BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY
DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING

EEE 312 (January 2025)

Full Title of the Course Laboratory

Final Project Report

Section: C2

Speech Emotion Recognition using Machine Learning

Video Link: *Click to go to Video*

Github Link: *Click to go to Github*

Course Instructors:

1. **Shafin Bin Hamid, Assistant Professor, Department of EEE, BUET**
2. **Rafid Hassan Palash, Lecturer, Department of EEE, BUET.**

Signature of Instructor: _____

Academic Honesty Statement:

IMPORTANT! Please carefully read and sign the Academic Honesty Statement, below. Type the student ID and name, and put your signature. *You will not receive credit for this project experiment unless this statement is signed in the presence of your lab instructor.*

"In signing this statement, We hereby certify that the work on this project is our own and that we have not copied the work of any other students (past or present), and cited all relevant sources while completing this project. We understand that if we fail to honor this agreement, We will each receive a score of ZERO for this project and be subject to failure of this course."

Full Name: Md. Ahsanul Karim **Student ID:** 2106172	**Full Name:** Ratul Das **Student ID:** 2106176
Full Name: Saumik Banik **Student ID:** 2106179	**Full Name:** Mushfiqul Amin **Student ID:** 2106194

Table of Contents

1 Abstract

This project presents a method for speech emotion recognition in both English and Bangla by extracting features from speech samples and comparing them with a feature database. The approach utilizes a combination of machine learning algorithms to classify five distinct emotional states: happy, sad, angry, surprised, and natural for Bangla and six distinct emotional states: angry, fearful, happy, natural, sad and surprised. The key features extracted include MFCC, chromagram, Mel-spectrogram, Spectral roll-off, and Spectral centroid. To enhance the details of the spectral features, the Teager Energy Operator (TEO) and Higher Order Statistics is applied, improving the algorithm's accuracy. Simulation results demonstrate that the proposed method outperforms existing speech emotion recognition techniques.

2 Introduction

Audio sentiment recognition has emerged as a promising research field in recent years. Accurately detecting sentiments through speech is not only valuable for understanding human emotions but also holds significant potential for various practical applications. One such application is in marketing, where customer sentiments greatly influence the market success of products. While sentiment detection via text data is widely explored, the vast amount of unused audio data offers new opportunities for emotion detection through voice alone, paving the way for a range of innovative AI-based applications.

The rise of social media has led to a concerning increase in the use of audio threats, making threat analysis a critical area of focus. Detecting threats through speech signals can significantly contribute to enhancing cyber security. Additionally, in customer service contexts such as call centers, emotion-aware systems could play calming music or adjust their tone when detecting anger or panic in a caller's voice. Another key application is in psychological evaluation, where detecting depression through audio sentiment analysis is a growing area of research. Depressed individuals often speak tensely or uncomfortably, using brief verbal expressions with a discouraged or exhausted tone, which can be identified through advanced analysis of speech patterns.[1]

Despite the advancements in voice-based sentiment recognition in languages like English, there is a notable scarcity of resources for Bengali language audio emotion recognition research. To address this gap, we have also developed our own Bengali speech-based dataset, encompassing five distinct emotion classes: Happy, Sad, Angry, Surprised and Natural. Speaker-independent emotion recognition presents a significant challenge due to the variability in speech patterns across individuals.[2]

Our sentiment detection model revolves around two critical phases: identifying effective audio features and constructing an appropriate mathematical model. In the initial phase, we explore six different audio features, which are then evaluated for their effectiveness across various machine learning (ML) models. Based on this analysis, we identify the best combination of features and models to optimize the accuracy and reliability of our sentiment detection system.

3 Design

3.1 About Dataset

3.1.1 English Dataset

For this project, we utilized the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS**)**, a widely used and publicly available dataset. The dataset consists of high-quality speech-only audio files in 16-bit, 48kHz WAV format, carefully curated for research in emotion recognition.

This portion of the RAVDESS includes 1,440 audio files, generated by 24 professional actors (12 male and 12 female). Each actor recorded 60 speech samples, consisting of two lexically-matched statements spoken in a neutral North American accent. The dataset , we have worked on,has 1152 samples among the RAVDESS and covers six emotional expressions: natural, happy, sad, angry, fearful, and surprise, with each emotion performed at two levels of intensity (normal and strong), in addition to a neutral expression.

Catagories	Number
Happy	192
Sad	192
Angry	192
Natural	192
Fearful	192
Surprised	192

3.1.2 Bangla Dataset

The BanglaSER dataset is a specialized resource developed to support research in Bangla Speech Emotion Recognition (SER**)**. It consists of a diverse set of 1,467 speech audio recordings, capturing a range of fundamental human emotions expressed in the Bangla language. A total of 34 native speakers—comprising 17 male and 17 female participants, aged between 19 and 47 years—contributed to the recordings.

The dataset covers five emotional states: angry, happy, sad, surprise, and natural. For each non-neutral emotion, participants were asked to speak three distinct statements, each repeated three times, resulting in 1,224 recordings (3 statements × 3 repetitions × 34 speakers). For the neutral emotion, a slightly smaller group of 27 speakers participated, producing 243 recordings under the same structure. This consistent and well-balanced recording framework makes BanglaSER highly suitable for training and evaluating deep learning models aimed at recognizing emotions in Bangla speech.

