**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, are found popular in the recent years. People who are focusing on research domains like linguistics, engineering, psychology, etc. have been trying to correlate them with speech based scientific experiments in diverse human languages around the globe. 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 presents speech recognition systems for the Sylheti language to recognize isolated Sylheti 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, 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 is becoming 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, a tremendous amount of research has been observed on the use of machine learning for speech processing applications, especially speech recognition.Speech acts as a prime mode of communication among humans. Each uttered word (speech) in a human language contains linguistic contents (vowel and consonant speech segments) specific to the language. With the advances in machine and in the area of artificial intelligence, it has become more pertinent to use voice for man-machine interaction. By using of isolated words, speech recognition application is used 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 requirements of applications, ASR model works with various patterns of recorded speech like isolated words, connected words or continuous speeches stored in the databases varying from small vocabulary to large vocabulary [1],[2],[3],[17],[30],[38]. ASR systems are found in two forms: speaker dependent and speaker independent [2],[56]. Technologies like document preparation or retrieval, command and control, automated customer service, etc. use speaker independent speech recognizers while speaker dependent systems associate with the applications like interactive voice response system, computer game control, etc. An ASR system usually works passing through three stages [5]:

i) *Signal Pre-processing*: Extracting of 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:* Deriving various features from each voiced part in the pre-processed signal. Some popular speech features used 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 being proportional to the logarithm of the linear frequency below 1 kHz, it closely imitates the human perception.

iii) *Classification:* Matching 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 used 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 [9],[47],[51], etc. ANN is still a popular choice in designing ASR system [5],[6],[19],[20],[21],[22],[23],[36],[60], [61] because of its following characteristics [24]:

* Non linearity: ANN has the ability to acquire and build the non-linear and complex relationships between the inputs and outputs.
* Robustness: It can continue to process even if any component of the network fails due to its intrinsic parallel nature.
* Adaptability: It has the built-in self-learning capability which does not require reprogramming in a dynamic environment.

Although remarkable progress of the ASR system in well-resourced human languages such as English, Russian, German, Japanese, Mandarin, etc. has been witnessed due to noteworthy demand of the application of ASR system in various domains, 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 of writing system, absence or 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 reported speech recognition solutions for some of the low-resourced languages [28],[29],[58],[59].

Observing the challenging factors of a low-resourced language as reported in [31],[32],[33], [34], [41], the Sylheti language seems to be considered a low resourced language. Sylheti belongs to the Indo-Aryan language family [32] and more than ten million people speak Sylheti globally. It is spoken largely in the Sylhet Division of Bangladesh. A partial section of Sylheti speakers resides in the northern part of Tripura and the Barak Valley region of Assam, India. Sylheti is written in its unique Sylheti Nagari script and the alphabets are inscribed in [34]. Sylheti language is described by 32 alphabets including 5 vowels and 27 consonants. The study on phonetic level in Sylheti addresses 5 vowel- and 17 consonant- phonemes [32]. Scientific experiments [32] show that Sylheti has some unique characteristics such as distinctive way of pronunciation, de-aspiration and deaffrication, etc.

As a machine learning model, ANN can process a large collection of nonlinear information in the form of artificial neurons or nodes and it’s structure is organized in 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.

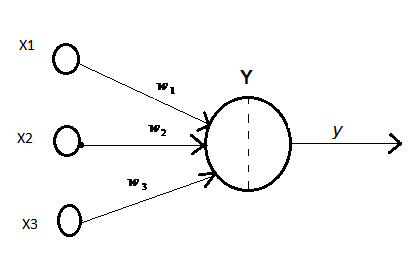
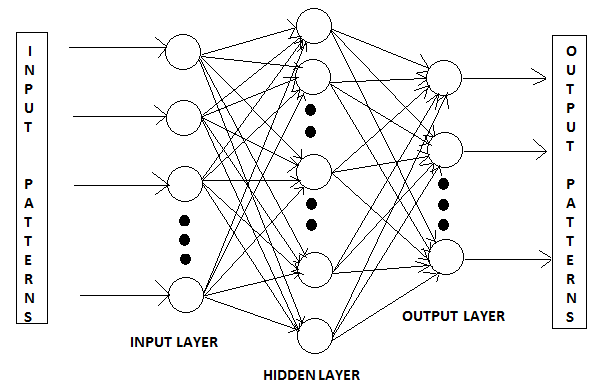


