EMOTION RECOGNITION FROM ISOLATED BENGALI SPEECH

[1]JYOTIRMAY DEVNATH, [1]MD. SABBIR HOSSAIN, [1]MOSHIUR RAHMAN, [2]HASI SAHA, [2]AHSAN HABIB, [2]MD. NAHID SULTAN

[1,1,1,2,2,2]Hajee Mohammad Danesh Science and Technology University, Department of Computer Science and Engineering, Bangladesh

E-mail: [1]jyotirmay.pulak@gmail.com,[1]hossainsabbir070@gmail.com, [1]mrmohim74@gmail.com, [2]hasi.cse3@gmail.com, [2]ahsan.habib@hstu.ac.bd, [2]nahid.sultan@hstu.ac.bd

ABSTRACT

In past few eras, emotion recognition from speech is one of the hottest research topic in the field of Human Computer Interaction. Many researches are going on various types of language, but for Bengali language, it is still very novice. In this work, 4 emotional state have been recognized i.e. happy, sad, angry and neutral from Bengali Speech Dataset. Proposed approach uses Pitch and Mel-frequency Cepstral Coefficient (MFCC) feature vectors to train k-Nearest Neighbor classifier for this work. A self-built Bengali emotional speech dataset has been used for both training and testing purpose. The dataset consists of consist of 50 people with 400 isolated emotional sentences. Using this dataset and above technique, we achieved 87.50% average accuracy rate, with detection accuracy each emotion (happy, sad, angry, neutral) respectively 80.00%, 75.00%, 85.00% and 75.00% in this work.

Keywords: *Speech Corpus, Noise Removing, Pitch, Mel-frequency Cepstral Coefficient (MFCC), K-Nearest Neighbor (KNN) classifier*

1. INTRODUCTION

Day by day technologies are greatly increasing its abilities. From the simple radio to now the smartphones, technology influenced our life and society in many ways. Nowadays it simplified people's lives by the interaction of different connections. For human machine interaction, one of the medium of interaction is speech. For human machine interaction the key challenge is detecting the proper emotional state and do like the command of the human. Human can easily recognize the emotion of a speaker but by a machine it is quit hard to recognize properly. Emotion can be detected from speech, facial expression, text, body gesture, physiological monitoring, visual appealing etc. Among all of these speech is the most fundamental and natural communication means of human being in communication. So, the speech can be more efficient and fast method of interaction between machine and human. Though, there is a significant improvement in speech recognition but still researchers are away from natural interplay between computer and human, since computer is not capable of understanding human emotional state. Speech emotion recognition is mostly beneficial for applications, which need human-computer interaction such as speech synthesis, customer service, education, forensics, automatic security system, helping the blind, autistic people, observation in call centers and medical analysis [1]

Now-a-day several voice assistant products like Alexa (Amazon), Siri (Apple), Google Assistant (Google) and Cortana (Microsoft) look like to represent a cool, fast and easy way to execute normal actions to help people attain a better class of personal life. If we can apply the emotional stage in these type of products, then they will be more friendly and smooth.

Emotion Recognition System is a kind of automatic speech recognition system. Automatic Speech Recognition systems can be classified into various templates. One of them is like the following types:

a. Narrator Dependent system: These systems can recognize speech from a particular speaker. They are easier to process, low-cost and more perfect but not as flexible and adaptive to real applications containing multiple speakers.
b. Narrator Independent system: These systems are developed to work with any speaker. They are more complex, high-cost and have less perfection but more flexible and adaptable for realistic applications that need to handle the auditory changeability in speech from multiple speakers.

Speech recognition can also be categorized into the following types:

1. Isolated Speech Recognition: The speech recognized by such systems are detached by pauses or contain utterances of one single word at a period. They are easier to build because the end points of the speech signals are easily noticeable and the articulation of a word is not affected by other words that precede or follow. They can also be quite robust since all probable patterns for the inputs are recognized. Isolated word recognition systems can be designed and manufactured for certain application oriented words such as, digits recognition for phone dialing, navigation related words e.g. left, right, forward, backward etc.
2. Continuous Speech Recognition: A continuous speech recognition system works on words that are linked without pauses. It identifies the normal flow of speech. The improved complication of such systems arises because of a number of aspects. Firstly, it requires recognition of start and end points of each word. Another is that, since the phonemes are related together, utterance of each word is affected by the words that surround it. This is known as "co-articulation". It is also affected by the speediness and rate of dialogue.

