Bengali Speech Emotion Recognition: A hybrid approach using B-LSTM

Abstract

Bengali Speech Emotion recognition is one of the major outstanding and novel pieces of work in the Bengali signal processing and the field of Bengali Artificial Intelligence (AI). A plethora of applications such as interactive consultancy, AI customer care services, etc requires identifying the emotion of the caller to give an appropriate response. The audio of human speech consists of various tones from the cochlea and vocal cords. Mel-frequency Cepstral Coefficients (MFCC), Chroma, Gamma -Frequency Cepstral Coefficients (GFCC), and Mel spectrogram are already state-of-the-art techniques to extract various features from human speech audio. In our study, a recently published dataset called "SUST Bengali Emotional Speech Corpus (SUBESCO)" is used to get the raw audio files for different emotions in the Bengali Language. MFCC, Chroma, and Mel-Spectrogram techniques are used to extract the features for each audio file in order to apply different Classifiers and RNN i.e. LSTM (Long Short Time Memory). We test and also compare 3 approaches: 1) Our proposed Neural Network model, 2) Support Vector Machine (SVM) Classifier, and 3) Multi-Layer Perceptron Classifier (MLPClassifier). Using our self-developed architecture, we achieve an overall accuracy of 83.33% with better average precision and f1-score than the SVM classifier and MLPClassifier in detecting 7 different emotion classes.

INTRODUCTION

Emotion recognition is crucial in social research and human-machine interaction because it allows us to see human actions and their correlations. Over the past two decades, research into SER (Speech Emotion Recognition) has gotten a lot of attention. Demonstration of SER models outperformed other models in the languages in which they were taught. To make the creation of SER easier, SUBESCO wants to create a distinct dataset for the Bengali language. SUBESCO dataset may be used to train ML (Machine Learning) models that identify fundamental emotions in Bengali speech. This would also enable fascinating linguistic comparisons between Bangla and comparable regional languages such as Hindi and Punjabi(Sultana et al. 2021b).As a result, many signals like facial expression, voice tone and body gestures are recognized in order to determine a person's emotional state. This crucial information about a person's mental state allows a machine to make a digital choice(Sharma and Singh 2014). Moreover, Deep learning has also discovered potential prospects in the field of Speech Emotion Recognition. The ability to identify emotional content in human speech has several applications and advantages. Maintaining a machine that can comprehend and maybe manufacture emotions will have a significant impact on how people interact with robots. We propose many upgrades to the present modern architecture for unimodal Speech Emotion Recognition (SER) in order to fully use the potential of audio signals. We present a new LSTM variation that can handle these two data sequences at the same time and use their complementing information(Wang et al. 2020). In order to establish a learning environment, a teacher, for example, can rely on an emotion recognition system (automated) from speech to rationally resolve content or topic primacy and invent techniques for manipulating students' emotions. In order to do this, one of the study's main goals is to look into the student's emotional (mental) condition.

MOTIVATION

Despite receiving little attention from researchers, emotion detection has a substantial influence on a variety of critical areas of our everyday lives. In the telemedicine period, through mobile platforms patients are examined, the expertise of medical to detect emotional experience of a patient through automation process is valuable for diagnosing and prescription..A client service conversation may also be used to examine customer service employees' personality psychology with clients and, as a consequence, enhance service quality.Furthermore, making acceptable contributions to the commercial sector by proposing different internet items to target customers depending on their opinions about the products. One of the most commonly spoken native languages in the world is Bengali, and it is extensively used for everyday communication.In comparison to other languages, however, Bengali has a dearth of research on automated speech emotion identification. One of the key reasons for the failure of academic interest in Bengali is the shortage of Bengali materials. In this context, it is expected that this study will lead to the creation of smarter new devices that can automatically analyze human emotions, with implications extending to many facets of everyday life.

RELATED WORKS

Emotion recognition from speech is a hot topic in academia. Several research studies have been conducted to see how this field might be improved. In order to extract characteristics, A real-time emotion identification system that analyzes tonal features to recognize emotions from live recorded speech is demonstrated(Iqbal and Barua 2019). MFCCs, energy, spectral entropy, and other audio properties are extracted. Gradient Boosting is used to train models that classify emotions in this way. SVM (Support Vector Machine) and KNN (K-Nearest Neighbor) are two more classifiers used to examine the accuracy of test audio recordings. Two datasets were utilized to train the system: RAVDESS and SAVEE. Anger, happiness, sorrow, and neutral are the four emotions examined by this approach. The author of the study for the Bengali language created a new open access emotion speech corpus (audio only)(Sultana et al. 2021b), which spans 7 hours and contains 7000 utterances. 20 native speakers took part, each recording of the audio files contains ten words imitating seven different emotions, and fifty university students took part in the evaluation of the largest Bengali language corpus. The reliability of the corpus was assessed using a variety of statistical methods, including Statistics (Kappa) and correlation coefficient (intra-class) scores. For human perception testing, good results were obtained (up to 80 percent). The impacts of Gender and Emotion were investigated using a two-way ANOVA. Levene, Shapiro-Wilk Jarque-Bera, and Bartlette tests were used to examine the normality and homogeneity of data for these parameters. The BanglaSER dataset(Das et al. 2022) consists of 1467 Banglaspeech-audio recordings of five basic human emotions: angry, happy, neutral, sad, and surprise. For speech emotion recognition, machine learning models can be trained using this dataset. The BanglaSER dataset is a speech-audio dataset that was created primarily for the SER task. It has 34 nonprofessional actors in a Bengali accent pronouncing three lexically-matched Bangla phrases. For each emotional state, three trials were performed. The SER task in Bangla was the driving force behind machine learning. A fresh Bengali speech dataset was created and used in a study(Hasan and Islam 2020). The human voice (speech) is classified into six emotional states using recurrent neural network (RNN) technology. This research explores machine learning algorithms for categorizing human emotion into six major stages and presents an autonomous emotion recognition system based on Bengali speech. Because of the classification of the dataset, performance in detecting happy and furious emotions has somewhat deteriorated, but performance in detecting all other emotions has improved. Human emotive information from speech which is for automated recognition, modulation spectral features (MSFs)(Wu, Falk, and Chan 2011) have been proposed. MSFs are derived from a long-term spectro-temporal representation inspired by auditory input. For identifying seven emotion types, 91.6 percent overall recognition rate was obtained. When each feature type is utilized alone, MSFs surpass MFCC and PLP features in terms of FDR scores and recognition accuracy. When MSFs are paired with prosodic characteristics, overall recognition accuracy improves to 91.6 percent. Estimating activation and dominance yields promising results, but valence yields less promising results. A critical component SER(Wang et al. 2020) has emerged in the future. To predict emotions, the author offers a new duallevel model that employs raw audio data including MFCC features and mel-spectrograms. In multimodal data study, audio signals, lexical information, and, in rare cases, visual information are all utilized. Multimodal models make use of more data and are, on average, more accurate. Two collaboratively trained independent neural networks, they proposed a new dual-level model. On average, 72.7 % weighted accuracy and 73.3% unweighted accuracy, which is a great amount of improvement and it is 6% for multimodal models. A hybridization of GFCC (Gammatone Frequency CepstralCoefficients) and BPNN (Back Propagation Neural Network) will be used to discern speech emotions in a proposed work(Sharma and Singh 2014). Using continuous BPNN and GFCC approaches, the author provides two functioning engines that leverage both of the referenced possibilities. For audio wave files, the technology has been fully developed and tested. The remaining papers in this series detail the creation of a model for recognizing human emotions in speech automatically. The findings of using BPNN and GFCC to recognize emotions are significant. Some emotions appear to be more easily detected than others, which could be due to our brains' perception that they are more prominent.