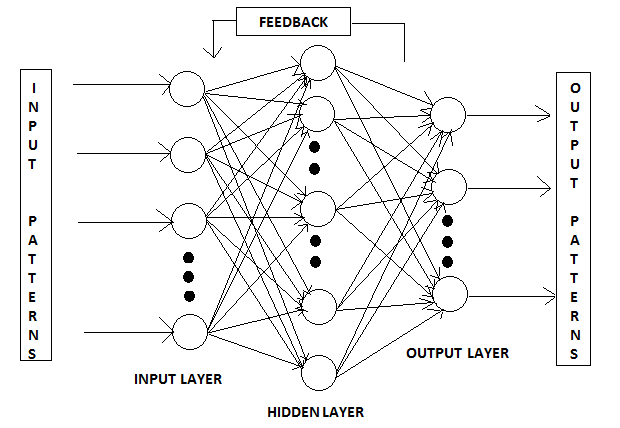
Figure 1. Model of an artificial neuron

The activation function used in neural network does the non-linear transformation to the input which decides whether a neuron activates or not. Sigmoid function, hyperbolic tangent function, Softmax function, etc [5] are some of the activation functions employed in neural network model. 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 and feedback or recurrent [24] neural network are employed in various machine learning applications. Skeletal structure of these two patterns are presented in Figure 2a and 2b showing one input -, one hidden- and one output layer. A feed-forward neural network (FFNN) broadcasts data from input to output through the hidden nodes wherein a recurrent neural network (RNN) uses the concept of feedback memory as shown in Figure 2(b). Due to the use of feedback, RNN has the ability 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) Architecture of FFNN



(b) Architecture of RNN

Figure 2. Types of ANN

Time Delay Neural Network (TDNN) is another technique of ANN which has the ability to capture 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. It provides a simple way to represent a mapping between past and present values. A sketch of TDNN architecture is presented in figure 3.



Figure 3. TDNN architecture

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.

# **LITERATURE REVIEW**

As a subject of interest over the last few decades, Speech recognition work started its journey by introducing a speech recognition tool in 1952 which could recognize ten digits (0-9) spoken by one speaker [35]. In continuation of this research, ASR systems for isolated words, connected words and continuous speeches in various languages are reported by employing different static and dynamic features and classification models [6],[18],[36],[37],[39],[44]. Although people employed hidden Markov model (HMM) in ASR as classifier during 1980s and 1990s, technology evolution has directed towards the use of ANN model significantly and also towards DNN in recent times.