Several speech recognition systems may only need to identify a few words, for example the digits, while others need a large set of words reliant on the application. Though automatic speech recognition systems can be further classified based on the size of the vocabulary that is recognized by the system like small, medium and large.

All the categories can be defined as the following Fig. 1.

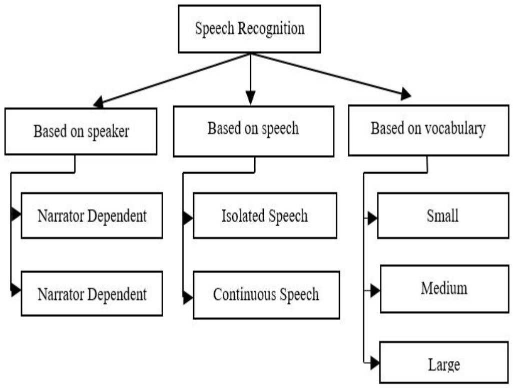


Fig. 1: Categories of speech for recognition process

Various kind of emotions exist in our daily conversations. Physiologist have tried to recognized the variants of emotions that people practice. Physiologist Paul Eckman recognized six basic emotions that he proposed were unanimously practiced in all human cultures. The recognized emotions were happiness, sadness, disgust, fear, surprise, and anger. Far along he extended his list of basic emotions to include such things as pride, shame, embarrassment, and excitement [2].

Recently, scientists acknowledged different categories of emotion. Rather than being totally distinctive, however, the researchers found that people practice these emotions along an ascent [3]. Some of the emotions are:

Happiness: Happiness is expressed through an upbeat, pleasant tone of voice with a smiling facial expression for the feeling of joy, satisfaction, gratification and well-being [4].

Sadness: Sadness describes a "negative" emotion, voices are made of gravel, clear tone are undercut with a choking heaviness resulting from the loss of someone or something important [5].

Angry: Angry emotions describe tone of voice such as speaking gruffly or yelling. It is generally an elevated Voice, high-pitched and loud [6] [7].

Fear: It is a powerful emotion that can play an important role in survival. Physiological reactions such as rapid breathing and heartbeat occurs while talking in a fearful situation [7].

Neutral: Emotional neutrality is the concept of removing happy, sad, fear and other human emotions. This are just a statement which are used for general purpose conversation [8].

To observe emotional state of human voice, we need to know about the characteristics of human voice which are necessary for speech emotion recognition. The human voice and associated speech patterns can be characterized by a number of attributes, the primary ones being pitch, loudness or sound pressure, timbre, amplitude, speech rate etc. In another way, various features are used such as LPC (Linear Predictive Coding), LPCC (Linear Prediction Cepstral Coefficient), MFCC (Mel Frequency Cepstral Coefficient). Our proposed system will be learned with dataset that produce the feature of the speech such as Pitch and MFCCs (Mel Frequency Cepstral Coefficients) [10]. A natural language dataset has been used in this work which is consist of voice records of 50 students of computer science and engineering department of Hajee Mohammad Danesh Science and Technology University, Dinajpur-5200. Our proposed system will take the testing input voice signal and extract the feature from that. After extracting the features system will classify the signal and match it with dataset. For classification and recognition, a classifier the k-NN (k Nearest Neighbor) has been used in our work.

2 ANALYSIS OF EXISTING SYSTEM

2.1 Emotion Recognition System

Automatic Speech Recognition (ASR) is the process of recognizing and reacting to the distinct sounds produced by human speech. It enables a program or a machine to translate the spoken words or sentences in machine readable format and identify them.
The speech samples are taken as input. Then the first thing to be done with the speech samples is the pre-processing where noise from the sample is removed and the purified sample is ready to use. Now from the noise free samples desired features are extracted. These features are then further pass on to the classifier. The classifier thus classifies the emotions accordingly and outputs the emotions.
General flow of the steps of a speech emotion recognition system is showed in Fig. 2.