Catagories	Number
Happy	**216**
Sad	**216**
Angry	**216**
Surprised	**216**
Natural	**216**

3.2 Feature Extraction

3.2.1 MFCC

Full form of MFCC is Mel Frequency Cepstral Coefficient. MFCC is mainly used in speech recognition and in speaker or speaker recognition. Parameters in MFCC are well suited to the speech or audio signals. [5] They come from the following hypothesis:

$$x_n = g_n * b_n$$

where, x_n is the signal, g_n is the input and b_n is the filter characterizing the path. It mainly indicates that the speech signal is the convolution between a filter (vocal tract) and excitation (vocal cords). A homomorphic transformation makes it possible to transform this product into a sum. The sum is then filtered and we get MFCC features. MFCCs allow to a deconvolution between the source of the sound produced (characteristics of the speaker) and the oral duct (whether or not coupled to the nasal duct)

$$\widetilde{x_n} = \widetilde{g_n} + \widetilde{b_n}$$

The homomorphic transformation breaks down into three main stages [6]:

- A passage in the spectral domain by calculation of the modulus of the Fourier Transform
- An application of the algorithm
- Return to the time domain by calculation of inverse FFT

An accentuation of the treble is present because the high frequency the high frequency components are always weaker than the bass. So, a simple high pass filtering is made. Transfer function of the filter is given below:

$$H(z) = 1 - 0.98 \times z^{-1}$$

The Mel nonlinear scale is known to account for human perception. The coefficients are called MFCC because in spectral domain the change of scale is carried out by us ing perceptual Mel scale. They have the property of being strongly decorrelated. In speech recognition systems, the first coefficient is often used to define energy. Usually, a cepstral subtraction is done on MFCCs to deconvolve the signal from the channel noise and obtain a parametrically denoised signal [6]. The noise comes from different sources, like recording sources as microphone, telephone channel etc. This operation results from the fact that the cepstral coefficients of any speech have zero mean. To remove the noise, it suffices to subtract at each cepstral coefficient of the noisy signal by their average, representative of the

average of cepstral coefficients relating to noise alone. The number of MFCCs varies between 12 to 40. In this project we used 18 coefficients. Flow chart diagram for MFCC is given below:

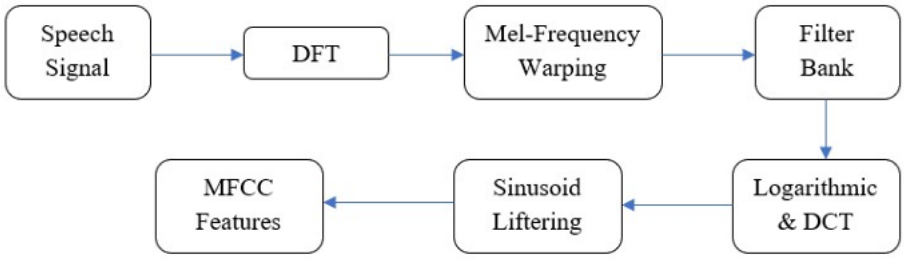


Figure : Flow Chart of MFCC

3.2.2 Chromagram

The main idea of chroma features is to aggregate all spectral information that relates to a given pitch class into a single coefficient. Given a pitch-based log-frequency spectrogram $Y_{\mathrm{LF}}: \mathrm{Z} \times [0:127] \rightarrow R_0$ as defined in [7], a chroma representation or chromagram $\mathrm{Z} \times [0:11] \rightarrow R_0$ can be derived by summing up all pitch coefficients that belong to the same chroma :

$$C(m,c) = \sum_{(p \in [0:127] \mid p \bmod 12 = c)} Y_{LF}(m,p)$$

For C in the range of [0:11]

3.2.3 Mel-Spectrogram

In this algorithm, the audio input is first buffered into frames of number of samples. The frames are overlapped. The specified Window is applied to each frame, and then the frame is converted to frequency-domain representation with FFT . Each frame of the frequency-domain representation passes through a mel filter bank. The spectral values output from the mel filter bank are summed, and then the channels are concatenated. The summer output is Melspectrogram.

3.2.4 Spectral Roll-off

Spectral roll off point is the cutoff frequency below which 95% of the power of a signal is concentrated. Spectral roll off is higher for audio files with high frequencies (mainly unvoiced data with sounds) than audio files with voiced data (energy concentrated in mainly low frequency regions). That's why, this feature enables us to find the voiced and unvoiced alternations of speech [3] .

In our project, we used frame length for FFT as 1024 and the hop length was 512. As spectral centroid, we took mean value of the whole data for each audio files after normalizing by their corresponding magnitudes as our feature. The spectral rolloff point measures the bandwidth of the audio signal by determining the frequency bin under which a

given percentage of the total energy exists [4]: The spectral rolloff point is calculated as i, such that:

$$\sum_{k=b1}^{i} s_k = \boldsymbol{\kappa} \sum_{\boldsymbol{k=b1}}^{\boldsymbol{b2}} \boldsymbol{s_k}$$

where,

- s_k is the spectral value at bin k
- b_1 and b_2 are the band edges, in bins, over which to calculate the spectral spread.
- $\boldsymbol{\kappa}$ is the percentage of total energy contained between b_1 and i.

