International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 12, No.1 (Oct-2022)

https://dx.doi.org/10.12785/ijcds/120195

Assamese Speech-based Vocabulary Identification System using Convolutional Neural Network

Dipankar Dutta[1], Ridip Dev Choudhury[2] and Utpal Barman[3]

[1]*Department of Information Technology, Gauhati University, Assam, India*
[2]*Dept. of Computer Science, KKHSOU, Assam, India*
[3]*Department of Computer Sc. and Engineering, The Assam Kaziranga University, Jorhat, Assam, India*

Received 22 Jan. 2021, Revised 15 Jul. 2022, Accepted 23 Jul. 2022, Published 31 Oct. 2022

Abstract: Though the machine learning techniques were being used in Assamese Language Automatic Speech Recognition (ALASR) system over the last five years, but the applications of Convolutional Neural Network (CNN) are very limited in ALASR. The present study introduces a Convolutional Neural Network (CNN) enabled ALASR system for the Assamese language by collecting 35 isolated words in five different prime emotions as Normal, Angry, Happy, Sad, and Fear from five native male and five native female speakers. During the experiment, the Mel Frequency Cepstral Coefficient (MFCCs), Spectral Centroid (SC), zero-crossing rate (ZCR), Chroma Frequencies (CF), spectral roll-off (SRO), and intensity are extracted and analyzed using CNN with convolution layers and max-pooling layers. To examine the consequences, other model such as Feed Forward Artificial Neural Network (FFANN) is likewise applied in ALASR. The evaluating results of CNN with an accuracy of 98.4 % outperformed the ANN accuracy of 86.4 %.

Keywords: Automatic speech recognition, Mel Frequency Cepstral Co-efficient, Convolutional Neural Network, Feed Forward Neural Network,Pooling, Zero-crossing-rate.

1. INTRODUCTION

Over the last two decades, Automatic speech recognition (ASR) in the Assamese language is becoming very challenging research in the area of speech processing. According to the census report, the Assamese language is the mother tongue of 70 % population of Assam. It belongs to the Indo-Aryan family. Indo-Aryan group is a subgroup of Indo-European and Indo Iranian subgroups. The linguistic experts have proved that the Assamese language was evolved from the ancient Indian language Sanskrit. It grew out from Magadhi Prakrit and enriched its glossary from all Non-Aryan languages. The Assamese language is also influenced by other neighborhood languages of Assam. In northeast India, approximately 39 local languages are found among different tribes. The Assamese language is the mother tongue of approximately 3 crore population of Assam and also spoken by lots of people in the North East region of India such as Arunachal Pradesh, Meghalaya, Manipur, Nagaland, Mizoram, etc. Making the Assamese language interactive communication media between computers and users has now become the most necessary and challenging task. To speed up the process of interaction, lots of researchers are working on speech processing regarding the Assamese language.

It is observed that the accent of a language may vary with region. In the case of the Assamese language, 2 prime dialects have been reported by renowned Assamese literate Dr. Banikanta Kakati [1]. Nowadays 4 dialects in the Assamese language have been identified such as Kamrupi, Eastern, Goalparia, and Central Group. Assamese is considered a phonetically rich language with 8 vowel phonemes and 22 consonant phonemes

As of now making a system with natural language interaction is the utmost necessity. It is a very complex and challenging task. Our human brain can process lots of signals consciously or with an unconscious mind also. From a raw speech signal, a human being can classify emotions, the context of the speech, speaker, distance of the speaker, age approximation, tribe of the speaker, region of the speaker, and many more. But to interpret a speech signal through a machine and classify all these characteristics efficiently and effectively is very complex. The development of the ASR is mostly examined and analyzed for highly populated languages like English, Mandarin, Spanish, and Japanese, etc. In India, scenario is different because of the number of local languages spoken by different tribes and people from different location. The total number of languages in India is 1369, and 22 languages under scheduled languages and 99 languages under non-scheduled languages are accepted as official languages for different locations

E-mail address: dipankar.jrt1982@gmail.com, ridip_choudhury@rediffmail.com, utpalbelsor@gmail.com

by Indian Constitutions [2]. Traditionally ASR system uses different statistical models including Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) for speech processing. GMM model cannot be used to detect or identify data present on the boundary line.

With the time, the Artificial Neural Network (ANN) becomes a very strong replacement of these statistical models with the ability to model data in a boundary line. Along with the ANN, the past researchers used ANN with back propagation, ANN [3], [4], Recurrent Neural Network (RNN) [5], Long short-term memory (LSTM) [6], Deep Feed Forward Neural Network (DFFANN) [6], [7], [8], LSTM [9], CNN [10], [11], [12], [13], [14], [15], [16], [17] Convolutional Deep Belief Neural Network (CDBN) [18], DNN [19], [20], [21], [22], [23], [8], Deep Convolutional Neural Network (DCNN) [24], [25], [26], DCNN-DTPM [25] in ASR system.

Michael et al. [27] presented an emotional corpus of speech data from Assamese movies using Random Forest classifier, Gradient Boosting, and Linear Support Vector Machine (LSVM). In Assamese language, different emotional speeches were analyzed using statistical GMM architecture [28]. [29] presented an Assamese Language Automatic Speech Recognition (ALASR) system using ANN, and Distributed Time Delay Neural Network (DTDNN). A. Vergara et al. used 3 different time-frequency spectrogram of speech signals to classify phonemes using CNN and RNN [30]. In the year 2017, Satt et al. [31] presented an emotion detection system using CNN and LSTM, for EEMOCAP corpus. In [32], J. Camilo et al. proposed a Parkinson patient detection system from recorded voice using STFT and continuous wavelet transform and reported 89 % accuracy rate. To detect pathological voice disorder, spectrogram of pathological and normal speech is used as input signal to the DNN architecture proposed by Huiyi et al. [33]. Alhussein et al. [34], presented a transfer learning model for pathology voice detection and reported 97.5% accuracy. In [35], Han et al. used DNN model to recognize robust speech for the CHIME-2 dataset. In 2018, [36] DNN was successfully used by Cernak et al. to detect nasal phone from the LIBRI speech dataset.

In the research area of ASR, deep learning tools have used effectively for different languages across the globe, but its application is limited in Assamese language, especially in Assamese emotional speech recognition system. It has been observed that, for noisy speech, there is a limited number of study using Deep Learning technique in emotional speech recognition for the Assamese language. In this work, we have considered it as a challenge.

The current study forwards the following contributions towards an emotional ASR system:

- The study presents an ALASR system to recognize the speech from the raw Assamese speech signals in different emotions.
- The social networking application “Whatsapp” is used as a low cost speech recording techniques in the ALASR system.
- A CORPUS of Assamese emotional speech is designed and developed by recording the speech signals in a noisy environment through the WhatsApp.
- Two different neural network structures are used to classify the raw speech signals uttered in the Assamese language.