Audio Features Deep Learning

Audio features can be extracted in various ways. The methods of audio extraction can be MFCC, Chroma, Mel Spectrogram, GFCC, etc. Audio is a file where fluctuation is happening in a very little span of time. So, if we want to capture the data from the audio such as emotion, background noise, etc can only be captured accurately if the frame size is very much small. So, MFCC can help us in this manner. In MFCC, framing units are used to approximate the energy spectrum and the Fourier transform, which are then expressed on a scale based on the Mel-frequency scale. Framing units are also used to approximate the Fourier transform. A discrete cosine transform (DCT) of the Mel-log energies was then illustrated. It should be noted that the first twelve coefficients of DCT give the MFCC values necessary for the categorizing procedure. The chroma feature is considered as an audio source (musical) reflecting tonal components. For this reason, chroma characteristics are a must prerequisite for semantic analysis (High Level) which can be chord identification or harmonic similarity estimates. When retrieving the feature vector from the magnitude spectrum, a short-time Fourier transform (STFT), Constant-Q transform (QCT), Chroma Energy Normalized (CENS), and other techniques are used. A spectrogram is a visual representation of the frequency content of a signal across time. Human auditory systems are able to use a linear scale provided by the Mel scale which is connected to Hertz by the following formula:

$$m = 2595log_{10}(1 + \frac{f}{700}) \quad (1)$$

In GFCC feature extraction(Zhao, Shao, and Wang 2012) sixty-four-channel grammtone filter banks have been used.

Here, filter response has been rectified and decimated to 100 Hz. This results in a cochleagram like time-frequency (T-F) representation.

MFCC and delta MFCC

The Mel Frequency Cepstral Coefficients are static coefficients that are one of the most renowned feature extraction techniques and one of the standard approaches for speech analysis. The voice waveform is converted into a series of discrete acoustic vectors. MFCC is a simple model of auditory processing that conducts reasonably quick acoustic analysis and depicts the ear model, resulting in good results in speaker recognition when a large number of coefficients are utilized(Baroi et al. 2019). The MFCC step-by-step competition is as follows: To acquire the MFCC coefficients, take the log scale of the window function waveform, smooth it out with triangle filtration, and then compute the DCT of the waveform. s(n) which indicates the voice signal is first passed via a high pass filter:

$$S(n) = s(n) - a * s(n-1) \tag{2}$$

Where the output signal is S(n) which is between 0.9 to 1.0. Compensating method like pre-emphasis which is for the high-frequency component that was suppressed during the human sound generating mechanism. It can also be used to emphasize the significance of high-frequency formants(Singh and Ghangas). Differential and acceleration coefficients are sometimes known as Deltas and Delta-Deltas. The MFCC feature vector just provides the power spectrum envelope of a single frame, however, it appears that speech has dynamics as well. Delta features are used to indicate the changes in cepstral characteristics across time. Each delta feature is received as MFCC feature's first derivative, which represents where the change occurs between frames. Delta features benefits over MFCC features are that they are used to characterize temporal data(Singh and Ghangas). Calculating the MFCC dynamics and adding them to the original feature vector significantly improves ASR performance. The following formula is used to determine the delta coefficients:

$$d_t = \frac{\sum_{n=1}^{N} n(C_t + n - C_t - n)}{2\sum_{n=1}^{N} n^2} \tag{3}$$

here, dt is a delta coefficient derived in terms of the static MFCCs Ct+n to Ct-n from frame t(Amin and Rahman 2015).

Chroma The chroma feature extraction approach is applied to the voice part of the audio stream. The chroma feature is a visual representation of a feature that pertains to the "color" of a sound pitch. Chroma-based characteristics are a strong technique for assessing sound whose pitches can be meaningfully classified (typically into 12 categories) and whose tuning is close to equal-tempered(Kattel et al. 2019). We can utilize Chroma feature visualization to determine how prominent a pitch's features are in the sampled frame. Because the natural vocal folds of all areas are so identical, it's difficult to tell them apart. However, when speakers use their vocal folds differently, as is the case when people speak in distinct dialects although speaking the same language, it makes a difference(Badhon et al. 2021).