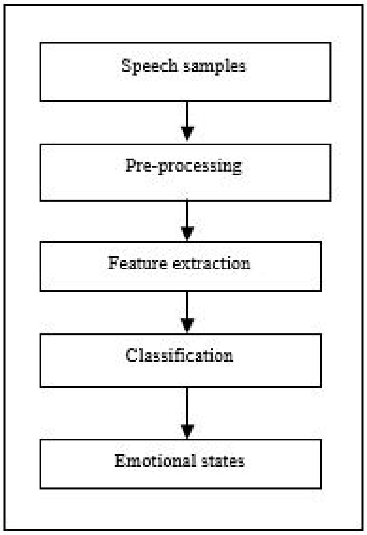


Fig. 2: Emotion Recognition System from Speech

2.2 Dataset

The proper selection of speech database (corpora) shows a very significant role in the arena of emotion detection. A suitable emotional speech database is preferential for a good emotion recognition system. Mainly three types of corpora are used for developing a speech system [20],

1. Elicited emotional speech database: In this type of database the data are collected from the speaker using artificial emotion.
2. Actor based speech database: The data are collected from the experienced and trained artist.
3. Natural speech database: This are created from the real world data. It includes the data recorded from call center conversations, cockpit recordings during abnormal conditions, conversation between a patient and a doctor, conversation with emotions in public places and similar interactions [21].

Some of the Databases are- Danish emotional database, Berlin emotional database, Natural ESMBS, SUSAS, MPEG-4, Beihang University [9]. It has been observed that for database creation nonprofessional and professional actors give their contribution by simulating emotions. [1] used the emotion corpus of 720 utterances included of 6 native Burmese and 6 native Mandarin language speaker for the training and testing process of six emotional classes. [10] detect fron the EMA (Electromagnetic Articulography) of speech related research of USC (University of South California) for training include of 3 speakers (1 male and 2 female) and 10 types of sentences to differentiate to emotional stage. In [11] the Berlin Emotional Database have been used for the training and testing purpose. The Berlin emotional database consists of 10 speakers (5 males and 5 females). Each one of the speakers is asked to speak 10 different texts in German. The database consists of 535 speech files. A Semi-natural Telugu database was used for speech emotion recognition which was recorded in Telugu language and it contained 7 numbers of students from IIIT Hyderabad for emotions such as anger, happy, neutral and sad [22]. In [12] there have been used two type of speech: a) Acted Emotion Speech and b) Real Emotion Speech. Acted Emotion Speech consists of 4 actors (2 males and 2 females) where each actor was made to speak 15 sentences in Telegu Language. The real emotion speech was collected for the Telegu movie of a single male actor.
In term of Bengali language, there is no such quality assured and a standard dataset to determined actual emotion for the natural speech. Some of researcher worked on gathering some data for Bengali speech but, still now more reliable and customary dataset is needed for Bengali speech. In [23], a standard dataset named as “SHRUTI” was used for an automated speech recognition system.
It is a read speech corpus designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems.

Communication Empowerment Lab, Indian Institute of Technology, Kharagpur (IITKGP) with collaboration with Media Lab Asia. The speech was recorded, transcribed and has been maintained, verified at IITKGP. This file contains a brief description of the SHRUTI Speech Corpus. SHRUTI contains a otal of 7383 unique sentences. Sentences are spoken by 34 speakers from region of standard Bngali colloquial language of the West Bengal. The percentages of male and female speakers are 75% and 25% respectively. The speaker age in the corpus varies in between 20 to 40 yrs. A speaker's dialect region is the geographical area of the West Bengal, India where they lived during their childhood years. Text material is collected from “Anandabazar patrika”. The domains are sports, political, general news and geographical news. Total sentences are recorded by 34 speakers with two sessions. The phonetically-compact sentences were designed to cover most of the frequent speaking word in Bengali language. Each speaker read variable number of sentences which were collected in ITRANS format.
Another version of “SHRUTI” was introduced in [24] as “SHRUTI-II” for a speech recognition system of visually impaired community. Further some of volunteer dataset are found in web like [26] [27] [28], but they not so reliable as we want in our research. All of the above were implemented in different purpose for various speech recognition system but the datasets were not for more efficient for any emotion recognition process. For emotion recognition process, we need some data which are more correlated to emotional sentences or speech. So for the better result, a natural dataset which consisted with emotional speech need to be generated.