2. Material and Methods

A. About the Assamese Corpora

Five prime emotions considered for the proposed work are - Angry, Happy, Normal, Fear, and Sad. The following figures from 1 to 5 show the waveform and their spectral centroid (SC) and spectral roll-off (SRO) feature graph of the Assamese word “ahiba” in five selected emotions. The graph in the figures 1, 2, 3, 4, 5 shows that the feature values for the same word in different emotions are different. Assamese is the most popular language in North East India. The Assamese dataset used in this study is designed and developed in Gauhati University, Assam, India by collecting 35 isolated words such as - ’kio’, ’khai’ , ’maghat’, ’ranga’, ’saah’, ’bajaroloi’, ’kotha’, ’de’ , ’dhuli’, ’naam’, ’paani’, ’phul’, ’ubhati’, ’maa’, ’tomaar’, ’duwar’, ’si’, ’pharilai’, ’monot’, ’gol’, ’aami’, ’xuniba’, ’nalage’, ’jabaneki’, ’tumi’, ’moor’, ’kuneo’, ’karibi’, ’egilaas’, ’ahise’, ’koloi’, ’goisil’, ’kiman’, ’ahibi’, ’ahiba’ in five different prime emotions - Happy, Normal, Fear, Angry, and Sad. The raw speech signals for the data-set have recorded for each of the five emotions, from native male and female speakers of Assam. Each speaker is asked to simulate one sentence 10 times in one emotion.

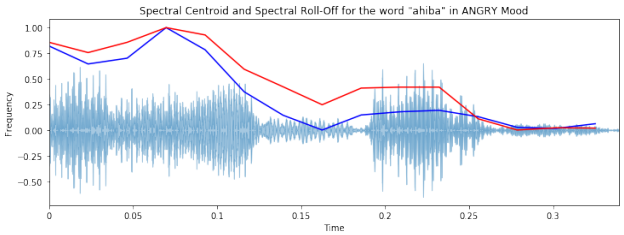

Figure 1. Spectral Roll-off and SC for angry-ahiba.wav

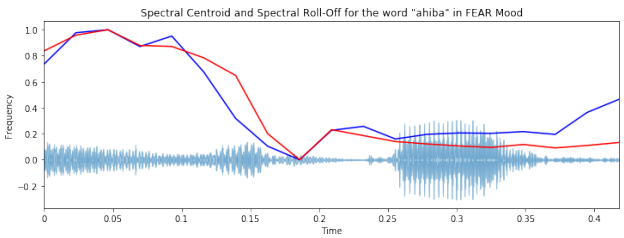

Figure 2. Spectral Roll-off and SC for fear-ahiba.wav

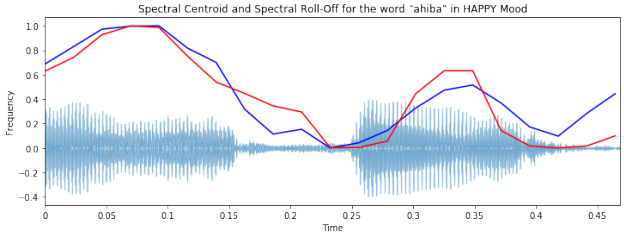

Figure 3. Spectral Roll-off and SC for happy-ahiba.wav

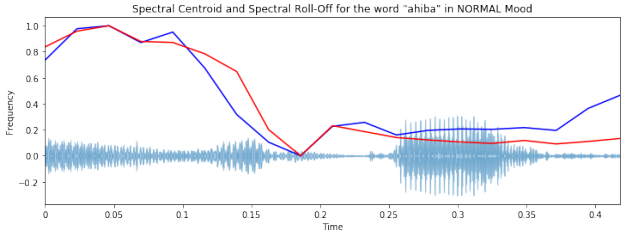

Figure 4. Spectral Roll-off and SC for normal-ahiba.wav

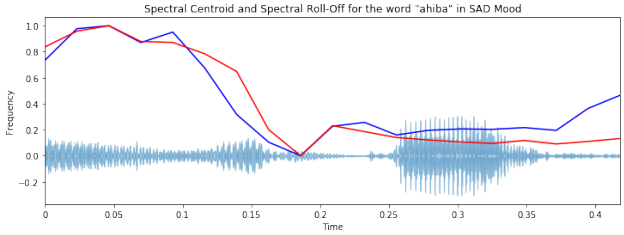

Figure 5. Spectral Roll-off and SC for sad-ahiba.wav

The recordings are done through the most popular mobile app "WhatsApp" without concerning about the surrounding environment. The recordings are with lots of unwanted background noise. To get the actual and real environmental feelings during experiments, recordings are kept in original stage for the experiment. Finally, the isolated words are separated manually through an application "PRAAT". The total number of raw speech data used in this experiment is 12250.

B. Assamese automatic speech recognition using CNN

CNN model is the mostly used and efficient deep learning tool for the last few years. It is efficiently used in speech processing. The figure 6 represents the applied CNN model architecture. The first convolution layer in the model is linked to a drop-out layer with a dropout rate of 0.5. The dropout is used to avoid the overfitting of the model during training [37]. The features detected in the convolutional layer are filtered in the pooling layer. Here, two convolutional layers are used. The amalgamation of convolution and the pooling layer is linked with the fully connected (FC) layer and the output layer or the classification layer of the model as shown in figure 6. The output layer makes the prediction for the input signal and gives the specified classification.

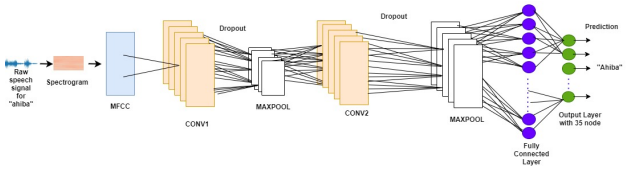

Figure 6. Architecture of the applied CNN model

- Feature Extraction

For the proposed work seven features have been extracted and analysed for each speech file. Spectral Centroid, Specral bandwidth, rollof, Zero crossing rate, Chroma stft., rmse and MFCC. In the following figure 7, it has shown a sample of .wav file for the vocabulary 'ahiba' in angry mood. We have used this .wav file to show other extracted features.