Mel Spectrogram Auditory frequencies are perceived in a logarithmic manner by humans. Mel-scale is a logarithmic scale representation of signal frequencies, which is comparable to this concept(Sultana et al. 2021a). The spectrogram illustrates the evolution of frequencies over time. The time-frequency form of this signal is critical for many studies in which time is the only parameter. The frequency-domain representations alone do not give sufficient categorization information(Sejdić, Djurović, and Jiang 2009). Each Mel's power spectrogram is represented by a Mel spectrogram versus the passage of time, and it can be used to demonstrate the relative value similar to how humans use different frequency bands a sense of hearing The mel spectrum and its linkThe signal frequency fHz and the frequency fmel are defined as:

$$f_{mel} = 2595.log_{10}(1 + \frac{f}{700}) \tag{4}$$

Gammatone Frequency Cepstral Coefficients (GFCCs) In recent studies(Shao and Wang 2008) - (Shao, Srinivasan, and Wang 2007), the GFCC features have shown very good robustness against noise and acoustic change. In a hearing periphery model, GFCCs' main approach can be found where it mimics the filtering mechanism (human cochlear). To mirror the hearing model, the gammatone filter bank breaks down the raw voice into a T-F format. Numerous psychophysical and physiological studies in the human auditory peripheral led to the development of the Gammatone filters(Bouziane, Kharroubi, and Zarghili 2021).

DATA DESCRIPTION

In this research study, we used SUBESCO which was published by the Department of CSE of SUST (Shahjalal University of Science and Technology). SUBESCO is a Bangla language audio-only emotive speech corpus. The corpus has a total duration of over 7 hours and contains over 7000 statements, making it the biggest emotional speech collection accessible for this language. The gender-balanced collection included twenty native speakers, each of whom recorded ten words imitating seven different moods. In SUBESCO, there are 7000 audio files representing 7 different emotions. The collection of six primary emotions, anger, disgust, fear, happiness, sadness, and surprise, as well as a neutral emotional state, were examined for corpus development(Sultana et al. 2021b). SUBESCO data was made after an inspiration of RAVDESS dataset(Livingstone and Russo 2018), which is working as an audio-visual resource that is for emotional speech and music in American English. We allocated 85% to training and 15% to testing, with 15% cross-validation applied to the 85% train data. In this research, we work on a real-time voice dataset to find those 7 emotions. All 7 types of emotion spectrogram and waveform are given below with examples.

Example 1:
Bangla sentence: মৌমাছির চাক দেখে কুকুরটি ঘেউ ঘেউ করছে।
Emotion: Happy.

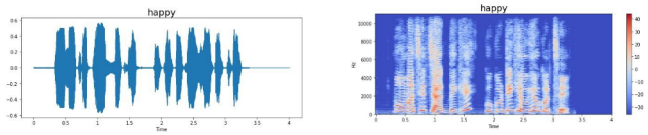


Figure 1: Spectrogram and waveform of Happy emotion audio speech.

Example 2:
Bangla sentence: ডাকাতেরা ঢল তলোয়ার নিয়ে এলো।
Emotion: Angry.

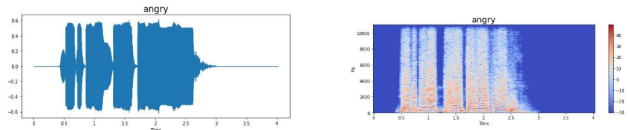


Figure 2: Spectrogram and waveform of Angry emotion audio speech.

Example 3:
Bangla sentence:সে কোন কিছু না বলেই চলে গেছে।
Emotion: Surprise.

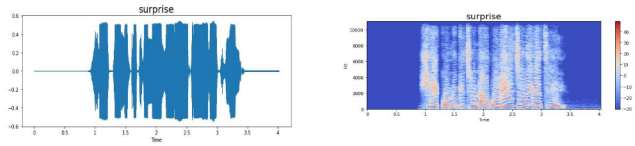


Figure 3: Spectrogram and waveform of Surprise emotion audio speech.

Example 4:
Bangla sentence:তোমার কাজটা করা ঠিক হয়নি।
Emotion: Fear.

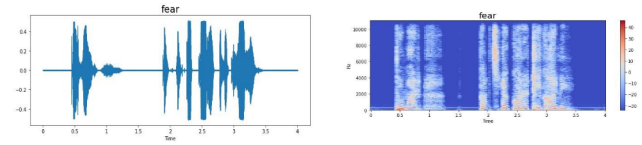


Figure 4: Spectrogram and waveform of Fear emotion audio speech.

Example 5:
Bangla sentence: দরজার বাইরে কুকুর দাড়িয়ে আছে।
Emotion: Disgust.

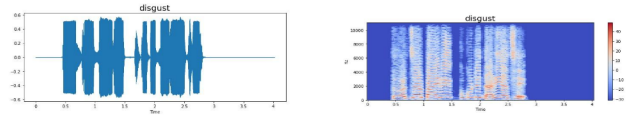


Figure 5: Spectrogram and waveform of Disgust emotion audio speech.

Example 6:
Bangla sentence: একদিন পরেই তার বিয়ে।
Emotion: Sad.

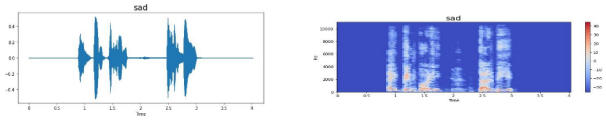


Figure 6: Spectrogram and waveform of Sad emotion audio speech.

Example 7:
Bangla sentence: তোমাকে এক্ষুনি আমার সাথে যেতে হবে।
Emotion: Neutral.