2.3 Feature

The speech processing stage is based on a number of signal processing stages. These are speech such as endpoint detection, windowing of the speech signal, filtering the speech [13] samples so that there would be no noise left in the speech signal, linear predictive coding of speech, computing the cepstral coefficients and then perform the vector quantization of the signal to obtain the codebook, which is used in the pattern recognition stage. Secondly, the frame blocking and windowing is a process to decompose the speech signal into short speech sequences called frames to conduct speech analysis. There are several windows that can be utilized such as the rectangle window, triangular window, but the Hamming window is often chosen as it softens the edges created due to framing, again emphasizing on simplicity [14]. Third is the feature extraction. The features to be extracted are various, but they can be grouped into 4 distinct groups, namely continuous, qualitative, spectral, and TEO-based features [15]. These features can be used as a sole determinant, but often they are used in combination to generate a more distinguishable pattern for the system.
We selected MFCC due to its suitability for N-based classifiers. Moreover, many researchers have used MFCC as the audio features. So we hope our proposed system could be benchmarked with other research.
Table. 1 shows the strength and weaknesses of various audio features [16] [17].

Table 1: Summary of literature review on audio features for SER

Feature	Strength	Weakness
LPCC	One of the most traditional features which implies that it is widely recognized and used.	LPC on its own has is not as reliable, as seen that it is often combined with other feature extraction methods.
MFCC	Tuned in a scale that is suitable for the human ear. Alongside with LPCC, is considered one of the standard features extracted, even more-so in SER. Best suitable for N-way classifier.	MFCC being in spectral form is sensitive towards noise.
TEO	Nonlinear approach, which is suitable for speech. Superior detection in stress-levels of emotion.	More complicated computations as compared to LPC.

2.4 Classifier

After the speech emotion recognition (SER) system extracts the desired features from the audio speech data, the next step is to pass the data on to the classifier. The primary job of the classifier is to determine the unrevealed emotion of the user by using a set of defined algorithms and functions.

Usually these classifier evaluations are performed using a single database or dataset, under one language. Up until now, there has been no agreed standard of which classifier is the best, but many have been evaluated to achieved better recognition [18].

Machine learning algorithm (classifier) can be divided into following categories [29].

Table. 2 Categories of Machine Learning Algorithm

Artificial Neural Network Algorithm (ANN)	**Instance-based Algorithms**
▪ Perceptron ▪ Back-Propagation ▪ Hopfield Network ▪ Radial Basis Function Network (RBFN)	▪ k-Nearest Neighbour (kNN) ▪ Learning Vector Quantization (LVQ) ▪ Self-Organizing Map (SOM) ▪ Locally Weighted Learning (LWL)
Bayesian Algorithm	**Regression Algorithms**
▪ Naive Bayes ▪ Gaussian Naive Bayes ▪ Multinomial Naive Bayes ▪ Averaged One Dependence Estimators (AODE) ▪ Bayesian Belief Network (BBN) ▪ Bayesian Network (BN)	▪ Ordinary Least Squares Regression (OLSR) ▪ Linear Regression ▪ Logistic Regression ▪ Stepwise Regression ▪ Multivariate Adaptive ▪ Regression Splines ▪ Locally Estimated Scatterplot Smoothing
Regularization Algorithms	**Deep learning**
▪ Ridge Regression ▪ Least Absolute Shrinkage and Selection Operator ▪ Elastic Net ▪ Least-Angle Regression (LARS)	▪ Deep Boltzmann Machine (DBM) ▪ Deep Belief Networks (DBN) ▪ Convolutional Neural Network (CNN) ▪ Stacked Auto-Encoders
Clustering algorithms	**Association rule learning algorithm**
▪ k-Means ▪ k-Medians ▪ Expectation Maximisation (EM) ▪ Hierarchical Clustering	▪ Apriori algorithm ▪ Eclat algorithm
Decision Tree Algorithm	**Dimensionality reduction algorithm**
▪ Classification and Regression Tree (CART) ▪ Conditional Decision Trees ▪ M5 ▪ Decision Stump ▪ Chi-squared Automatic Interaction Detection ▪ C4.5 and C5.0 (different versions of a powerful approach) ▪ Iterative Dichotomiser 3 (ID3)	▪ Principal Component Analysis (PCA) ▪ Principal Component Regression (PCR) ▪ Partial Least Squares Regression (PLSR) ▪ Sammon Mapping ▪ Multidimensional Scaling (MDS) ▪ Projection Pursuit ▪ Flexible Discriminant Analysis (FDA)