3.2.5 Spectral Centroid

Full form of PSD is Power Spectral Density. PSD of a signal can be calculated by taking Fourier Transform of the autocorrelation of the signal. The spectral centroid is the frequency center of gravity of PSD of a signal. Mathematically, we can represent as[3] :

$$C(i) = \frac{\sum_{n=1}^{N} w_n . S_i(w_n)}{\sum_{n=1}^{N} S_i(w_n)}$$

3.3 Higher Order Statistics

3.3.1 Skewness and Kurtosis

Higher order statistics refers to the use of higher power order of moments and cumulants. In machine and deep learning approach, such higher order can help the model to be insensitive to outliers while learning the useful features which could help the output prediction.[12]

In the proposed method, third and fourth order moments, skewness and kurtosis respectively are applied. Skewness measures the symmetry of a random variable or more precisely, the deviation of the variable's distribution from the normal distribution. The given distribution might be skewed either to left or to right. It is measured as :

$$S = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^3}{(N-1)\sigma^3}$$

Here $\bar{X}$ is the mean of the random variable X, N is the total number of data and σ is the standard deviation.

Kurtosis is fourth order moment and a measurement of the tailedness of a distribution. It can exhibit whether tail data of a given distribution are more or less extreme than the normal distribution. It can be measured as

$$K = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^4}{(N-1)\sigma^4}$$

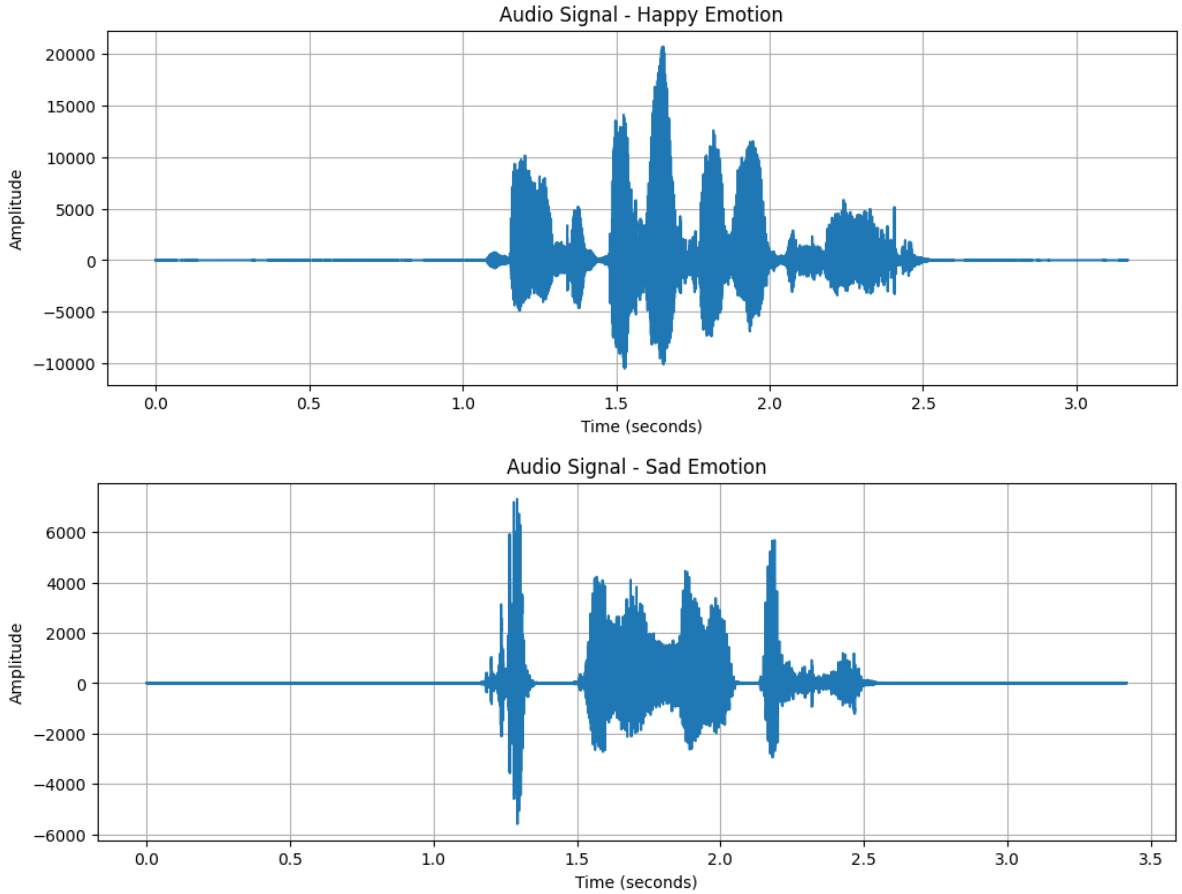
Here $\bar{X}$ is the mean of the random variable X, N is the total number of data and $\boldsymbol{\sigma}$ is the standard deviation.