A neural network based ASR approach was introduced for recognizing isolated English digits by B. P. Das and R. Parekh [6] which considered MFCC, LPC and short-time energy features for testing. An overall recognition accuracy of 85% was reported after carrying out the experiments. In [8], N. Seman et al. presented a recognition model for isolated Malay words where MFCC features were trained and tested deriving an average classification rate of 84.73%. An ASR model for Chinese digits was proposed by using MFCC features and neural network classifier where an average recognition rate of 97.4% was reported. I. Kipyatkova et al. introduced an RNN based ASR system for facilitating the Russian language [50]. The speech recognition models proposed in [48] for Arabic language employed two variants of neural network- multi-layer perceptron and Long Short-Term Memory (LSTM). For both the models, experiments reported recognition rates of around 95% in case of digits as well as words. To recognize isolated Turkish digits, a set of three ASR systems were proposed in [4] by applying different types of neural networks. Considering MFCC features of each speech signal, these systems presented recognition accuracies in the range from 98.12% to 100%. In [5], authors addressed three ASR systems to recognise isolated digits in Assamese by using LPC features. These systems used FFNN, RNN and Cooperative Heterogenous ANN (CHANN). An ASR model for recognizing isolated Bangla words was reported in [7] which derived MFCC features from preprocessed speech. In this experiment, authors used a semantic time delay neural network which resulted a recognition accuracy of 82.66%. J.T. Geiger et al. [39] in their research had applied RNN technique in a hybrid NN-HMM system architecture considering the medium-vocabulary speech dataset. P. Sarma and A. Garg proposed an ASR system in [40] to recognize Hindi words with a neural network classifier. MFCC and PLP features were used here which reported an average recognition accuracy of 79%. In [13], Marathi isolated words were considered to build an ASR system using neural network classifier which also presents an overall classification rate. Another ANN based ASR model for Malayalam isolated words was presented in [42] by using combined feature set comprising of MFCC, energy and zero crossing. This work reported a recognition accuracy of 96.4%. M. Oprea and D. Schiopu [12] proposed an ASR system to recognize Romanian language by using neural network classifier. S. Furui introduced a speaker-independent ASR system which facilitated to learn Japanese cities [44]. M. K. Luka et al. [10] employed ANN classifier to design an ASR system for Hausa language by extracting MFCC features. In [59], the authors introduced a neural network based ASR system for Gujarati words by using MFCC and real cepstral coefficients [59] and reported the comparison results. In recent times, people are employing DNN and its various variants in ASR system. Authors in [45] used a DNN classifier with MFCC features for English digit recognition. When tested with the database of English digits constructed by Texus Instrument, an average recognition accuracy of 86.06% resulted in this experiment. By considering German speech data, [43] presents an ASR system by employing convolutional neural network which derives a WER of 58.36% and a letter error rate of 22.78%. Authors in [46] use LSTM RNN to construct an ASR system to be able to learn Japanese speech. Another ASR system proposed in [47] considers three African languages Swahili, Amharic and Dinka which employs DNN classifier.

From the above discussion, it is observed that researchers are emphasizing to employ ANN technique and its variants in recent times to design ASR systems in some major languages [6],[48],[49],[50] due to their attractive characteristics as discussed in Section 1. ANN is also being popularly used for digit and isolated word recognition in low-resourced languages [4],[5],[7],[8],[10]. These ANN based ASR systems are reported to deliver good recognition rates.

# **Data set**

A speech database is a collection of utterances for a particular language and it is an integral and vital resource for building a speech recognizer. Each utterance in a human language contains the information about phonemes which describe the basic units of speech sound [52]. Phonemic status of a sound is dissimilar across languages. Moreover, the number of phonemes in one language varies from another language. As addressed in [32], Sylheti contains some unique phonemes which are not found in Bangla - or in mostly used English language. This scientific work also proves the nature of obstruent weakening which employs de-aspiration, spirantization and deaffrication in 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, out of which 7 are vowel and remaining 30 are consonant phonemes [52]. Five vowel phonemes /i/, /e/, /a/, /u/ and /ǝ/ in Sylheti 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 are also present in SCB [52] and they are /b/, /t/, /ɡ/, /m/, /n/, /ʃ/, /s/, /h/, /r/, /l/, /ŋ/, /t̪/, /d̪/. The other 4 consonant phonemes /z/, /x/, /ɖ/ and /Φ/ are unique in Sylheti language. The 17 consonant phonemes /p/, /ph/, /bh/, /th/, /d/, /dh/, / t̪h/, / d̪h/, /c/, /ch/, /k/, /kh/, /ɡh/, /w/, /j/, /Ɉ/, / Ɉh/ in SCB are absent in Sylheti.

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/ |

While comparing Sylheti with English, it is found that English contains a total of 12 vowel phonemes [52]. All the 5 vowel phonemes in Sylheti also exist in English. The remaining 7 vowel phonemes in English (/ə/, /æ/, /I/, /ɒ/, /ɜ/, /ʌ/, /ʊ/) are unique to the language. On the other hand, 12 consonant phonemes /b/, /t/, /ɡ/, /m/, /n/, /ʃ/, /s/, /h/, /r/, /l/, /ŋ/, /z/ in Sylheti [32] are also available in English [52]. Thus, Sylheti presents 5 language specific consonant phonemes /t̪/, /d̪/, /x/, /ɖ/ and /Φ/, whereas the English language has 12 specific consonant phonemes (/p/, /d/, /ɵ/, /k/, /f/, /v/, /Ʒ/, /tʃ/, /dƷ/, /w/, /j/, /ð/). From this observation it can be stated that there is enough scope to study the Sylheti language from the linguistic point of view [32],[41]. Further, it is stated that the technological study can also be carried out on Sylheti due to its unique speech features. The construction of a small vocabulary speech database in Sylheti considering isolated words is presented in the following.