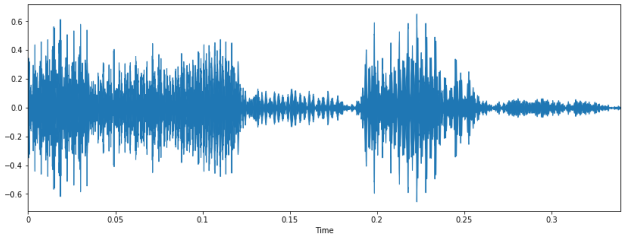

Figure 7. Sample Wav form of the speech file "ahiba-angry.wav"

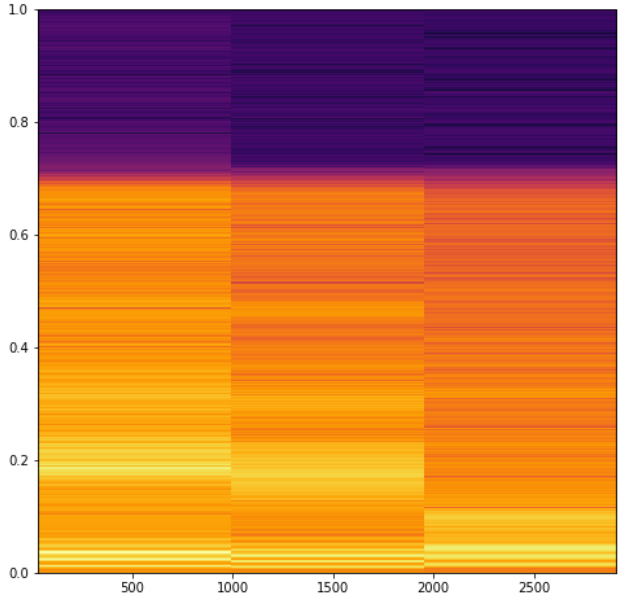

Figure 8. Spectrogram for the speech file "ahiba-angry.wav"

Mel Frequency Cepstral can be used to represent power spectrum and better representation of any sound. In the Mel

Scale, frequency bands are distributed uniformly. MFCC can be derived by applying Fourier transformation and discrete cosine transformation onto the Mel scale. The spectrum for the vocabulary 'ahiba' in angry mood is shown in the figure 8. The unit used to measure the center of mass of a spectrum can be defined as the Spectral Centroid. Digital audio signal and spectrum can be characterized by using the spectral centroid. Spectral roll-off restricts the frequency range outside from a specific range. High-pass or low-pass filters are used to roll off the frequencies beyond the range. The differences between the higher and lower frequency can be defined as the spectral bandwidth. In the figure 9, the spectral centroid and spectral roll-off is shown for the speech signal presented in figure 7. For each speech signal, 20 MFCCs are extracted to represent the overall shape of the spectral envelope as shown in the figure 10.

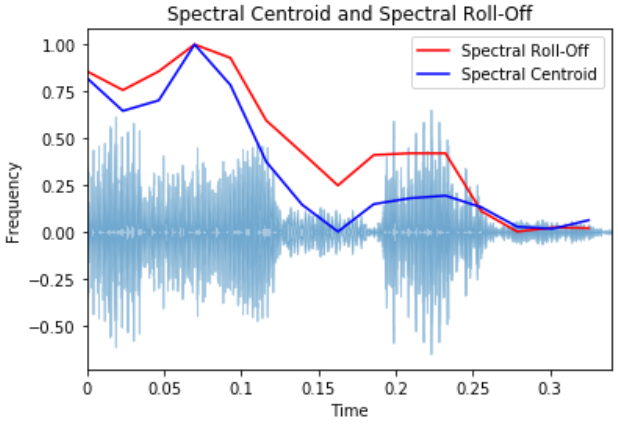

Figure 9. Spectral Centroid and Spectral Roll Off extracted from the file 'ahiba-angry.wav'

filename	chroma _stft	rmse	spectral _centroid	spectral _bandwidt	rolloff	zero_ crossing rate	mfcc1	mfcc2
ahiba_angry _01.wav	0.384458	0.095585	1316.778	1331.997	2191.418	0.040753	-229.929	147.8817
mfcc3	mfcc4	mfcc5	mfcc6	mfcc7	mfcc8	mfcc9	mfcc10	mfcc11
-42.6307	52.57816	-1.8683	8.755906	-32.3403	-7.67909	-21.074	-26.5504	-26.5504
mfcc12	mfcc13	mfcc14	mfcc15	mfcc16	mfcc17	mfcc18	mfcc19	mfcc20
-14.9318	6.133262	-16.1987	-2.61637	-5.95378	0.086543	-12.2652	2.097751	-5.98573

Figure 10. Sample of Extracted feature set from the Spectrogram of the file 'ahiba-angry.wav'

The set of MFCC and other feature values extracted and input into the convolution layer. In the model the convolution layer is followed by a pooling layer and the convolution layer output is considered as input for the pooling layer. Max pooling increases the efficiency and avoids the over fitting of the CNN model by lessening the number of arguments and computation steps [37]. In the proposed CNN structure, A convolution layer is followed by a pooling layer, and the total model comprises two pooling layers. Pooling layers summarizes the presence of features in patches in the feature map like a filter. The figure 10 depicts the extracted feature values from a sample spectrogram of the raw speech file 'ahiba-angry-01.wav' for the Assamese vocabulary 'ahiba' recorded in angry emotion.

- Convolution layer

The first convolution layer receives the input signals in terms of Male Frequency Cepstral analysis (MFCC) features set. In figure 11. the 6x6 matrix shows the input signal values in the form of MFCC into the convolution layer.

The convolution operation in the model can be perform by using a 3x3 weight matrix over the 9x9 matrix. The final values of the figure 11, such as 77.8658 is obtained by performing the convolution operation between the weight matrix and input wave signal.Thus in the convolution layer, the input signal size is reduced from a higher dimension to some lower dimension without removing the significant features set.

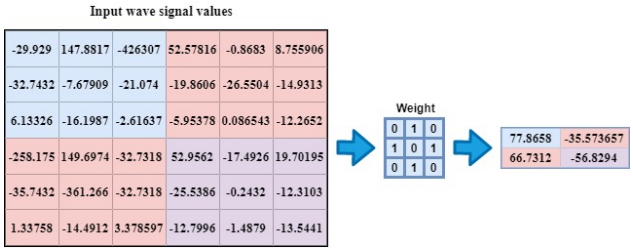

Figure 11. Filtering of input values by a convolution layer in CNN model

- Pooling layer

The set of MFCC featured values extracted in the convolutional layer is input to the pooling layer. It increases efficiency and avoids the over-fitting of the CNN model by decreasing the number of arguments and calculations of the model [37]. Two pooling layers are used in the entire model. Two types of pooling layer, such as average pooling and Max pooling layers computes the features in patches in the feature map like a filter. In this, the max pooling is used to filter the feature set.

- Fully connected layer

A set of significant features from the input speech signal are extracted by the convolution and pooling layers. To produce the required number of output classes, Fully-connected (FC) layer is used. Here we have 35 neurons in the flattened layer for identifying 35 different vocabularies from Assamese language. Neurons of FC layers have connections with all activations of the previous layer. A loss-function softmax cross-entropy is used in the FC layer and computes the error during prediction. In the model, once the forward passes are

finished, it starts Back propagation to reduce the error and loss from the prediction by updating the weight and biases.