3.3.2 Teager Energy operation (TEO)

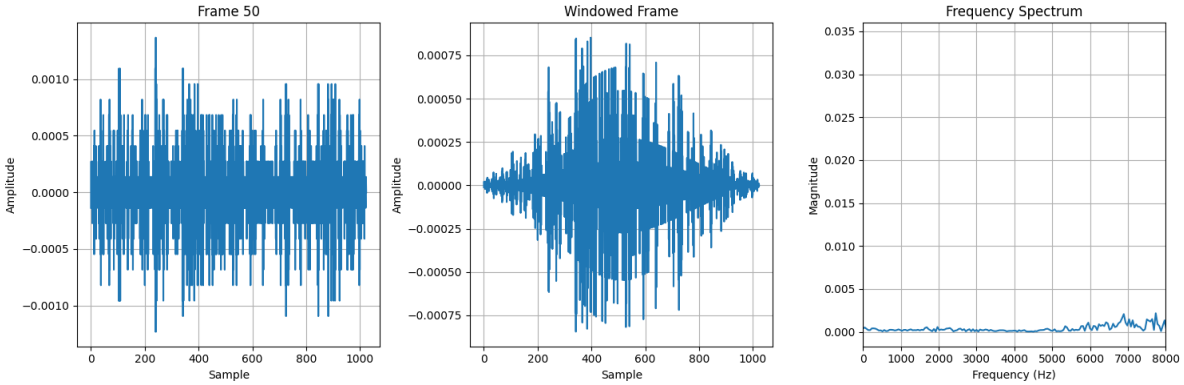
If the amplitudes of a feature is quite small, Teager energy operation (TEO) is performed on it to enhance feature details.

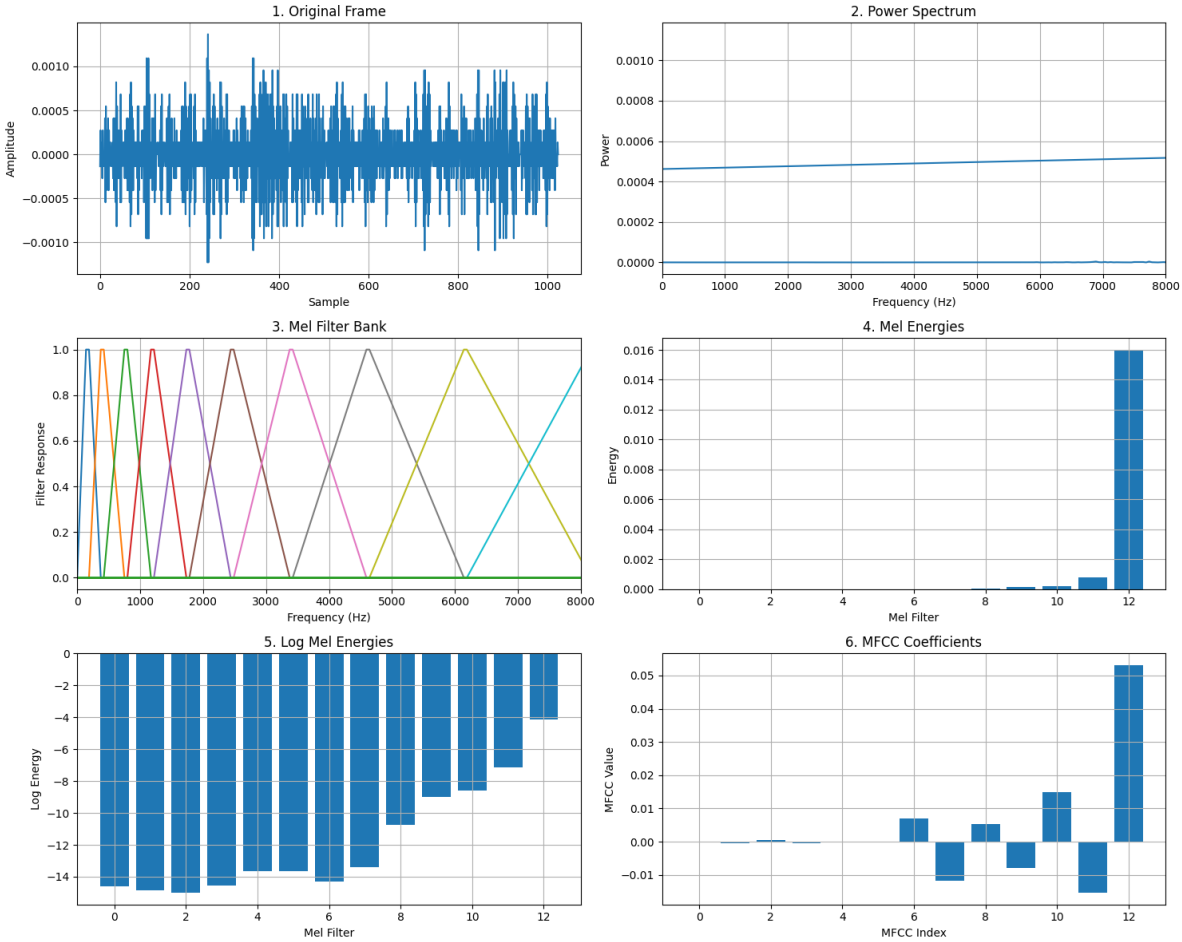
$$TE(x[n]) = x[n]^2 - x[n-1]x[n+1]$$

Here a band-limited discrete signal x[n] is approximated by TE operator [13]. A Teager Energy Operated feature enhances detailing and approximation.



Now we'll visualize sample features-





4 Implementation

4.1 Used Machine Learning Models

After all the necessary feature extraction, the dataset was run on different machine learning models. In our model we have implemented four machine learning models. They are-

- K-Nearest Neighbours Classifier
- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Classifier

No neural network models have been implemented on the dataset.
K-Nearest Neighbours Classifier:
A non-parametric, instance-based learning technique called K-Nearest Neighbors uses a distance measure to classify a data point according to the majority class of its k nearest neighbors in the feature space. It is easy to use and intuitive, but for large datasets, it can be computationally costly. The most common metric is Euclidean distance, defined as:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Here x *and* y are datapoints and n is the number of features.