In constructing a speech database in Sylheti, 30 mono syllabic phonetically rich Sylheti words are considered. Among 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 speech utterances, the following setup is used:

* Microphone: iBall Rocky unidirectional microphone (frequency range from 20Hz to 20KHz)
* Laptop: Intel Core i3 processor and 6 GB RAM, manufactured by Lenovo
* Operating system: Windows 10
* Voice recording software: PRAAT (version praat5367\_win64)

Further the following parameters are formulated during recording:

* Sampling rate for speech signal: 16 KHz
* Channel mode: Mono
* Environment: Noise-free closed room environment
* Encoding format: WAV with 16 bit PCM
* Distance of microphone from speaker's mouth: 10-12 cm

The ages of the participating speakers are in the range from 25 to 70 years. Six speakers are in the range 25 to 45 years with graduate degree. The other 4 speakers are in the age group 46 to 70 years and are undergraduates. The proportion of the speakers includes 8 male and rest 2 female speakers. These speakers do not have any history of speech disorders. Due to the variations in speech features with age and gender, speakers are selected in different age group [55]. Apart from Sylheti, speakers can also speak Bengali, English and Hindi. The samples are recorded and stored according to:

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

The construction of speech database primarily involves speech acquisition and labeling [53]. The acquisition of speech utterances may be from read out speech or from spontaneous speech [53],[54]. In this case, the read-out speech is chosen here. Further, 10 native Speakers of Sylheti speaking areas like Karimganj, Silchar and Guwahati of India are asked to read out each Sylheti words 10 times as shown in Table 2 and accordingly the utterances are recorded and stored. Thus, a total of 300 utterances are recorded for each speaker. Consequently, this exercise derives a speech database of having 3000 speech samples of isolated Sylheti words. Duration of recording for this speech database is approximately 5 hours. In the labeling process [53], the recorded utterances are verified carefully by listening the target words and presence of any irregular noise or quiet segments in the recorded samples are examined. For each recorded utterance, the voiced parts are extracted by selecting their beginning and end points, by eliminating the unwanted silent parts. This is accomplished by using PRAAT software.

IV. **PROPOSED ASR SYSTEM FOR SYLHETI LANGUAGE**

Most of the ASR systems [4],[6],[7],[8],[42],[49] 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.

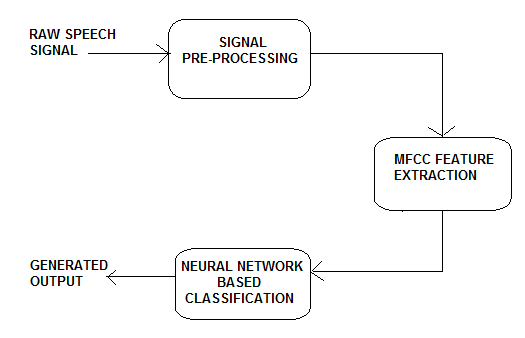


Figure 4. Architecture for ASR system employing MFCC features and ANN classifiers

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 analysis like analog-to-digital (A/D) conversion, end point detection, pre-emphasis filtering and windowing is carried out. In A/D conversion process, the input speech signal is sampled at 8 KHz and quantized with 16 bits/sample to derive a digital signal. The voiced part of the speech signal is extracted from the digital signal by locating the beginning and end points in the utterance. One popular method the zero-crossing rate, which is measured when the speech signal crosses the zero amplitude line by transition 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 applied to spectrally flatten the signal. Author considers the following FIR filter for pre-emphasis [5]:

 (3)

where, ** is the input signal (voiced part) to the pre-emphasis filter and is the output.

Thereafter, voiced part of the speech signal is divided into short segments called *frame*s usually from 5 ms to 100 ms [5] due to its time varying nature. Frames are considered stationary and hence, speech analysis is carried out on the frames. In this proposed system, a frame duration of 32ms with an overlap of 10ms is considered.