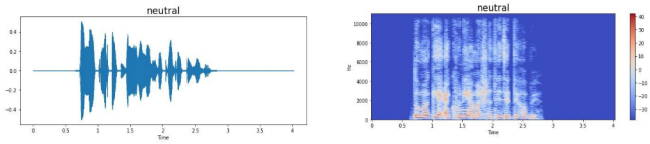


Figure 7: Spectrogram and waveform of Neutral emotion audio speech.

We already mentioned that there are 7000 emotional voice speeches and each emotion has 1000 voice speeches. This makes the dataset more suitable for our model and work plan. A bar chart is given below to show the voice datasets number visually.

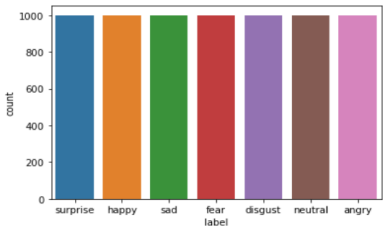


Figure 8: Class Balanced Nature of the dataset.

METHODOLOGY

In our experiment, we have used the SUBESCO dataset which contains 7000 audio files. First, we have extracted 75 features/attributes from the audio files using MFCC, Chroma, and Mel Spectrogram. Since we have extracted features from each and every input audio file, the training dataset contains 85% and the test dataset contains 15% of the total dataset. After the audio files have been split to train and test datasets, the train set contains 5950 audio files and the test dataset contains 1050 audio files. Then, we have run our proposed hybrid model of deep neural network using B-LSTM, MLP Classifier and SVM Classifier on the train and test dataset, and we observe the outputs of the 7 different emotions and then for experimental comparison we have calculated recall, f1-score, support and precision for each model and choose the best model based on the given parameters. Finally, we have run the best model on a real-

time dataset of audio files of 7 different emotions and observed how the model works on it. The figure below shows the workflow of our experiment:

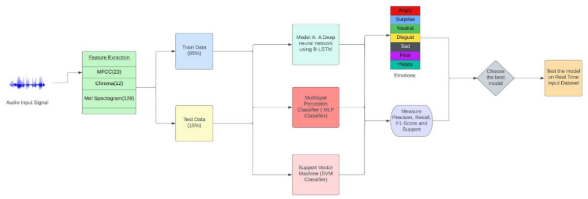


Figure 9: The Flowchart of our experimental workflow.

Features Extraction From Audio DATA

MFCC: In this study, a Mel-Frequency Cepstral Coefficients for each audio were calculated for 23 mfccs using a function of the librosa framework. The Mel frequency scale, which is nearly linear for frequencies below 1 kHz and logarithmic for frequencies above 1 kHz, was employed as an approximation for the non-linear frequency scale. This is due to the fact that when the frequency rises beyond 1 kHz, the human hearing system becomes less frequency-selective. Transformation (linear cosine) of log power spectrum used in sound processing, which is measured in nonlinear Mel-frequency scale is a short -term power spectrum representation(Amin and Rahman 2015).

Chroma: In this work, we used 12 different pitch classes to distinguish different emotions. For analyzing audio files, these pitch class profiles are quite valuable. The term chromagram refers to putting all of the pitches in an audio recording in one location so that we can understand how to classify them. Pitches are a feature of any sound or signal that enables the frequency-related scale to be used to arrange files. It's a metric for measuring sound quality that allows you to categorize sounds as higher, lesser, or medium.

Mel Spectrogram: In our research, we used Mel-spectrogram for audio feature extraction. Librosa framework's function was used to calculate each sentence's Mel-spectrogram where the scale was for 40 Mel filter banks. Transforming the spectogram's magnitude in decibels (use log scale), power log Mel-spectrogram is generated.

Model Description

Support Vector Machine (SVM) SVM which is a supervised machine learning algorithm is used for regression and classification. It's a straightforward procedure. This is most often used for classification, but in certain instances, it may also be used for regression. SVM is used to find a hyperplane, which acts as a boundary between the various kinds of data. This hyper-plane is just a straight line when seen in 2-D space. Every data point in SVM is plotted in an N-D space, with N denoting the size of the data's features/attributes. Decide on the best hyperplane for splitting the data after that. As a consequence, binary classification is conducted by SVM. Other techniques to handle situations involving many classes, on the other hand, exist. When working with multi-class situations, to a classifier model for each type of data by applying SVM to it. Each classifier's two outputs are as follows: The data point is a member of that category OR There is no such class for the data point in question. To accomplish a multi-class categorization of a category of fruits, for example, to create a classifier model for each fruit in the group. When it comes to the "orange" class, for example, a classifier will determine whether or not something is an orange. The SVM's output is determined by choosing the classifier that has the highest score. The criteria of separable dataset (linear) has worked wonderfully by SVM. Separated data (linear) can be plotted on a 2-D graph space which is segmented by classes through a straight line drawn between the data points.

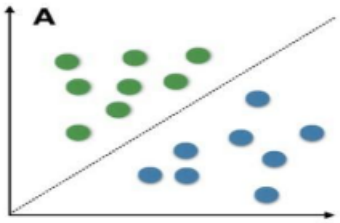


Figure 10: A: Linearly Separable Data

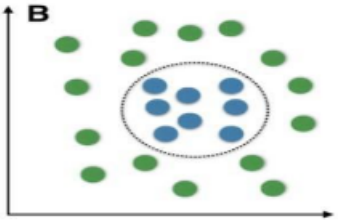


Figure 11: B: Non-Linearly Separable Data.