- Softmax cross-entropy

At the last of the neural network model, the softmax function is applied to convert the outputs into a normalized probability distribution. Using softmax it has been analysed the derivative of the loss function for all the weight values of all the input signals in the training set. After several iterations during the training of the model, the weight values are updated and come closer to the preferred values, and increase the prediction accuracy of the model. The mathematical definition for softmax is defined in equation (1), where z_i are the input vector elements and $\bar{z}$ is the input vector. On each of the input vector elements, the conventional exponential function e^{z_i} is applied. The normalisation factor in the denominator ensures that the output prediction is in between the range (0,1) and the sum of the probability distribution always as 1. The value k represents the number of classes in the case of a multiclass classifier.

$$\sigma(\bar{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \tag{1}$$

C. *Assamese automatic emotional speech recognition using FFANN*

In recent years all the neural network models in machine learning comprised of artificial neurons mimicking the structure of a biological neuron. The authors presented an ANN model [7], [8] to recognize speech signals. The authors of [29] used the ANN structure Recursive Neural Network to analyse Assamese speech signals and classify speech emotions Angry, sad and happy.

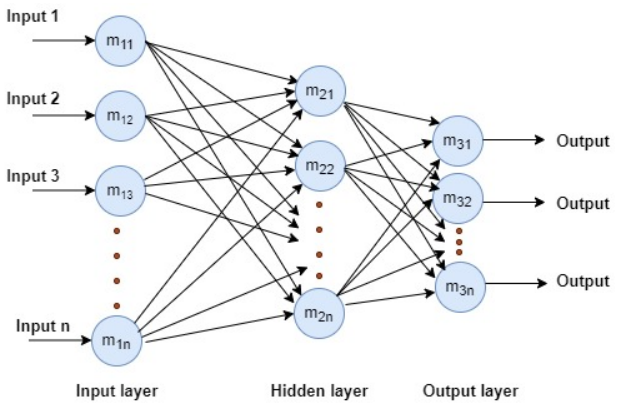

Figure 12. Architecture of Feed forward neural network model

In this proposed system a neural network has been designed with one input layer consist of 64 nodes, two intermediate layers with 32 and 16 nodes respectively, and an output layer with 35 nodes. The output layer contains several nodes equal to the classes we need to classify.

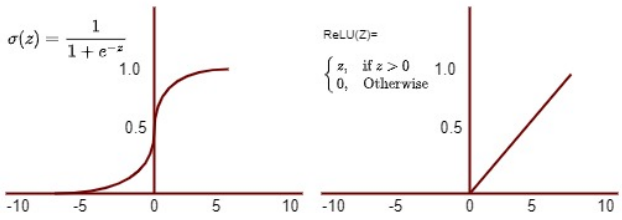

Figure 13. Activation functions used in Neural network

An architecture of Feed Forward Neural Network model is shown in the figure 12. To sum up the input values of one level of neuron and decide on the firing the node and transforming the summed weight into the next layer of neuron, the activation function RELU is used. RELU behaves as a linear function for positive values of input values but shows non-linear properties for negative values and returns zero as depiced in equation 2. RELU does not require any complex exponential functions as in sigmoid or TANH during activation as shown in the equation 3 and 4 respectively. The figure 13 is used to depict the graphical representation of the equations 2, 3 and 4 . RELU allows more than one true zero value for negative inputs of hidden layers and this sparse representation increases the learning capacity and ease of the model.

$$ReLU(Z) = \begin{cases} z, & \text{if } z > 0 \\ 0, & \text{Otherwise} \end{cases} \tag{2}$$

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{3}$$

$$\sigma(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}} \tag{4}$$

For doing the recognition, two machine learning approaches FFANN and CNN have been used during the experiment. The neural network FFANN is subsists of an input and a output layer, and one or more intermediate layers. The structure of the neural network that has been designed for this experiment consists of one input layer with 64 nodes, two intermediate layers with 32 and 16 nodes respectively, and an output layer with 35 nodes. The output layer contains several nodes equal to the classes we need to classify. Spectrograms of each audio file from the dataset were extracted and converted to.png an image file format of size 576 X 576. From the spectrogram, the required information such as MFCCs, SC, ZCR, CF, and spectral roll-off has been extracted and save into a file with .csv format.

It has been observed that the smoothness of a sound depends on the contents. A voiced sound is smoother in comparison to an unvoiced sound. So smoothness of a speech signal is considered an informative characteristic. To measure the smoothness of the input signals zero-crossing

rate is calculated. The formula to calculate the zero-crossing rate (ZCR) is:

$$ZCR = \sum_{a=-\infty}^{\alpha} | [f[z(a)] - f[z(a-1)]] | w(b-a) \quad (5)$$

where

$$f[z(a)] = \begin{cases} 1, & \text{if } z(a) \geq 0 \\ -1, & \text{if } z(a) < 0 \end{cases} \quad (6)$$

and $w(b)$ is the windowing function with a window size of B numbers of samples

$$W = \begin{cases} \frac{1}{2B}, & \text{if } 0 \leq b \leq B-1 \\ 0, & \text{Otherwise} \end{cases} \quad (7)$$

3. RESULTS AND DISCUSSION

A. Results of CNN

During the machine evaluation the recognition of 35 isolated words in five different emotions, angry, happy, sad, fear, and normal has been analysed by using the Convolutional Neural Network model. 20 MFCCs has been used as features set for the raw speech signals. CNN can be fed with raw speech signals [11]. A speech signal of 1-second length can have approximately 16000 features, which is a very huge amount. The time consumption is more during the processing of such raw speech signals. In this regard, the calculation of a fixed 20 MFCCs features has been considered during the experiment. For this purpose, a total of 12,250 emotional speech data with five different emotions have been considered. These words are spoken by 5 native male and female speakers of the Assamese language.

During the recognition process, the selected sub dataset is split into two sections: training and testing, in the ratio 50:50 respectively. The training set has been composed of 6125 speech recordings and the test set has been composed of 6125 speech recordings.The division of train and test set is done by manually. One speaker used to utter a word 10 times repeatedly in one emotion. During the division of the train and test dataset five utterances were kept in train set and remaining 5 utterances were kept in the test set. In each of the iteration, 100 speech files have been used for training the dataset. Total number of iterations have been considered as 750, 1000, 1250, and 1500 separately during the training session and 50 iterations have been considered during the testing session. The CNN model has been implemented with One fully connected layer, one output layer, and two convolution layers as shown in figure 6.