Finally, the windowing stage minimizes the spectral discontinuities at the boundaries of the frame. The windowing operation is executed as [5]:

 ;  (4)

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],[36]. The coefficients of the Hamming window are computed according to:

; (5)

**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 concatenated to derive the features for the input speech signal.

ASR systems for different languages use MFCC coefficients as features due to their high similarity with human hearing system [4],[5],[6],[8]. The mel frequency scale is approximately linear up to about 1000Hz in the frequency and approximates the sensitivity of the human ear. Considering this, the proposed ASR systems for Sylheti language also use a set of MFCC coefficients as the features for the speech sample. The block diagram for computing the MFCC coefficients at frame level is presented in Figure 5.

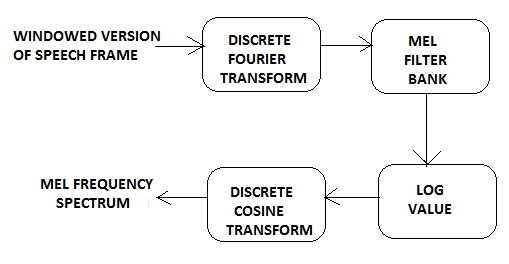


Figure 5. Computation of MFCC coefficients

The first block in MFCC computation finds the discrete Fourier transform (DFT) coefficients from the windowed version of an input speech frame deriving the amplitude spectrum. The DFT coefficients are usually obtained by employing the fast Fourier transform (FFT). The mel filter bank converts the frequency scale to the mel-scale, which is performed according to:

 (6)

where  is the mel frequency corresponding to the linear frequency . Finally, log value is computed from the output  and discrete cosine transform (DCT) is applied to it to obtain the magnitudes of the resulting spectrum [4]. Davis and Mermelstein introduced this methodology of the MFCC features extraction from the frames of a speech signal in 1980.

As the first 12 to 13 MFCC coefficients contain maximum information of a speech signal [40], we here consider the first 13 MFCC coefficients of a frame as features to represent the frame. Let represent the first 13 MFCC coefficients corresponding to the mel frequencies for the frame of an utterance. For a mel frequency , the mean of all the MFCC coefficients derived from the frames of an utterance is computed according to:

 (7)

The set of mean values acts as the features for the utterance.

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

Due to its inherent characteristics as described above, ANN classifiers are still in use in designing speech recognition model. The present work proposes to employ three variants of ANN classifiers for building ASR systems for Sylheti language. The role of the ANN classifier is to classify an input speech by measuring its similarity with a reference pattern derived through training phase. The proposed ASR systems for isolated Sylheti words employ FFNN, RNN and TDNN techniques separately for classification. Each of the neural networks is designed with one input layer, one hidden layer and one output layer. The number of neurons in input layer is taken as 13 as the 13 coefficients of MFCC feature set are used to represent an utterance. The output layer uses 30 neurons corresponding to the 30 different words to represent 30 different classes. The selection of an appropriate number of neurons in the hidden layer is challenging in the design of neural network. Using too few neurons in hidden layer results in underfitting whereas a large number of hidden neurons may cause overfitting [24],[57]. There are three rule-of-thumb methods to select the number of hidden neurons [57]:

* 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.

However, these rules may not work to find an optimum hidden layer. Therefore, trial and error approach with backward or forward selections is generally adopted to find the optimum network architecture [57]. In the present work, the number of neurons in hidden layer is decided empirically as discussed above. The observed performances for the FFNN, RNN and TDNN network models establish that 46 neurons in the hidden layer show the optimal performance.

The non-linear activation (transfer) functions logsigmoid and tansigmoid are used respectively 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 be 1 once the word is detected and 0 otherwise. Again, in tansigmoid function, it’s output is zero centered 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 [4],[5]. Many other authors have also used this training algorithm due to the above said advantage [6],[12]. As a supervised algorithm, this back-propagation method optimizes the weights of the neurons by using a loss/cost function [5] and produces faster convergence than other methods.