Decision Tree Classifier:
To classify data, the Decision Tree Classifier creates a model of decisions that resembles a tree using feature splits. Class labels are represented by leaves, and each node denotes a feature condition. Splits are selected to optimize information gain, which is frequently determined by entropy:

$$Entropy(S) = -\sum_{i=1}^{c} p_i \log_2(p_i)$$

Where (p_i) is the proportion of class (i) in set (S), and (c) is the number of class. Although it can be interpreted, without trimming, it is prone to overfitting.

Random Forest Classifier:

Using random selections of features and data, the Random Forest Classifier is an ensemble technique that builds several decision trees and combines predictions via majority voting. It increases robustness and decreases overfitting. The average reduction in impurity can be used to determine the significance of a feature. For high-dimensional data, it is flexible.
Support Vector Classifier:

Using support vectors, the Support Vector Classifier maximizes the margin between classes to determine the best hyperplane for class separation. The decision function for a linear SVC is

$$f(x) = w^T x + b$$

where the sign of $(f(x))$ determines classification, (w) is the weight vector, and (b) is the bias. Non-linear data is handled with kernel functions (like RBF). Although it is sensitive to parameter adjustment, SVC works well in high-dimensional spaces.

5 Design Analysis and Evaluation

5.1 Design Considerations

Both English and Bangla dataset was used in this project. In order to implement rigorous analysis, we've implemented several feature extraction processes and then used simple machine learning algorithms to train and classify from the dataset. Total dataset was split into three parts. 70% training data, 20% test data and 10% for validation.

5.2 Investigation and Result Analysis

5.2.1 Results for Bangla Dataset

In the Bangla SER dataset acquired from Kaggle, there were total 1467 audio.wav file as data. After carefully extract the features, dataset was run of four different machine learning models. Each one showed accuracy around 50% and KNN algorithm showed highest accuracy among them, marking 66.2% when run with k=5. When decreasing the value of k, better accuracy was noted. Highest accuracy was got using k=1, which was 78.24%. ML classifier's performance are sorted in the table below-

Table I- ML Classifiers' Performance for Bangla Dataset

Classifier	**Accuracy for Bangla Dataset**
Decision Tree Classifier	53.24%
Random Forest Classifier	53.7%
Support Vector Classifier	50%
K Nearest Neighbours Classifier	
k=1	78.24%
k=3	71.3%
k=5	66.2%
k=7	65.75%
k=9	60.65%

The confusion matrix and ROC curves are attached here for a better visualization of the accuracy.