In the first two convolution layers, The activation function ReLU (Rectified Linear Unit)$f(y) = max(0, y)$ is used. Dropout has been used to avoid overfitting the network. To measure the loss during training, a function has been designed using softmax-cross-entropy and Adam Optimizer has been used to minimize the loss. For speech in five common emotions, the training accuracy rate and loss during training has been presented in figure 14 and 15 respectively.

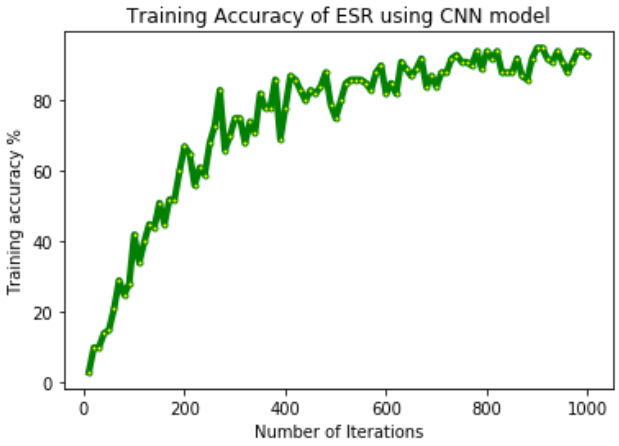

Figure 14. Training accuracy using CNN model

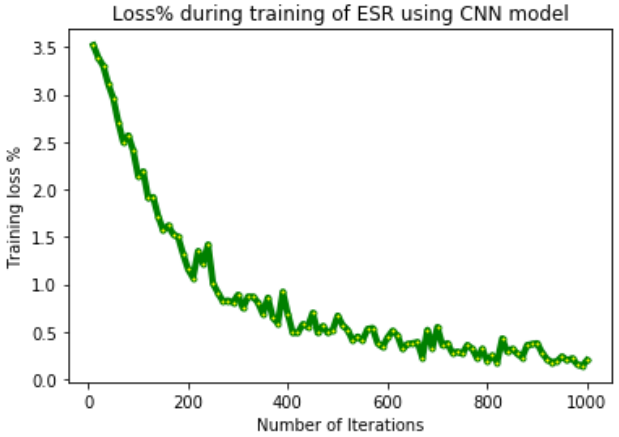

Figure 15. Loss % during training using CNN model

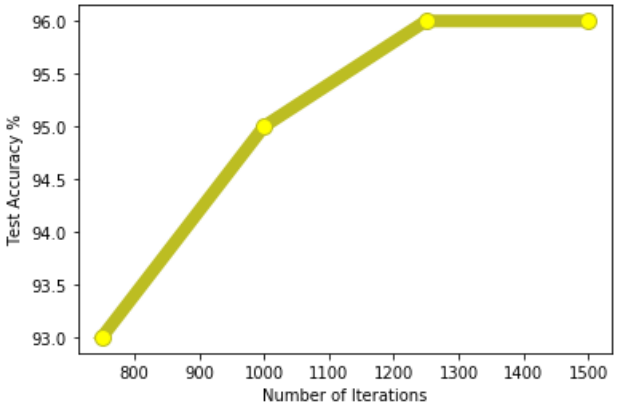

Figure 16. Training accuracy using CNN model

TABLE I. Training accuracy % at different epoch using CNN

Epoch	Accuracy %	Loss %	Run Time/step
750	93.4	0.6	38ms
1000	95.3	0.4	41ms
1250	96.1	0.2	39ms
1500	96.1	0.2	38ms

The value of epoch plays a vital role during training and testing of neural network models. Accuracy % are found as very poor for lower epoch during the training. The accuracy % is gradually increases up to a certain epoch value during training. Beyond this, the training accuracy % got a saturation state as shown in the table I. From the Table II, it has been observed that, the testing accuracy % have reached a satisfactory level for a very low epoch value. Beyond this point, the testing accuracy is got saturated.

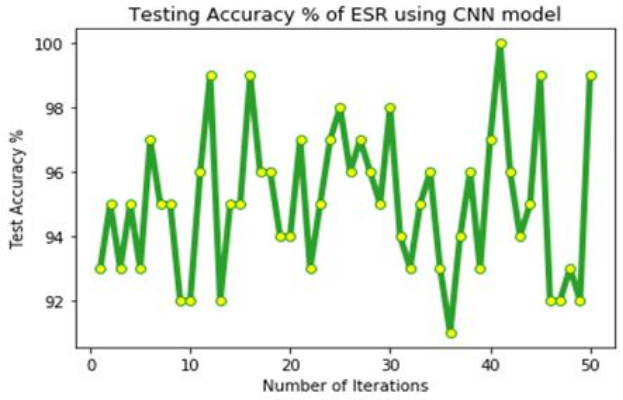

Figure 17. Testing accuracy using CNN model

TABLE II. Testing accuracy % at different epoch using CNN

Epoch	Accuracy %	Loss %	Run Time/step
10	92.5	0.6	42ms
20	94.3	0.4	39ms
30	97.2	0.3	38ms
40	96.7	0.2	40ms
50	98.4	0.2	42ms

The graph in figure 17, testing accuracy for the speech signals in five emotions has been depicted and the values scattered within the range of 92 to 100. The average test recognition rate was recorded as 98.4% for the emotional raw speech signals. Figure 17 depicts the iteration-wise test accuracy percentage for speech signals in five different emotions. The loss percentage during testing is shown in figure 18. The summary of the testing accuracy observed for different epoch using CNN is depicted in table II.

B. *Results of FFANN*

The energy level of each pitch class contained in the signal is normally represented by a 12-element feature vector called a charateristic vector or Chroma feature. After completion of the feature computation phase, these are loaded from the .csv file for pre-processing such as label encoding, feature scaling, and splitting the dataset have been performed. Generally, the raw data are saving with labels such that human beings can understand them easily. But to make it a machine-readable format, label encoding process is used such that ML algorithms can use it efficiently.