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

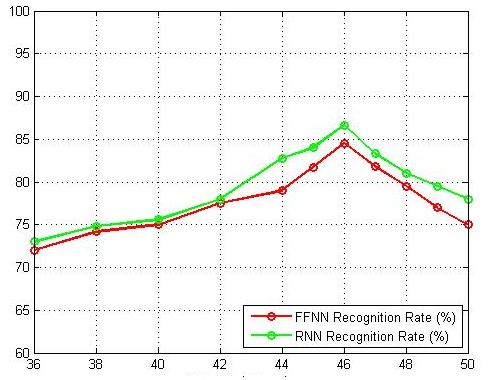
# **EXPERIMENTAL RESULT AND ANALYSIS**

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 ASR models. 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.
3. Activation functions: tansigmoid for hidden layer and logsigmoid for output layer.
4. Training and testing datasets: The database for Sylheti language presented in Section 3 has a total of 3000 utterances of 30 words, where each word is uttered 10 times by each speaker. Out of the 3000 utterances, 1500 utterances comprising of 50 utterances of each word are considered for training the networks. The other 1500 utterances are used for testing.
5. Convergence: Targeted mean-squared error (MSE) of 0.001 during training.
6. Performance measure: The performances of ASR systems are studied in terms of Percentage recognition rate (%RR), which is computed according to:

 (8)

The performances of the proposed ASR systems change when the number of neurons in the hidden layer is varied. To achieve superior performances, the trial and error approach is adopted in backward and forward directions by taking 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

A neural network model stops its training when any one of two conditions are met: a) the maximum number of epochs is reached, or b) performance is converged to the goal. In the presented work, 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. It presents the MSE values between the corresponding network outputs and targeted values [5],[36]. 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. Due to speech variability in age variation (which affect the performance of any ASR system) in the constructed Sylheti speech database, there is a minor deterioration of recognition results in the presented ASR systems. Thus, from the generated results it can be determined that the observed performances of the ASR systems for Sylheti presented above are comparable to the performances of similar systems available for other languages [6],[7],[8],[40],[50] 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 also 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.

**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] K.H.Davis, R. Biddulph and S. Balashek, "Automatic Recognition of Spoken Digits", *Journal of the Acoustic Soc. of America,* vol. 24, no. 6, 1952*,* pp. 627-642.

[36] 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.

[37] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for Spoken word recognition", *IEEE Trans. Acoustic, Speech Signal Processing,*  vol. 26 , no.1, Feb 1978 , pp. 43-49*.*

[38] 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.

[39] J.T.Geiger, Z.Zhang, F.Weninger, B. Schuller and G. Rigoli, "Robust speech recognition using long short term memory recurrent neural networks for hybrid acoustic modeling", *Conference of the International Speech Communication Association,* 14-18 September 2014, Singapore INTERSPEECH 2014

[40] 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.

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

[42] 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.

[43] J. Kunze, L. Kirsch, [I. Kurenkov](https://arxiv.org/search/cs?searchtype=author&query=Kurenkov%2C+I), A. Krug, [J. Johannsmeier](https://arxiv.org/search/cs?searchtype=author&query=Johannsmeier%2C+J), and S. Stober, "Transfer Learning for Speech Recognition on a Budget", *Proceeding of the 2nd workshop on Representation learning for NLP,* Vancouver, Canada, August 3, 2017, pp. 168-177.

[44] 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.

[45] D. Dhanashri and S.B. Dhonde, "Isolated word speech recognition system using Deep Neural Networks", *Proceedings of the International Conference on Data Engineering and Communication Technology,* vol. 1, 2017, pp. 9-17.

[46] T. Hori, C. Hori, S. Watanabe and J. R. Hershey, "Minimum word error Training of Long short-term memory Recurrent Neural Network Language models for Speech recognition", *41st* *IEEE International conference on Acoustic, Speech and Signal processing, Shanghai, China,* vol. 2016-May, 2016, pp. 5990-5994.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

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

[53] 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.

[54] 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.

[55] 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.

[56] M. Dua, R. K. Aggarwal and M. Biswas, "Performance evaluation of Hindi speech recognition system using optimized filterbanks", *Engineering Science and Technology, an International Journal,* Vol. 21, no. 3, June 2018, pp. 389-398.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.