In pattern recognition stage, the codebooks made by the feature extraction are used as an input to the recognizer. The recognizer used in this stage some classifier such as the maximum likelihood Bayes classification, kernel regression, VQ (Vector Quantization), HMM (Hidden Markov Models), GMM (Gaussian Mixture Model), k-nearest neighbor (K-NN) classifier, Support Vector Machine (SVM), Deep Neural Network (DNN), Artificial Neural Network (ANN). The codebook is taken as the input and the related method is used for training purposes and then the training parameter is stored for future to test the same sample speech [18] [19].

From the Berlin emotional database and the features extracted from these utterances are energy, pitch, ZCC, entropy, Mel Frequency cepstrum coefficients (MFCC) [25] used the K Nearest Neighbor (KNN) algorithm for the classification of different emotional states.

A typical SER consisted of at least feature extraction, classifier, and speech emotion database. From the critical literature review, of the various audio features we selected MFCC and pitch due to its popularity and suitability, while k-NN (k-nearest neighbor algorithm) was selected as the classifier due to its higher accuracy if more data is available. A comprehensive emotion dataset mentioned earlier was used for the training and testing purpose.

The research landscape for Bengali Speech Recognition is relatively nascent in comparison to the rich history of ASR system development involving other languages and that is the motivation of this work.

3 PROPOSED SYSTEM DESIGN

The design is the depiction of any system with necessary elements. These elements are categorized as Flow charts, Block diagrams, DFD (Data Flow Diagram) etc. How the system takes the inputs, process them and finally detect several attributes of emotions from the speech and by the detected attributes of emotions are classified by some classifier that will show the category of the detected emotion.

There are various datasets on different language as we discussed it in 2.1 and various research had been employed on those datasets. But for Bengali language the development of automatic speech recognition is quit slow and deferred. So we focused on Bengali language.

The main advantage of the proposed system is dataset and accuracy. In this work the dataset looks like more natural, diverse and effective than the previous. The performers of the dataset are general peoples, so the emotions are expressed in natural formats. This reason helps the classifier to improve the result in natural audio processing systems. It will also determine the emotion more accurately on natural data. Our goal is to make the communication between human to human and human to computer more realistic. This will help mostly in the field of robotics. The robots will be able to recognize the emotion of a person from the speech and can do works more fluently.

3.1 Dataset Processing

As mentioned earlier some datasets of Bangla speech like "SHRUTI", "SHRUTI-II" existed but they are not enough for detecting emotional state of a speech. All of the datasets are collected from movie dialogue or spoken from the page of a book or paper. These dialogue or pronounced speech can't express the proper emotional state of our everyday life. So, some emotional speech corpora needed to be generated for the purpose of emotion recognition. Thus in our proposed system we used a dataset which are collected form general dialogue of common people. The sentences of this dataset contain different emotions in various categories.

The proposed system used a dataset which was created by the recorded voice of 12 male and 8 female students of department of Computer Science and Engineering of Hajee Mohammad Danesh Science and Technology University. The voices were in Bengali language and the age limit of the students was about 23-28 years. Every student spoke 8 sentences which consist of four ("Happy", "Sad", "Angry" and "Neutral") emotions. For each emotion we recorded two sentences. We used total 160 data for training and testing sections. Among them 128 were used for training and 32 others were used for testing.

Table. 3: Dataset Introduction

Total data (160)			
Training(128)		Testing(32)	
Male(80)	Female(48)	Male(16)	Female(16)
Happy (20) Sad (20) Angry (20) Neutral(20)	Happy (12) Sad (12) Angry (12) Neutral(12)	Happy (4) Sad (4) Angry (4) Neutral (4)	Happy (4) Sad (4) Angry (4) Neutral(4)

3.1.1 Preprocessing

The recordings were recorded with a sampling rate of 44100Hz and in stereo recording format. Then the noise of the sample audio data was removed by audacity software.

The noise was filtered as below:

1. Noise reduction level(dB): 14dB
2. Sensitivity level: 6.00
3. Frequency smoothing(bands): 3.00

The Training and testing data were collected in an air proofed room and Table. 3 shows an overview of the dataset.