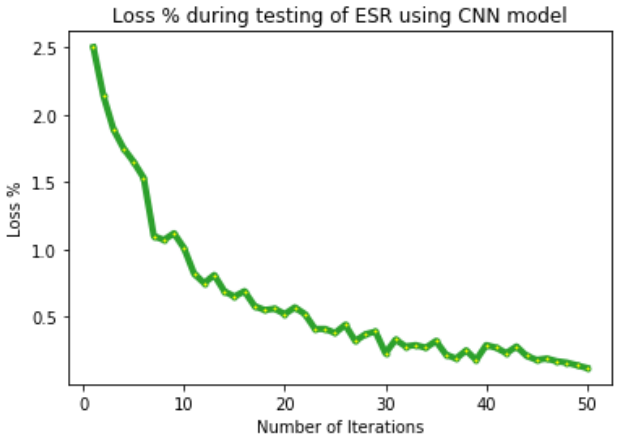

Figure 18. Loss % during testing using CNN model

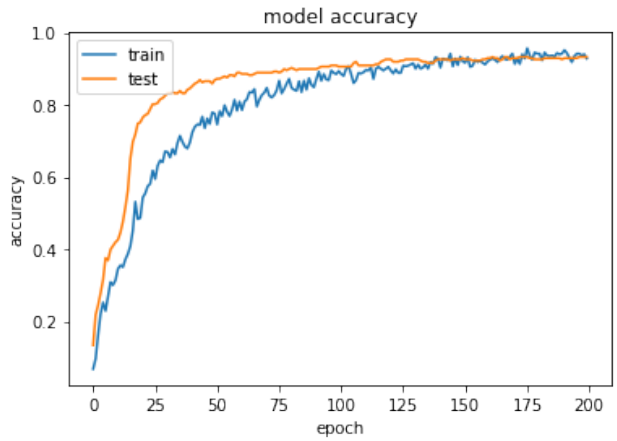

Figure 19. Training and testing accuracy % using FFANN model

To standardize the independent feature range feature scaling is performed. Finally,the entire dataset is split into two equal data sets: train and test.The test set is utilised during the evaluation phase and the train set is used during the training of the FFANN model. During training, a dropout rate of 0.2 is used to reduce the overfitting of the model. Adam optimizer is used to tune the loss that occurs during training and the loss function ”sparse-categorical-crossentropy” is used to calculate the loss during training.

TABLE III. Training and testing accuracy and loss % using FFANN model

Epoch	Training Accuracy %	Testing Accuracy %	Training Loss %	Testing Loss %
50	47.3	60.3	1.5	1.2
75	56.7	75.5	1.2	1.1
100	61.4	74.1	1	0.7
125	68.2	79.7	0.9	0.6
150	78.3	82.6	0.8	0.5
175	82.5	85.4	0.6	0.5
200	85.1	86.4	0.5	0.4
225	86.3	86.1	0.4	0.3
250	86.5	86.3	0.4	0.4

The recognition accuracy during training and testing the model is shown in figure 19. And the loss during training and testing is shown in figure 20. The test accuracy is found as 86.4% using the FFANN model when the epoch is 200. As the epoch value is increasing , the training and testing accuracy is also increasing up to a certain epoch as shown in figure 19. Beyond this value, the training and testing accuracy % got saturated. According to the table III, at epoch 200 or higher, the training and testing accuracy lies within the range 85 % to 86.5% using the FFANN model. The model loss is decreasing as the epoch value is increasing as shown in figure 20. The result of FFANN model is depicted in the table III.

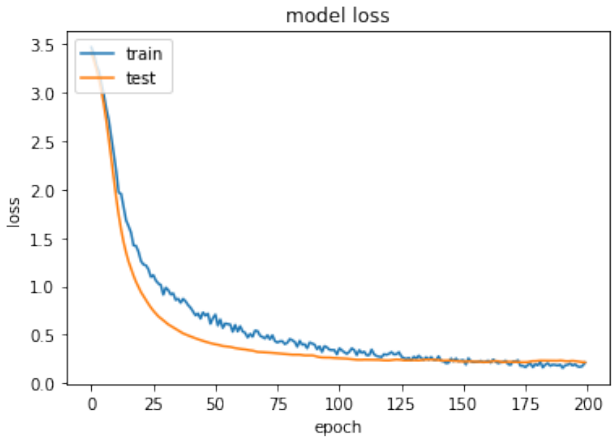

Figure 20. Recognition Loss % during training and testing using FFANN model

C. Discussion

FFANN is a basic artificial neural network and successfully used in various Automatic speech recognition. Before input, a speech signal into the ANN model, lots of preprocessing tasks such as noise removal, removing unvoiced signal, computation of feature set, etc. has to be complete. CNN is an advanced version of the ANN consists of convolution layers, pooling layers, a fully connected layer, and dropout layers. CNN can receive raw speech signals as input and extracts the required feature set automatically. In recent years, ANN structure is used successfully in the identification of Parkinson's disease by Ramzi et al [3] in 2019. In 1997, Sepp Hochreiter et al suggested the architecture of a DNN called RNN and LSTM [4], [5]. LSTM and RNN structure was able to avoid the vanishing gradient problem [38]. In the year 2009, the concept of a deep feed-forward network for acoustic modeling has been proposed by Geoffrey Hinton et al in the University of Toronto. Common problems faced in a traditional neural network such as gradient diminishing, weak temporal correlation, handling a large dataset had overcome by this new Deep feed-forward neural network architecture. A deep neural network is a type of ANN with single input and output layer, and two or more intermediate layers [37]. Due to the dense layers in the DNN structure, chances of overfitting are more in comparison to ANN. Due to the effectiveness of auto features detection and classification of bi or multi-class with less error, the DNN become the most popular acoustic modeling in the era 2010 [22]. In recent years, in the field of voice recognition system, DNN emerges as a revolutionary technique. Simultaneously lots of new DNN architectures have been evolved such as LSTM and CNN etc. Some other popular hybrid models of deep learning HMM/DNN, GMM/DNN, Deep Belief Neural Network (DBNN), etc. are also used successfully in the ASR system and other classification tasks. It has been seen from the study in 2012 by Deng et al [7], [20] and in 2011 by Seide F et al [21] that as an acoustic model, DNN improves its performance in the field of speech classification significantly. In the year 1998, Yann LeCun et al have been proposed a handwritten character recognition system using a new Deep learning structure called Convolutional Neural Network (CNN) [5]. CNN starts with a convolutional layer associated with activation function and followed by pooling layers. More than one combination of convolutional and pooling layers can exist in one structure. The output layer with a softmax layer is known as a Fully Connected Layer. In 2012 CNN become popular among the researchers as it overcomes the drawbacks of simple DNN structure and got succeed in the classification of large vocabulary speech datasets [20]. In the field of speech recognition, CNN has shown to be successful and effective. In 2016 [9] W.Lim et al, 2015 [10], [11], Palaz et al. Suggested

CNN based acoustic models to accept raw speech data directly as input. In [10] Palaz et al. demonstrated that the CNN model can derive class conditional probabilities from a raw speech input. Lots of research work has been performed for the prediction of speech content from speech signals using CNN architecture. In 2014, Huang et al. proposed a semi CNN model and tested its efficiency on four different publicly available datasets SAVEE, EMO-DB, DES, and MES [12]. They reported the accuracy rate of the proposed Semi-CNN architecture for Single speaker, Speaker dependent, and Speaker Independent as 91.1%, 81.27%, and 79.25% respectively. Rana et.al in 2016, used a deep learning approach to recognise emotions from noisy speech and claimed an error rate of less than 10% [19]. In 2016, Lim et al. used CNN, LSTM, and Time Distributed CNN and has shown accuracy rate as 86.06%, 78.31%, and 86.65% respectively on a noiseless emotional dataset [8]. In 2017, Badshah et al. described an experiment to recognize emotion using Deep Convolutional Neural Network (DCNN) from spectrograms achieved from some emotional speech and reported an 84.3% accuracy rate [22]. S. Zhang et al. in 2018, presented a deep neural architecture DCNN-DTPM and has been reported the accuracy rate on the datasets EMO-DB, RML, and BAUM-1 as 87.31%, 75.34 and 44.61% respectively [24].