Multi-Layer Perceptron Classifier (MLP Classifier) MLP is a non-linear neural network of neurons with non-linear input-output mapping called perceptrons. Several interconnected layers of neurons are stacked hierarchically. Input and output layers, as well as one or more hidden layers packed with multiple neurons, make up a Multilayer Perceptron. The MLP network begins with an input layer and continues through hidden layers to an output layer. The network's hidden layers supply the network's computational processing, which results in the network output. In a Perceptron, neurons must use an activation function like ReLU or sigmoid to enforce a threshold, however neurons in a Multilayer Perceptron can use any activation function they like. Because inputs are mixed with initial weights in a weighted sum and then applied to the activation function.Weights are the links between the layers, and they are generally specified as a number between 0 and 1. The Multilayer Perceptron, like the Perceptron, falls into the category of feedforward algorithms. On the other hand, each linear combination is sent on to the next layer. The result of each layer's computation, or internal representation of the data, is passed on to the next. This holds true for both the output and hidden layers. But there's a lot more to it. The algorithm would be unable to learn the weights that maximize the cost function if it just

calculated the weighted sums in each neuron, transferred the results to the next output layer, and then stopped there. If the algorithm just calculated one iteration after another, there would be no learning.

Bidirectional Long Short Term Memory (Bi-LSTM) Any neural network can maintain sequence data in both backward (future to past) and forward (present to future) orientations using bidirectional long-short term memory (b-lstm), giving it additional flexibility in data storage (past to future). The fact that a bidirectional LSTM's input can flow in both directions and so it differentiates the regular LSTM at the same time. Standard LSTM inputs flow in one direction. It can be backward or forward. On the contrary, info can flow in both directions not only in the future but also in the past (same time). Let's look at an example to help you understand the concept. It is impossible to complete the phrase "boys go to..." by filling in the blanks. In any event, we can simply forecast the previously vacant space and have our models apply the same method when we have a future phrase, such as "boys come out of school," and bidirectional LSTM enables the neural network to do this task. BI-LSTM is typically used for activities that need the execution of sequence after sequence. Text classification, speech recognition, and forecasting models are just a few of the applications that could profit from such a system. After that, we'll design a bi-directional LSTM model with Python.

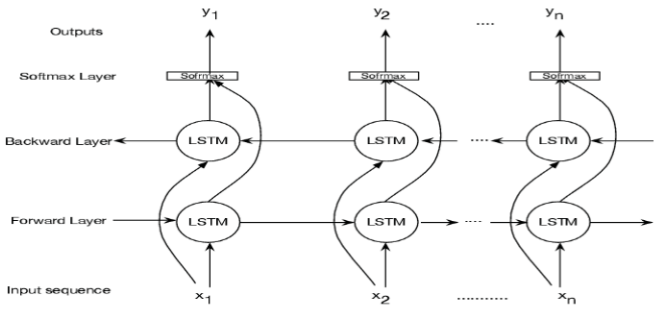


Figure 12: Model Architecture of B-LSTM.

The input is traveling in both the forward and backward layers in the LSTM cells, as seen in fig11, allowing it to record the sequence of input. Speech recognition, text categorization, stock market research, and weather forecasting models all use these types of networks when sequences are required for future predictions.

Proposed Model/ Model Architecture

In our model, we implemented a hybrid deep neural network of Bi-LSTM and dense layers. The model is sequential. Here, we have used B-LSTM as it is better than LSTM and traditional recurrent neural networks. First, we have extracted features (MFCC= 13, Chroma=12, Mel Spectrogram=40 total 75 extracted features) from our 7000 input audio files. Then we split the 7000 input audio files into two sets: train set (5950 audio files) and test(1050 files), since we have extracted 75 features from each audio file, as a result, we make a data frame of dimensions (5950,75: train set) and (1050,75: test set). However, the input for LSTM architecture is a 3D array, that's why we have expanded the dimension of the train and test data frame by 1 and made it (5950,75,1) and (1050,75,1). The timeseries_size used for the inputs of the B-LSTM network is 75 as we have extracted 75 features and the encoding used was 1. After that, the train and test data frame is passed through the two layers of B-LSTM, the batch size used in B-LSTM layers is 256, then the fully connected layer (two dense) is given an input by the output of the next B-LSTM layer, the activation function used in the fully connected dense layers is ReLu, here we have used ReLu because of the training productivity and faster merging. Finally, the output of the fully connected dense layer is passed into the output layer of the softmax activation function. The dropout used in our model is 20%. Lastly, we have compiled our model, for training the model using the optimizer and the loss function. The optimizer we have used for compiling is Adam. The input and output of each layer in our model is shown in figure12. The total number of parameters used in our model is 810,042. The architecture of our model is shown in figure 13.

Layer (type)	Output Shape	Param #
bidirectional (Bidirectional	(None, 75, 306)	189720
bidirectional_1 (Bidirection	(None, 306)	563040
dense (Dense)	(None, 153)	46971
dropout (Dropout)	(None, 153)	0
dense_1 (Dense)	(None, 64)	9856
dropout_1 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 7)	455

Total params: 810,042
Trainable params: 810,042
Non-trainable params: 0

Figure 13: The total number of input-output parameters used in our proposed model.

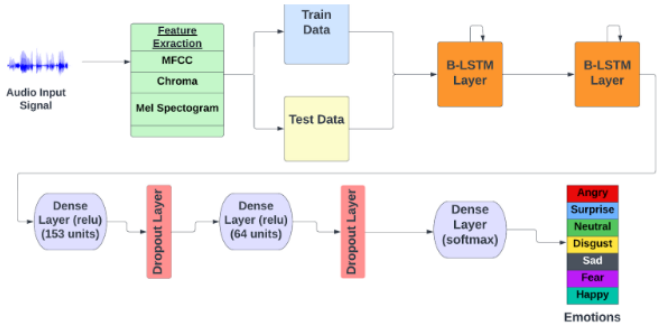


Figure 14: The model architecture of the hybrid deep B-LSTM network.

RESULT AND ANALYSIS

Experimental Evaluation

It was possible to assess the success of the classifying job on two different levels: overall precision and class precision.