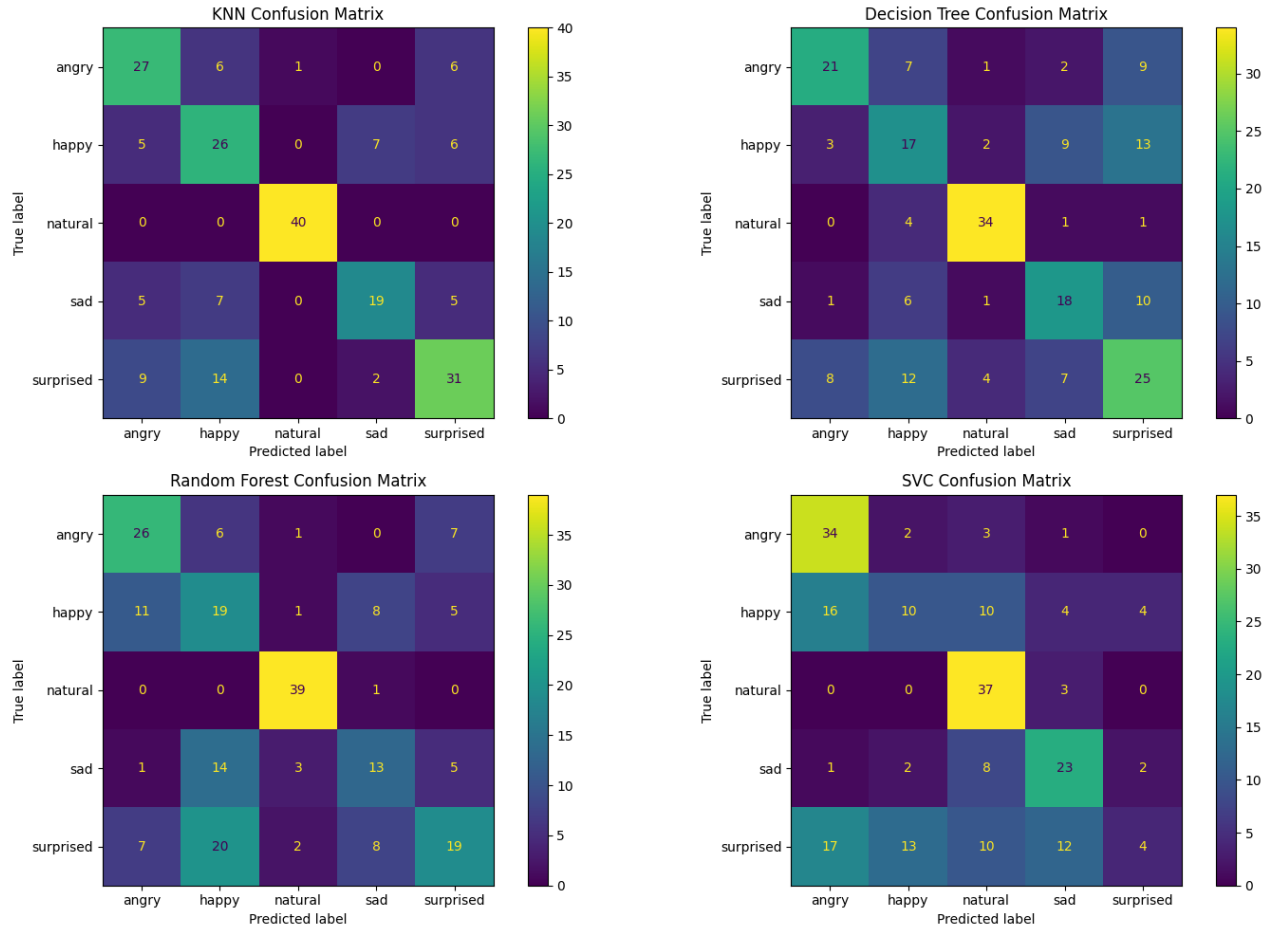


figure: Confusion Matrix

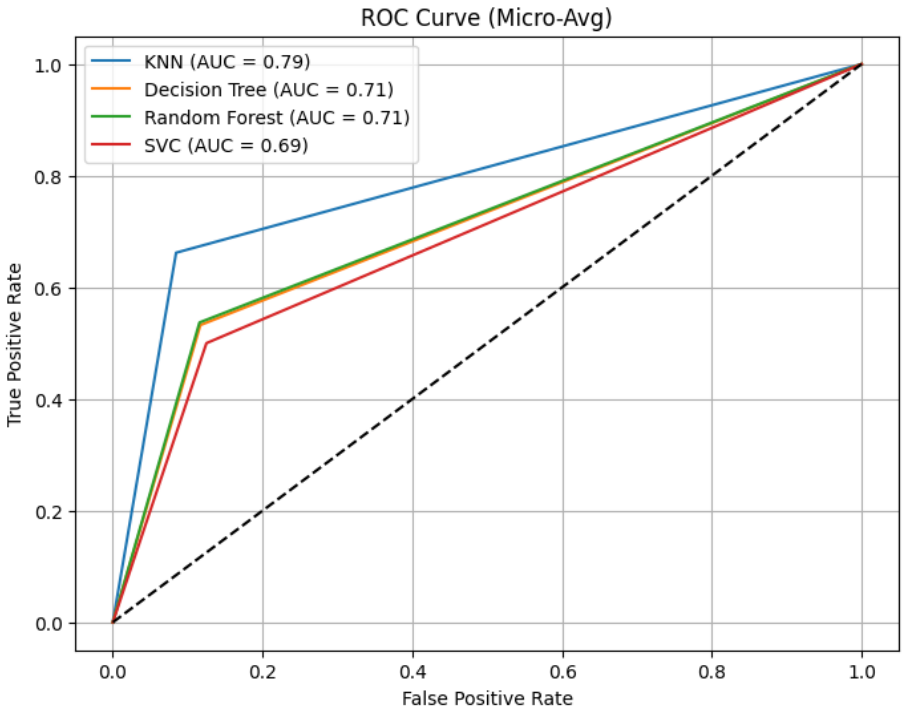


Figure: ROC Curve

Table II- Class wise metrics for Bangla Dataset

Class	Precision	Recall	F-1 Score	Support
Angry	0.667	0.8	0.7273	40 samples
Happy	0.7333	0.75	0.7416	44 samples
Natural	1	1	1	40 samples
Sad	0.7	0.5833	0.6364	36 samples
Surprised	0.8113	0.7679	0.789	56 samples

5.2.2 Results for English Dataset

For English speech emotion recognition Ravdess dataset was used which was acquired from Kaggle. We preprocessed total 1152 audio.wav file at six different emotion class. After carefully extracting the features, dataset was run of four different machine learning models. Decision Tree and Random Forest Classifier showed an accuracy around 40% whereas KNN and SVC model showed accuracy around 50%. KNN marked the highest accuracy (53.25%) among them with k=5. However, when decreasing the value of k, better accuracy was noted. Highest accuracy was got using k=1, which was 58%. ML classifier's performance are sorted in the table below-

Table III- ML Classifiers' Performance for English Dataset

Classifier	Accuracy for English Dataset
Decision Tree Classifier	39.96%
Random Forest Classifier	43.29%
Support Vector Classifier	50%
K Nearest Neighbours Classifier	
k=1	58.01%
k=3	54.11%
k=5	53.25%
k=7	50.22%
k=9	53.25%

The confusion matrix and ROC curves are attached here for a better visualization of the accuracy.