In 2014, Abdel et al. presented a CNN model and described how to use the CNN model in Automatic speech recognition. The error rate has been reduced by 6% to 10% in the CNN model as compared to other DNN and neural network models [13]. Lee et al [18] presented a Convolutional Deep Belief Neural Network (CDBN) and successfully used in the evaluation and classification of feature learning representation on different types of unlabeled audio data such as speech, and music. The input speech signals have been converted into a spectrogram at first and passed to the CNN model as input signals. In 2016, there has been a proposal for a CNN model that is aware of the Signal to Noise Ratio (SNR) by Szu-Wei Fu et al and successfully used for speech enhancement (SE) [14]. They proved the better performance of the CNN model in SE through extracting TF features from speech signals than the DNN model. In another research work in 2015, CNN was successfully presented an Audio Visual recognition system by Kuniaki Noda et al [15] using the AURORA dataset. An airport inquiry system in the Telugu language has been developed by D. Nagajyothi et al in 2018 using CNN architecture [16]. Li Deng et al in [25] presented a Deep Convolutional Neural Network (DCNN) structure containing heterogeneous pooling layers and a fully connected multilayer neural network.It was employed in speech spectrograms to enable regulated frequency-shift invariance. In 2019 [23], and 2018 [24] presented Deep Convolutional Neural Network (DCNN) to recognize speech emotions from the spectrogram of speech signals. In 2018, Zhang et al presented a Deep Neural Network (DNN) comprised of CNN for a complete emotion classification and detection [19] from the raw recorded speech signals. During the last decade, lots of research work has been done on automatic speech processing in the Assamese language. Most of the experiments carried out in normal speech signals and a few works have been done in emotional raw speech. Sharma et al [28] 2016, presented a work using RNN and DTDNN model and showed that the RNN recognition rate is better than DTDNN with 86.55% accuracy. They used an Assamese emotional speech database with 1440 samples recorded in a noiseless room environment in three common emotions such as normal, loud, and angry. In 2015, Michael et al. have built an emotional corpus of speech data from Assamese movies using three different models such as Random Forest, Gradient Boosting, and Linear Support Vector Machine (SVM) [26]. In 2008, Kandali et al. have presented a project work on emotional speech in the Assamese language in seven different emotions where seven sentences were recorded from 14 male and 13 female native speakers in a quiet environment and reported 74.4% of recognition accuracy using GMM architecture [27]. In 2015, Gogoi and Kalita have collected 120 samples from 10 actors with 4 emotions in the Assamese language and build an ASR system using MFCC feature and GMM model [39]. All the mentioned works in recognized emotions attached to the speech signals [26], [27], [28], [39].

In this paper, it has been tried to recognize speech from raw speech signals in different emotions recorded in a normal noisy environment through WhatsApp using FFNN and CNN model architecture. In this proposed work, we have considered MFCC feature extraction for both the model and it has been observed from the experimental result that - Emotional Assamese speech recognition in a noisy environment, the CNN model shows better performance with 98.4% testing accuracy in comparison to FFANN model. The testing accuracy is recorded 86.4% using the FFANN model. A comparative study between different machine learning approaches used for Assamese language is summarized in the following table 5.

D. Conclusion

The convolutional neural network model has shown an effective result in the recognition of emotional speech in a noisy environment regarding the Assamese language. Lots of research issues are still there to explore such as recognition rate for male and female speakers separately, phoneme recognition, speaker identification, age detection from speech, disease diagnosis, etc. Our next experiment will be on Phoneme and speaker identification for emotional speech in the Assamese language and build a corpus from different age groups.

References

[1] G. Goswami, *Structure of Assamese*. Department of Publication, Gauhati University, 1982.

[2] "Languages in india - map, scheduled languages, states official languages and dialects," https://www.mapsofindia.com/culture/indian-languages.html, (Accessed on 03/17/2022).

TABLE IV. Models used in emotion and emotional speech recognition

Reference	Model Name	Used Language	Recognition Objective	Recognition Accuracy
[9]	LSTM	-	Emotion recognition	78.31%
	Time Distributed CNN	-	Emotion recognition	86.65 %
[13]	Semi-CNN (Single speaker)	-	Emotion recognition	91.1%
	Semi-CNN (Speaker dependent)	-	Emotion recognition	81.27%
	Semi-CNN (Speaker Independent)	-	Emotion recognition	79.25%
[14]	DNN	American English	Emotion recognition	37.1%
	CNN			34.2
[19]	DBN	German	Emotion recognition	80%
	CNN	-	Emotion recognition	86.06%
[22]	CNN	German	Emotion recognition	84.3%
[24]	DCNN	German	Emotion recognition	86.30%
[27]	GMM	Assamese	Emotion recognition	74.4%
[28]	DTDNN	Assamese	Emotion recognition	80%
[39]	GMM	Assamese	Emotion recognition	43.39%
	RNN	Assamese	Emotion recognition	86.55 %
Proposed work	FFANN	Assamese	Emotional Speech recognition	86.4%
Proposed work	CNN	Assamese	Emotional Speech recognition	98.4%

TABLE V. Comparison of recognition accuracy of different machine learning models for Assamese language

Reference	Model Name	Used Language	Feature set	Recognition Accuracy %
[26]	Logistic Regression	Assamese	23 MFCC Vector	45%
	Random Forest	Assamese	23 MFCC Vector	68 %
	Gradient Boosting	Assamese	23 MFCC Vector	50%
	Linear SVM	Assamese	23 MFCC Vector	40 %
[27]	GMM	Assamese	14-MFCC 14-delta MFCC 14-delta-delta MFCC	74.4%
[39]	GMM	Assamese	MFCC	43.39%
[28]	RNN	Assamese	MFCC-delta	86.55%
	DTDNN	Assamese	MFCC-delta	80%
Proposed work	FFANN	Assamese	MFCC	86.4%
Proposed work	CNN	Assamese	MFCC	98.4%

[3] R. M. Sadek, S. A. Mohammed, A. R. K. Abunbehan, A. K. H. A. Ghattas, M. R. Badawi, M. N. Mortaja, B. S. Abu-Nasser, and S. S. Abu-Naser, "Parkinson's disease prediction using artificial neural network," 2019.