The sentences used in our proposed system were in 4 different types of emotions and for each emotion we collected two sentences which are generally used for expressing the corresponding emotions. The sentences were recorded as an acted based platform. The speaker tried to speak the dialogue in proper emotion. The number sentences used in this system are eight (8). For happy emotion we used "Mom, I got A plus" and "Hurray! We have won the game", for sad "Mom, I have failed in the examination" and "We have lost the game", for anger "Damn, what are you doing?" and "Don't bother me" and lastly, for neutral we used "Bangladesh is a peaceful country" and "Two and two make four". The pronunciations of the sentences are described in Table 4.

Table. 4: Dataset used in proposed system

Emotions	Recorded Sentences	Pronunciation in English
Happy	1. মা আমি এ প্লাস পেয়েছি।	1. Mā āmi ē plāsa pēẏēchi.
	2. হুররে আমরা খেলায় জিতেছি।	2. Hurarē āmarā khēlāẏa jitēchi.
Sad	1. মা আমি পরীক্ষায় ফেল করছি।	1. Mā āmi parīkṣāẏa phēla karachi.
	2. আমরা খেলায় হেরে গেছি।	2. Āmarā khēlāẏa hērē gēchi.
Anger	1. ধুর, এইটা কী কাজ করছ।	1. Dhura, ē'iṭā kī kāja karacha.
	2. বিরক্ত করিস নাতো ।	2. Birakta karisa nātō.
Neutral	1. বাংলাদেশ একটি শান্তিপ্রিয় দেশ।	1. Bānlādēśa ēkaṭi śāntipriẏa dēśa.
	2. দুইয়ে দুইয়ে চার হয়।	2. Du'iẏē du'iẏē cāra haẏa.

Our proposed system can be divided into two major sections. Firstly, the extraction of the features of the training data and then the classification process by k-Nearest Neighbor (k-NN) algorithm.

3.2 Feature Extraction

Feature extraction is the first step in Automatic Speech Recognition which encodes the speech data as a set of quantifiable feature vectors that are fed into the acoustic models. Speeches are certain sounds that are shaped by articulation of the vocal cord, nose, tongue, teeth and other organs. Speech signal is basically a one dimensional waveform which has some discrete and measureable qualities to it, such as, energy level, certain frequencies and so on. It is important to select feature vectors in a way that minimizes redundancy, gets rid of unwanted or unimportant features and focuses on features that distinctly represent different sample classes.

Pitch: Pitch is the fundamental period of the speech signal. It the perceptual correlate of fundamental frequency. It represents the vibration frequency of the vocal cords during the sound productions.

Mel-Frequency Cepstral Coefficients (MFCC): MFCC is the most popular feature vectors for speech and voice recognition. The popularity of MFCC vectors is attributed to the fact that it closely mimics the way human ears perceive sound and respond accordingly.

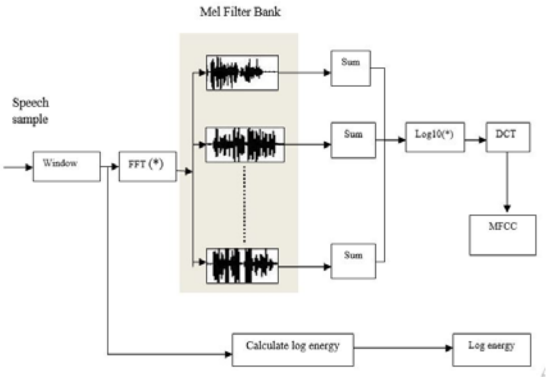
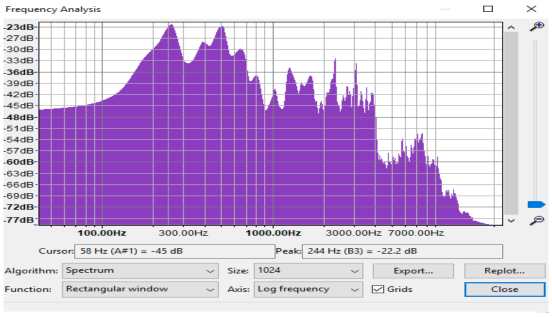


Fig. 3: Block Diagram of detecting MFCCs

Speech is a time varying signal, i.e. it is always changing. Therefore, the signal is divided into small segments, e.g. 20-40 ms frames where the signal can be considered to be statistically stationary. An aural frame has amplitude or loudness data at that certain point in time. If the frame is too long, the signal may change too much which will cause the feature vectors to be a lot less useful for accurate predictions. Frame steps are 10ms which causes some overlapping of frames. In this work, 30ms frames were used.