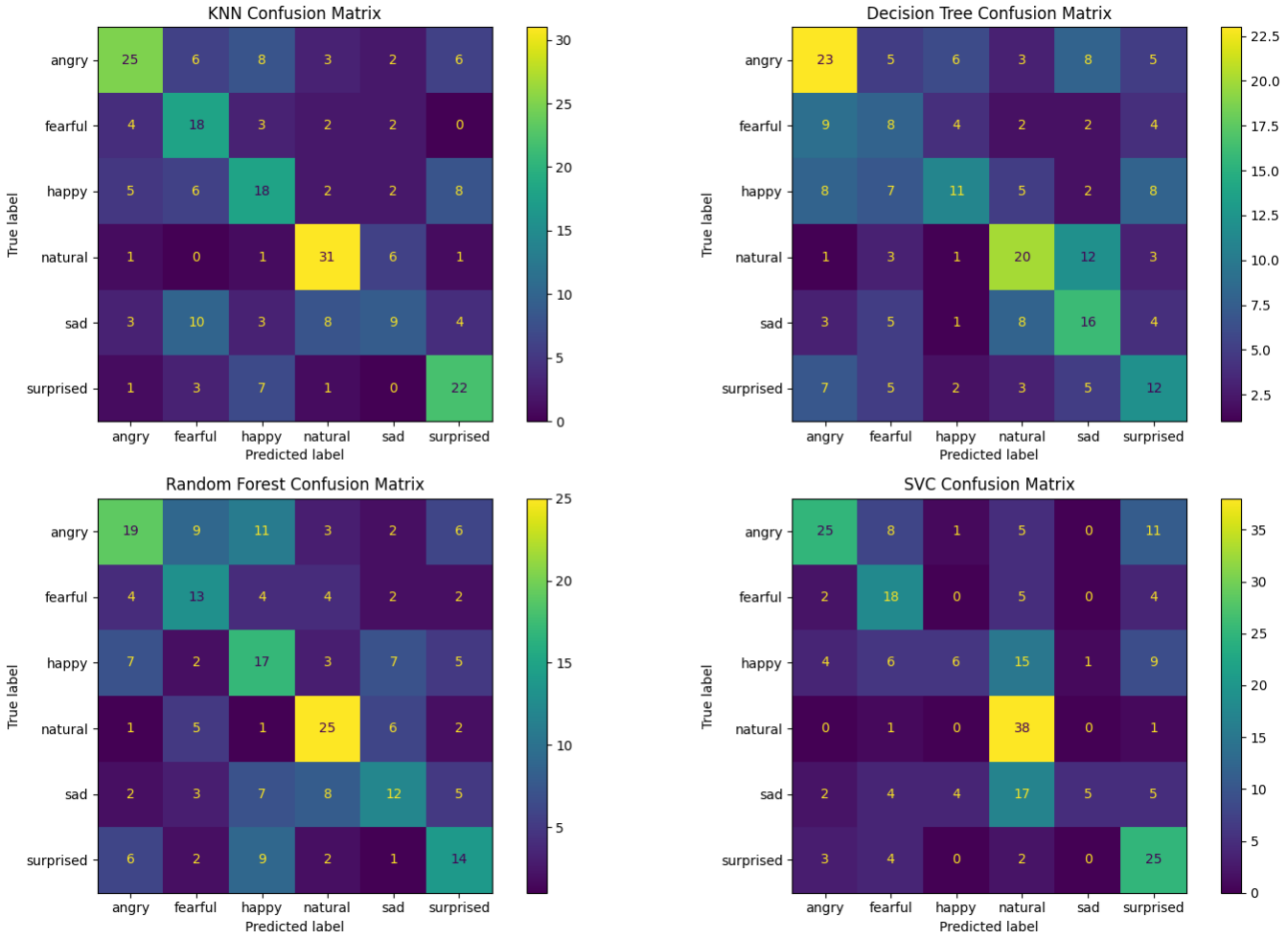


figure: Confusion Matrix

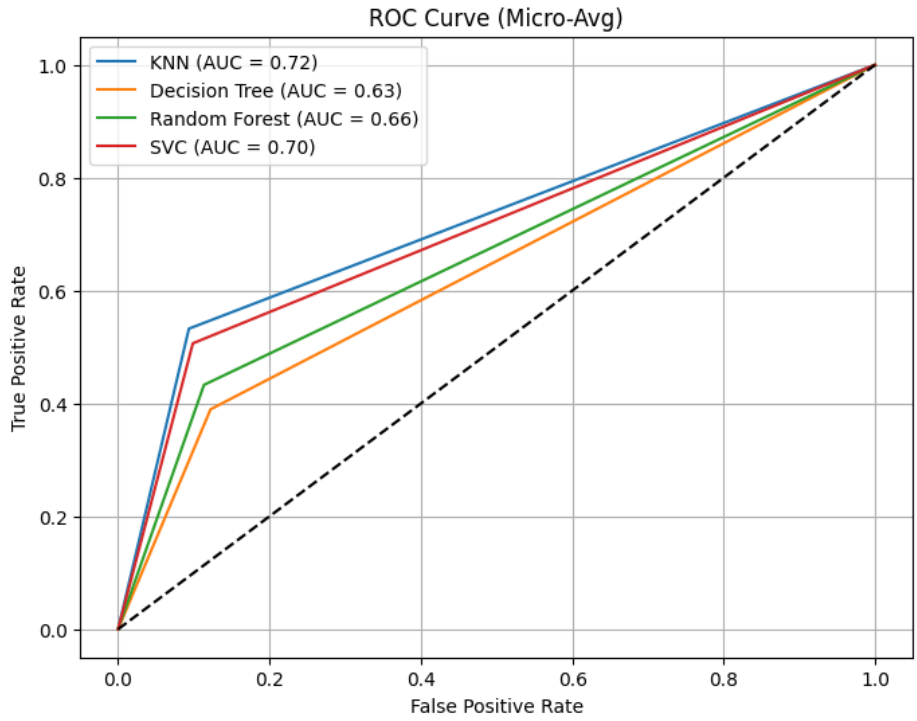


Figure: ROC Curve

Table IV- Class wise metrics for English Dataset

Class	Precision	Recall	F-1 Score	Support
Angry	0.6304	0.58	0.6042	50 samples
Fearful	0.5312	0.5862	0.5574	29 samples
Happy	0.4048	0.4146	0.4096	41 samples
Natural	0.7333	0.8250	0.7765	40 samples
Sad	0.4688	0.4054	0.4348	37 samples
Surprised	0.6765	0.6765	0.6765	34 samples