[4] U. Barman and R. D. Choudhury, "Smartphone image based digital chlorophyll meter to estimate the value of citrus leaves chlorophyll using linear regression, lmbp-ann and scgbp-ann," *Journal of King Saud University-Computer and Information Sciences*, 2020.

[5] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[7] A. Ries, C. Chang, S. Glim, C. Meng, C. Sorg, and A. Wohlschläger, "Grading of frequency spectral centroid across resting-state networks," *Frontiers in human neuroscience*, p. 436, 2018.

[8] L. G. Dahl, J. W. Stokes, and L. Deng, "Dong yu," *Context-*

dependent pre-trained deep neural networks for large vocabulary speech recognition, vol. 20, pp. 30–42, 2012.

[9] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA)*. IEEE, 2016, pp. 1–4.

[10] D. Palaz, R. Collobert, and M. M. Doss, "End-to-end phoneme sequence recognition using convolutional neural networks," *arXiv preprint arXiv:1312.2137*, 2013.

[11] D. Palaz, M. M. Doss, and R. Collobert, "Convolutional neural networks-based continuous speech recognition using raw speech signal," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4295–4299.

[12] D. Palaz, R. Collobert *et al.*, "Analysis of cnn-based speech recognition system using raw speech as input," Idiap, Tech. Rep., 2015.

[13] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using cnn," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 801–804.

[14] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[15] S.-W. Fu, Y. Tsao, and X. Lu, "Snr-aware convolutional neural network modeling for speech enhancement." in *Interspeech*, 2016, pp. 3768–3772.

[16] S. Sukittanon, A. C. Surendran, J. C. Platt, and C. J. Burges, "Convolutional networks for speech detection," in *Eighth international conference on spoken language processing*, 2004.

[17] D. Nagajyothi and P. Siddaiah, "Speech recognition using convolutional neural networks," *Int. J. Eng. Technol*, vol. 7, no. 4.6, pp. 133–137, 2018.

[18] H. Lee, P. Pham, Y. Largman, and A. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," *Advances in neural information processing systems*, vol. 22, 2009.

[19] R. Rana, "Emotion classification from noisy speech-a deep learning approach," *arXiv preprint arXiv:1603.05901*, 2016.

[20] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5089–5093.

[21] L. M. Q. de Santana, R. M. Santos, L. N. Matos, and H. T. Macedo, "Deep neural networks for acoustic modeling in the presence of noise," *IEEE Latin America Transactions*, vol. 16, no. 3, pp. 918–925, 2018.

[22] D. Yu, F. Seide, and G. Li, "Conversational speech transcription using context-dependent deep neural networks." in *ICML*, 2012.

[23] N. A. Kulkarni, "Speech recognition using convolutional neural network," 2019.

[24] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *2017 international conference on platform technology and service (PlatCon)*. IEEE, 2017, pp. 1–5.

[25] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2017.

[26] L. Deng, O. Abdel-Hamid, and D. Yu, "A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6669–6673.

[27] G. Michael, N. Kaphungkui, and P. K. Thakur, "Comparative study of various classifiers using assamese speech utterances."

[28] A. B. Kandali, A. Routray, and T. K. Basu, "Emotion recognition from assamese speeches using mfcc features and gmm classifier," in *TENCON 2008-2008 IEEE region 10 conference*. IEEE, 2008, pp. 1–5.

[29] R. Kaushik, M. Sharma, K. K. Sarma, and D. I. Kaplun, "I-vector based emotion recognition in assamese speech," *International Journal of Engineering and Future Technology*, vol. 1, no. 1, pp. 111–124, 2016.

[30] T. Arias-Vergara, P. Klumpp, J. C. Vasquez-Correa, E. Noeth, J. R. Orozco-Arroyave, and M. Schuster, "Multi-channel spectrograms for speech processing applications using deep learning methods," *Pattern Analysis and Applications*, vol. 24, no. 2, pp. 423–431, 2021.

[31] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms." in *Interspeech*, 2017, pp. 1089–1093.

[32] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, and E. Nöth, "Convolutional neural network to model articulation impairments in patients with parkinson's disease." in *INTERSPEECH*. Stockholm, 2017, pp. 314–318.

[33] H. Wu, J. Soraghan, A. Lowit, and G. Di Caterina, "A deep learning method for pathological voice detection using convolutional deep belief networks," *Interspeech 2018*, 2018.

[34] M. Alhussein and G. Muhammad, "Voice pathology detection using deep learning on mobile healthcare framework," *IEEE Access*, vol. 6, pp. 41 034–41 041, 2018.

[35] K. Han, Y. He, D. Bagchi, E. Fosler-Lussier, and D. Wang, "Deep neural network based spectral feature mapping for robust speech recognition," in *Sixteenth annual conference of the international speech communication association*, 2015.

[36] M. Cernak and S. Tong, "Nasal speech sounds detection using connectionist temporal classification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5574–5578.

[37] X. Huang, F. Alleva, S. Hayamizu, H.-W. Hon, M.-Y. Hwang, and K.-F. Lee, "Improved hidden markov modeling for speaker-independent continuous speech recognition," in *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.

[38] W. Zhang, A. Hasegawa, K. Itoh, and Y. Ichioka, "Error back propagation with minimum-entropy weights: a technique for bet-

ter generalization of 2-d shift-invariant nns," in *IJCNN-91-Seattle International Joint Conference on Neural Networks*, vol. 1. IEEE, 1991, pp. 645–648.

[39] N. J. Gogoi and J. Kalita, "Emotion recognition from acted assamese speech," *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 4, no. 6, pp. 4116–4121, 2015.

Dipankar Dutta Dipankar Dutta has been completed his Master of Computer Application (MCA) under Dibrugarh University, Assam, India in the year of 2007 and pursuing his Ph.D. under Gauhati University, Assam, India. He had been working as an Assistant Professor in North Eastern Regional Institute of Management (NERIM) since 2008. His research interest is in the field of speech processing and Expert systems.

Dr. Ridip Dev Choudhury Dr. Ridip Dev Choudhury received his Master of Science (M.Sc.) in Computer Science and Ph.D. in Computer Science from Gauhati University, Assam, India in the year of 2004 and 2014 respectively. Presently he is working as an Associate Professor in KKHSOU, Assam, India. His research interest is in the field of Speech Processing, Digital Image Processing and Expert System.

Dr. Utpal Barman Dr. Utpal Barman received his Master of Technology (M.Tech.) in Computer Science and Ph.D. in Computer Science from Gauhati University, Assam, India in the year of 2014 and 2021 respectively. Presently he is working as an Associate Professor at The Assam Kaziranga University, Jorhat, Assam, India. His research interest is in the field of Image Processing, machine learning, and deep learning.