After dividing the signals into small frames, the second step is calculating the power spectrum of each frame. This was inspired by the mechanism of cochlea, an organ inside human ear. A sample spectrum of a voice data is shown in Fig. 4.



Cursor: 58 Hz (A#1) = -45 dB
Peak: 244 Hz (B3) = -22.2 dB
Algorithm: Spectrum
Size: 1024
Function: Rectangular window
Axis: Log frequency
☑ Grids

Fig. 4: Spectrum of a sample data

It vibrates at different spots depending on the frequency which in turn fires different neurons. The periodogram spectral estimates similarly detect the frequencies present in the input signal. For each step, this was done by taking one of the most influential tools in digital signal processing which permits us to catch the continuum of a finite-duration signal named Fast Fourier Transform (FFT). Fast Fourier Transform is an algorithm which computes the Discrete Fourier Transform a sequential signal. To converts the data signal from its original domain to a representation in the frequency domain the FFT was used here. FFT formula is defined as below.

For a given length- N < ∞ sequence {x[n]}, the N-point FFT {X[K]} is

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i2\pi kn/N} \quad k = 0, 1, 2, \ldots, N\text{-}1,$$

Where a primeval Nth root of 1 is $e^{-i2\pi kn/N}$.

The initial periodogram estimated contain a lot of information that are unnecessary for recognition. The cochlea cannot effectively differentiate between two closely spaced frequencies. This is why the Mel Filter bank is used. The first filter grows from narrow to wider as it provides an estimate of the amount of energy present near zero hertz to higher frequencies. 20-40 (here 26 is used) triangular spaced filters are applied to the periodgram power estimates. This results in 26 vectors of length 257. Each filter bank is multiplied with the power spectrum and then the coefficients are added up to calculate the filter bank energies.

In the next step, the logarithm of the filter bank energies is taken. This is also influence by how

humans perceive sound. It is necessary for the large variation of energy to sound similar. So the perceive loudness is increased so that the variations are insignificant since the sound was already loud enough compared to the energy variations to make it ignorable. Therefore, the logarithm of each of the 26 energies from the previous step is taken.
The last step of the feature extraction process is calculating the Discrete Cosine Transform (DCT) of the log filter bank energies from the previous step. The filter bank energies are correlated to each other since they are overlapping. This step fixes that so that these can be used for diagonal covariance matrices. This step results in 26 cepstral coefficients only the lower 13 of which are kept.

3.3 Training Classifier:

Many researchers used various types of classifier for the classification task such as SVM, ANN, DNN etc. In this system we used KNN (k-Nearest Neighbor) classifier for classification because it gives better result in multiclass classification.

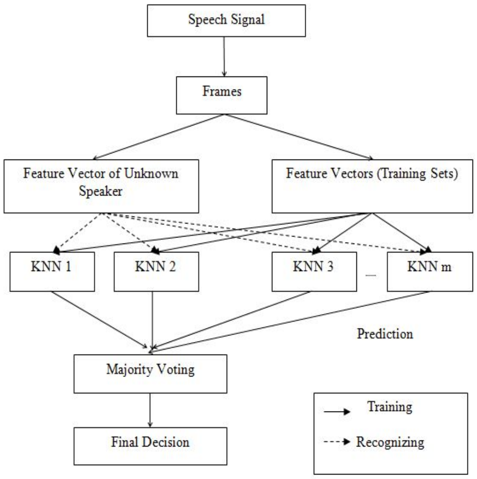


Fig. 5: k-NN Structure and Algorithm

For the classification task here in our system we used KNN (k-Nearest Neighbor) because by nature KNN classifier works better for multiclass classification. The hyper parameters for the nearest neighbor classifier include the number of nearest neighbors, the distance metric used to compute distance to the neighbors, and the weight of the distance metric. The hyper parameters are selected to optimize validation accuracy and performance on the test set. In this example, the number of neighbors is set to 5 and the metric for distance chosen is squared-inverse weighted Euclidean distance.
After the feature extraction, KNN gives each speech sample with the corresponding emotion class label. After that it will put them to the classifier and gain a model file by training the data set. When an unclassified speech sample come into this system, the system extracts the feature coefficients and use the model file to classify the speech emotion.