We tallied up accuracy and average F1 scores to get a sense of overall performance. The recall, precision, and confusion matrices were utilized to report class-wise accuracy.. For us to measure how good our model is, here are the matrices we use and how they work in more detail:

Weighted accuracy: It assigns a numerical value to every class based on the number of right predictions. The total of all the weighted classes is equal to the following:

$$WA = \sum_{i} \frac{correct_i}{instance_i} \tag{5}$$

Sensitivity or Recall: The fraction of correctly identified cases of total occurrences of that specific class in the dataset is called recall. The model's ability to recognize Positive samples is measured by the recall. It is also known as the True Positive rate(TPR) and it can be calculated in the following way:

$$recall = \frac{truepositive}{truepositive + falsenegative} \tag{6}$$

Precision: It shows the current amount of correction. Precision is evaluated as a fraction of positive samples (corrected ones) which is then divided by the number of negative samples added to positive samples.It is also known as the PPV (positive predictive value) and it can be calculated in the following manner:

$$precision = \frac{truepositive}{truepositive + falsenegative} \tag{7}$$

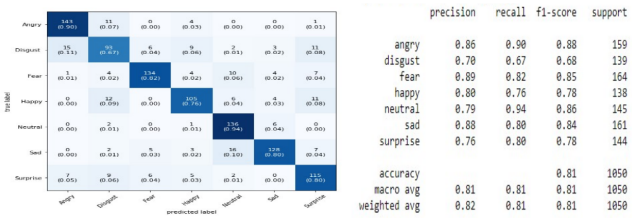
The F1 score is described as,

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{8}$$

Confusion Matrix: A confusion matrix is a table that shows how well a categorization method performs. The efficiency of a classification algorithm is shown and summarized using a confusion matrix.

Analysis of Results

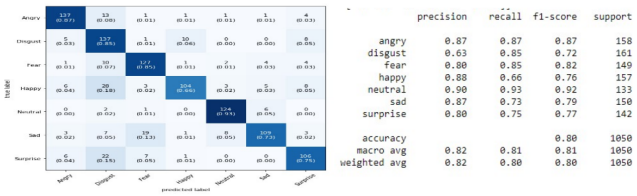
SVM(Support Vector Machine) The overall accuracy of the SVM model is quite satisfactory. We use different support cases for different emotions. For example, we use 159 support cases for anger,164 for fear, 138 for happiness, and so on. The SVM model shows 90% right prediction for angry emotion, 82% right prediction for fear, and 76% for happy. We evaluate the SVM model in the SUBESCO dataset. The overall accuracy we get from the SVM model is 81.33%. From this SVM model, neutral emotion shows the best performance. We use 145 support cases for neutral emotion, where 136 emotions are being successfully detected and the right emotion detection percentage is over 94%. On the other hand, disgust performs the worst for emotion detection in the SVM model. We use over 139 support cases for disgust emotion, but only 93 cases are being detected by the SVM model and the emotion detection percentage is only 67%. But in general SVM model perform better in our targeted voice dataset to find emotion.



	precision	recall	f1-score	support
angry	0.86	0.90	0.88	159
disgust	0.70	0.67	0.68	139
fear	0.89	0.82	0.85	164
happy	0.80	0.76	0.78	138
neutral	0.79	0.94	0.86	145
sad	0.88	0.80	0.84	161
surprise	0.76	0.80	0.78	144
accuracy			0.81	1050
macro avg	0.81	0.81	0.81	1050
weighted avg	0.82	0.81	0.81	1050

Figure 15: Confusion Matrix and Classification report of SVM Classifier.

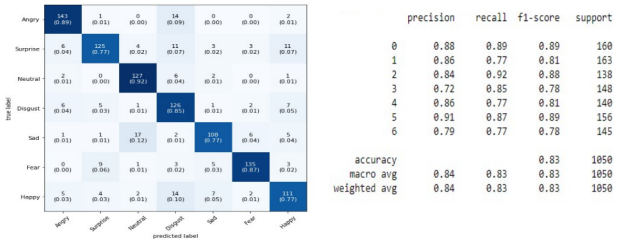
MLP Classifier Performance evaluation of the MLPC (Multilayer Perceptron Classifier) which is a class of ANN (artificial neural networks) explained over prepared dataset in this category. The total number of support cases is 1050 from which MLPC is later distributed among 7 emotions. Our MLPC model used hidden layers for better accuracy. The SUBESCO dataset obtained an average accuracy of 80.38 percent from MLPC. Overall, MLPC performed well in anger, disgust, fear, and neutral emotions, with the accuracy of 87 %, 85 %, 85 %, and 93 %, consecutively. However, it performed poorly on happy (66%), sad (73 %), and surprise (75%). Our MLPC model performed the best on neutral emotion, with a 93% accuracy rate. However, MLPC didn't do great when it came to predicting happiness, with just 66% accuracy. Only 104 out of 157 support cases were identified by MLPC for the happy emotion, whereas 124 out of 133 support cases were identified for the neutral emotion.



	precision	recall	f1-score	support
angry	0.87	0.87	0.87	158
disgust	0.63	0.85	0.72	161
fear	0.80	0.85	0.82	149
happy	0.88	0.66	0.76	157
neutral	0.90	0.93	0.92	133
sad	0.87	0.73	0.79	150
surprise	0.80	0.75	0.77	142
accuracy			0.80	1050
macro avg	0.82	0.81	0.81	1050
weighted avg	0.82	0.80	0.80	1050

Figure 16: Confusion Matrix and Classification report of MLP Classifier.

Our Proposed Model(B-LSTM) The Bi-LSTM model is designed by us. Here if we see the best emotion detection according to percentage is Neutral. The 127 neutral files were detected correctly whereas 138 files were in total. The lowest accuracy we find here is surprising, sad, and happy. In these aspects, 77% of accurate results were found here. Nevertheless, numerous numbers are there as a surprising total of 125, sad 108, happy 111 were correctly identified. If we look at the most successful criteria in our model which is angry in our case 14 samples were detected as disgust, 1 as a surprise, 2 as happy, but they all should be identified as angry. So, it is a very simple thing that no matter how much we train our model we will always find a little ambiguity in the result of detecting models.



	precision	recall	f1-score	support
0	0.88	0.89	0.89	160
1	0.86	0.77	0.81	163
2	0.84	0.92	0.88	138
3	0.72	0.85	0.78	148
4	0.86	0.77	0.81	140
5	0.91	0.87	0.89	156
6	0.79	0.77	0.78	145
accuracy			0.83	1050
macro avg	0.84	0.83	0.83	1050
weighted avg	0.84	0.83	0.83	1050

Figure 17: Confusion Matrix and Classification report of hybrid deep-Bi-LSTM network.

