, "Speech Corpora of Under Resourced Languages of North-East India", *Oriental COCOSDA 2018,* 7-8 May 2018, Miyazaki, Japan.

[52] Desai V. A and Dr. V. K. Thakar, "Neural Network based Gujarati Speech Recognition for Dataset Collected by in-ear microphone", *6th International Conference On Advances In Computing & Communications, ICAAC 2016,* 6-8 September,2016, Cochin, India, pp. 668-675.

[53] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. J. , “Phoneme recognition using time-delay neural networks”,

*IEEE Transactions on Acoustics, Speech, and Signal Processing,37*, 328–339, 1989.

[54] Peddinti, V., Povey, D., & Khudanpur, S., “A time delay neural network architecture for efficient modelling of long temporal contexts”, In *Sixteenth annual conference of the International Speech Communication Association*, 2015.