5.3 Limitations of Tools

Our main limitation of this work was not being able to use any Neural Network models and advanced algorithms. Also, *Teager Energy coefficients* generated from spectral features have been used to get more precise predictions Besides if we could use various built in libraries to extract features, the overall accuracy could have been improved to a great extent. But that's beyond this course capacity.

6 Future Work

In future we aim to work on these datasets with various Neural Network model and built in library functions to get better accuracy. Overall there are a lot of scopes with working these datasets.

7 Conclusion

In this project, both Bangla and English speech sentiment detection model is implemented to detect emotions from natural language audio data. Pre-built famous datasets are used from Kaggle for this purpose. Several ML algorithms are constructed to procure the model best suited to our purpose. This has a better accuracy rate compared to some of the works earlier published only using feature extraction and machine learning approach in this sector. In addition, our model is a speaker-independent method performing uniformly irrespective of gender division. One of the major objectives of this project is to detect hate speech and audio threat, which is becoming a monstrous issue in current technology-based society. This can also have a far- reaching consequence in psychology analysis and personality evaluation. Also, this project can be extensively used in the field of artificial intelligence and robot-human interaction.

7.1 Acknowledgement

Our project is supported by Department of EEE, Bangladesh University of Engineering and Technology. This project is done under the supervision of Shafin Bin Hamid, Assistant Professor, Department of EEE, BUET and Rafid Hassan Palash, Lecturer, Department of EEE, BUET. We would like to thank them for their constant support and supervision. We would also like to express our utmost gratitude to Dr. Steven R. Livingstone, who leads the Affective Data Science Lab, Dr. Frank A. Russo, who leads the SMART Lab, both at Ryerson University (now Toronto Metropolitan University) and Rajib Kumar Das, Nusrath Islam, and Md. Rakibuddin Ahmed, from Shahjalal University of Science and Technology, Bangladesh because the RAVDESS dataset and Bangla SER dataset that we used was constructed by them.

7.2 References

[1] Negi, Himani and Bhola, Tanish and S Pillai, Manu and Kumar, Deepika, ”A Novel Approach for Depression Detection Using Audio Sentiment Analysis,” *International Journal of Information Systems & Management Science*, Vol. 1, no. 1, 2018.

[2] Chen, L., Mao, X., Xue, Y., & Cheng, L. L. (2012a). ”Speech emotion recognition: Features and classification models*,” Digital Signal Process ing,* 22(6), 1154–1160.

[3] Julien PINQUIER. Indexation sonore, ”Recherche de composantes pri maires pour une structuration audiovisuelle. Human-Computer Interac tion*.” Universit´e Paul Sabatier*- Toulouse III, 2004

[4] Scheirer, E., and M. Slaney, “Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator*.” 1997 IEEE International Con ference on Acoustics, Speech, and Signal Processing, 1997.*

[5] J. Mariani, editor, *”Speech Analysis, Synthesis and Coding, Language Processing”* Spoken I. Hermes, 2002.

[6] Calliope*,” Speech and its automatic processing .”* Masson, Paris, France, 1989.

[7] C. Joder, S. Essid, and G. Richard*, ”A comparative study of tonal acoustic features for a symbolic level music-to-score alignment”.* in Proceedings of the 35nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Dallas, USA, 2010.

[8] Kumar P, Chandra M. Speaker identification using Gaussian mixture models. MIT International Journal of Electronics and Communication Engineering. 2011;1(1):27-30

[9] Mosa GS, Ali AA.,” Arabic phoneme recognition using hierarchical neural fuzzy petri net and LPC feature extraction.” Signal Processing: An International Journal (SPIJ). 2009;3(5):161

[10] Vıctor Su´ arez-Paniagua and Isabel Segura-Bedmar*,"Evaluation of pool ing operations in convolutional architectures for drug-drug interaction extraction"* The 11th International Workshop on Data and Text Mining in Biomedical Informatics Singapore. Singapore. 10 November 2017

[11] Nitish Srivastava,Geoffrey Hinton,Alex Krizhevsky,Ilya Sutskever,Ruslan Salakhutdinov*,"Dropout: A Simple Way to Prevent Neural Networks from Overfitting"* June 2014,Journal of Machine Learning Research 15(1):1929-1958

[12] M. Welling*, "Robust Higher Order Statistics."* Proc. Int. Workshop Artif. Intell. Statist(AISTATS) 2005,pp.405-412.

[13] S. Sultana, C. Shahnaz, S.A. Fattah, I. Ahmmed, W.-P. Zhu, M.O. Ahmad*, "Speech Emotion Recognition Based on Entropy of Enhanced Wavelet Coefficients*,