From the Epochs curve, we see that 256 samples were tested and if we observe the learning curve(Model Accuracy, Model Loss) no smooth direct thing is observed here. We see that model accuracy is constantly increasing and model loss is constantly decreasing.

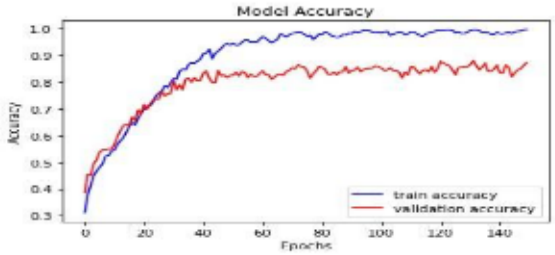


Figure 18: Learning Curve (Model Accuracy) of the Proposed Model.

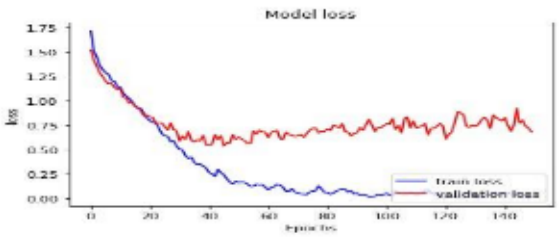


Figure 19: Learning Curve (Model Loss) of the Proposed Model.

Comparison among Models

In this section, we compared precision, f1-score, and recall for all three models (SVM, B-LSTM, MLP).

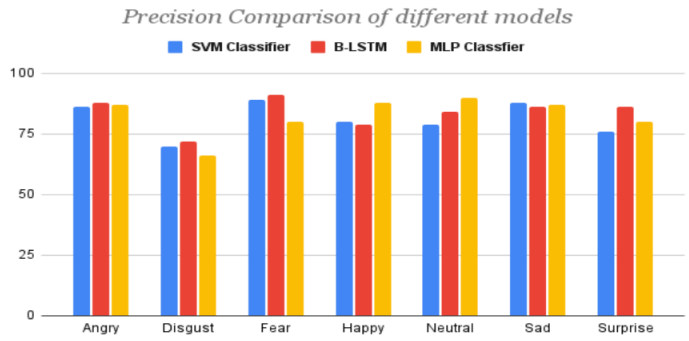


Figure 20: Precision comparison of different models.

Precision is defined as the quality or state of being accurate. We can observe that our model B-LSTM performed much better than the other models in general. In the case of angry emotion, all three models scored well (B-LSTM 88%,MLP 87%,SVM 86%), however in the case of the disgust emotion, all three models scored poorly (B-LSTM 72%,MLP %,SVM 70%) . In anger (86%), fear (89%), and sad (88%) emotions, SVM precision was excellent. However, our model (B-LSTM) outperforms both SVM and MLP in terms of anger (88%), disgust (72%), fear (91%), and surprise (86%). We can also observe that MLP outperformed both SVM and B-LSTM in happy (88%) and neutral (90%) feelings.

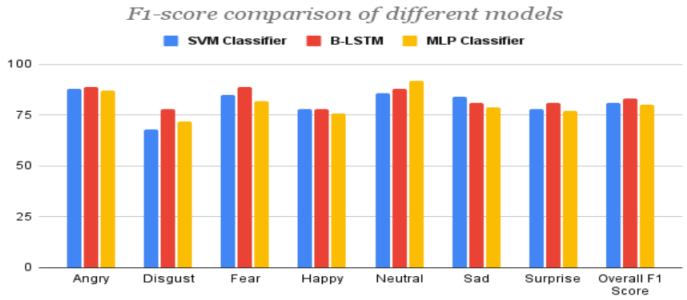


Figure 21: F1-Score comparison of different models.

The F1 score is the mean of precision and recall. It's a statistical metric for evaluating performance. Our data show that the total f1 score for all three models performed admirably and was almost similar. However, when compared to our model(B-LSTM), the SVM and MLP scored poorly on disgust (B-LSTM 78%,MLP 72%,SVM 68%). Nevertheless, our f1-score model outperformed the SVM and MLP by a little percentage.

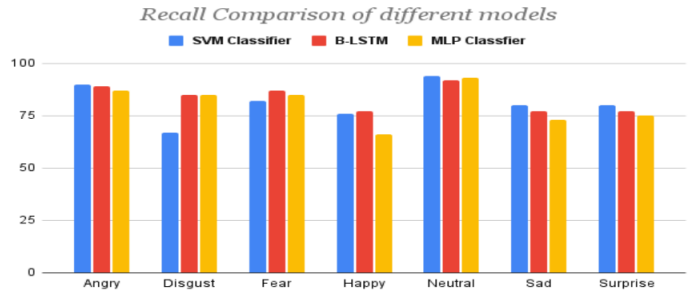


Figure 22: Recall comparison of different models.

The term "recall" refers to a quantitative measurement. We can observe that the recall score of the SVM was low (67%) on the disgust emotion, whereas the recall score of the other models was high (B-LSTM 85%, MLP 85%). MLP also received a low score(66%) in the happy emotion. Aside from that, all of the models performed fairly identically. We can also observe that the recall score for neutral is the highest (B-LSTM 92%,MLP 93%,SVM 94%) for all three models, but the recall score for happy is the lowest (B-LSTM 77%,MLP 66%,SVM 76%). Nevertheless, when compared to SVM and MLP, our model (B-LSTM) has a good overall recall score.

Model	Accuracy
SVM Classifier	81.33
MLP Classifier	80.38
B-LSTM	83.33

Table 1: Accuracy comparison between different models.

In table1, we can see that our proposed hybrid deep-B-LSTM model gave higher accuracy than the traditional MLP Classifier and SVM Classifier. Therefore, we can say that our hybrid deep Bi-LSTM model is the best model and outperformed these traditional methods.

CONCLUSION AND FUTURE WORK

To conclude, we found seven distinct emotional states in Bengali speech by experimenting with two different machine learning techniques: A hybrid deep learning network employing B-LSTM, as well as MLP and SVM classifiers. This was done by extracting features from the SUBESCO dataset's audio files using MFCC, Chroma, and Mel Spectrogram. In our investigation, we discovered that identifying joyful emotions surpassed all other emotions on the following parameters: weighted accuracy, recall, precision, and F1 scores for each model. The overall accuracy obtained in our experiment was 83.33, 81.33, and 80.38 from our model, SVM classifier, and MLP classifier. Since, the quality of recorded audio files mostly depends on loudness, pitch, frequency, audibility, background noise, and the external machines (headphones, microphones, and environment) used by the actors, improving work accuracy, we might extract the feature more specifically and used a more complex model of deep learning method on the SUBESCO dataset. Similarly, we have implemented our model on real time-audio files and observed the result of it.

After developing each of the models, we collected some real-time audio files and tried to detect the emotion in those files. Initially, those models were performing very poorly, however, after some tuning the results were improving. This could be a scope of future work related to this study. We might train the model with a larger trained dataset so that the model can work on the larger range of the audio file's loudness, pitch, frequency, audibility, and so on. Furthermore, we have already seen that there are some audio features that are not used in this study such as delta-MFCC, GFCC, etc. So another scope could be to study and apply those audio features and try to develop a more robust model that performs really well to detect the emotion of any real-time audio files.

References

Amin, R.; and Rahman, Z. 2015. *Bangladeshi Dialect Recognition using MFCC, Delta, Delta-delta and GMM*. Ph.D. thesis, East West University.

Badhon, S.; Rahaman, H.; Rupon, F. R.; and Abujar, S. 2021. Bengali accent classification from speech using different machine learning and deep learning techniques. In *Soft Computing Techniques and Applications*, 503–513. Springer.

Baroi, O. L.; Kabir, M. S. A.; Niaz, A.; Islam, M. J.; Rahimi, M. J.; et al. 2019. Effects of Different coefficients on MFCC and PLP for Bangla Speech Corpus using Tied-state Triphone Model. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 1–6. IEEE.

Bouziane, A.; Kharroubi, J.; and Zarghili, A. 2021. Towards an objective comparison of feature extraction techniques for automatic speaker recognition systems. *Bulletin of Electrical Engineering and Informatics* 10(1): 374–382.

Das, R. K.; Islam, N.; Ahmed, M. R.; Islam, S.; Shatabda, S.; and Islam, A. M. 2022. BanglaSER: A speech emotion recognition dataset for the Bangla language. *Data in Brief* 42: 108091.

Hasan, H. M.; and Islam, M. A. 2020. Emotion recognition from bengali speech using RNN modulation-based categorization. In *2020 third international conference on smart systems and inventive technology (ICSSIT)*, 1131–1136. IEEE.

Iqbal, A.; and Barua, K. 2019. A real-time emotion recognition from speech using gradient boosting. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 1–5. IEEE.

Kattel, M.; Nepal, A.; Shah, A.; and Shrestha, D. 2019. Chroma feature extraction. In *Conference: Chroma Feature Extraction using Fourier Transform*, 20.

Livingstone, S. R.; and Russo, F. A. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one* 13(5): e0196391.

Sejdić, E.; Djurović, I.; and Jiang, J. 2009. Time–frequency feature representation using energy concentration: An overview of recent advances. *Digital signal processing* 19(1): 153–183.

Shao, Y.; Srinivasan, S.; and Wang, D. 2007. Incorporating auditory feature uncertainties in robust speaker identification. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, IV–277. IEEE.

Shao, Y.; and Wang, D. 2008. Robust speaker identification using auditory features and computational auditory scene analysis. In *2008 IEEE international conference on acoustics, speech and signal processing*, 1589–1592. IEEE.

Sharma, S.; and Singh, P. 2014. Speech emotion recognition using GFCC and BPNN. *International Journal of Engineering Trends and Technology (IJETT)* 18(6): 321–322.

Singh, A.; and Ghangas, S. ???? SPEAKER RECOGNITION USING MFCC AND DELTA-DELTA MFCC AND CLASSIFICATION USING ARTIFICIAL NEURAL NETWORK .

Sultana, S.; Iqbal, M. Z.; Selim, M. R.; Rashid, M. M.; and Rahman, M. S. 2021a. Bangla Speech Emotion Recognition and Cross-Lingual Study Using Deep CNN and BLSTM Networks. *IEEE Access* 10: 564–578.

Sultana, S.; Rahman, M. S.; Selim, M. R.; and Iqbal, M. Z. 2021b. SUST Bangla Emotional Speech Corpus (SUBESCO): An audio-only emotional speech corpus for Bangla. *Plos one* 16(4): e0250173.

Wang, J.; Xue, M.; Culhane, R.; Diao, E.; Ding, J.; and Tarokh, V. 2020. Speech emotion recognition with dual-sequence LSTM architecture. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6474–6478. IEEE.

Wu, S.; Falk, T. H.; and Chan, W.-Y. 2011. Automatic speech emotion recognition using modulation spectral features. *Speech communication* 53(5): 768–785.

Zhao, X.; Shao, Y.; and Wang, D. 2012. CASA-based robust speaker identification. *IEEE Transactions on Audio, Speech, and Language Processing* 20(5): 1608–1616.