3.4 Full Process at a Glance

The block diagram in figure 6 shows the full procedure of the system.

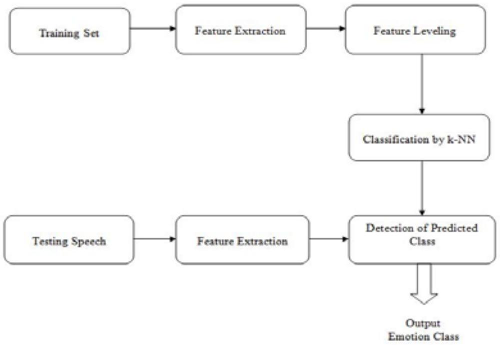


Fig. 6: Block Diagram of the process

Firstly, the speech training set of data is used for feature extraction for the classification. The features are extracted and leveled for smooth functioning of the feature extraction process of Pitch and MFCC. The features are classified by the k-Nearest Neighbor (k-NN) classifier. After the training and classification process is completed then the system is ready for the use. A raw input data is taken from a speaker then the features of that speech is extracted and thus classified by k-NN and then match and predict the class from the dataset.

4 TOOLS AND TECHNOLOGY

To develop a project, tools and technologies play the most important part. The tools denote the components that are used to create the environment for the programmers.
This basic tools and technologies described in this project are listed below:

- Language and Platform: MATLAB R2018b
- Algorithm: k-NN classifier
- Operating System: Windows 10 64 bit
- Minimum required RAM: 4 GB

5 RESULT AND DISCUSSION

In our work we have used our own dataset for both training and testing purposes. In our dataset 12 male and 8 female performed all the emotional speeches which were used for training and testing purpose. For the training section, 128 emotional data are used for four emotion classes, 32 for each emotion class. For the testing purpose 32 emotional data are used, 8 for each emotional class.

Among the testing data of happy emotion 8 data have showed the desire emotion. The 8 sad emotional data were tested and among them 7 data gave the exact result but the rest 1 data did not do that. For the 8 angry emotion data 6 of them give the exact result and rests of them give wrong answer. Lastly the 8 neutral emotions give desire result. Thus the performance of total system became 87.50%.

The confusion matrix of our work is given in Table 5.

Table. 5: Tabular Result of the Testing Data

		Target Class			
		Happy (8)	Sad (8)	Angry (8)	Neutral (8)
Output Class	Happy (8)	7	0	0	1
	Sad (8)	0	7	1	0
	Angry (8)	1	1	6	0
	Neutral (8)	0	0	0	8

In this system four types of emotional dataset like Happy, Sad, Angry and Neutral are used. The accuracy percentage was like table 6:

Table. 6: Accuracy measurement of the system

Emotions	Accuracy (%)	Overall System Performance
Happy	87.50	87.50%
Sad	87.50	
Angry	75.00	
Neutral	100.00%	

For the used dataset the accuracy level of neutral was perfect but the sad and happy was quite problematic and the angry was less accurate. If the data can be increased there will be some variations. The used dataset is not regional based. Similar feature and classifier was used in [26] and they recognize the accent verities of speakers and categorized them in groups. So, we believe that the classifier should be able to address different accent variation in the testing. This approach will show us difficulties if the data were not accent free. In [27], similar classifier was used for the classification of different emotional states from the Berlin emotional database and the features were extracted from the energy spectrum, entropy and Mel Frequency cepstrum coefficients (MFCC). They got a result of around 60 to 70 percent.

Here we used the pronunciation of the sentences according to "Promito Bangla Grammar" which is the basic standard of pronunciation of Bangla language. Various projects have been developed on various methods. In [26] and [27], using this method the accuracy is around 78%.

On the basis of the dataset discussed in 2.1 and extracting various features as mentioned in 2.2, many researches had been occurred using various classifier. Among them the accuracy range is approximately 65% to 90%. This project has an accuracy of 87.50% with the experiment of four emotion class. We think that if we add more emotion classes like depression, nervous etc. then we can get similar performances. These extended features are the future work of our project.

6 CONCLUSION

In the current past widespread effort have been depleted by scholars in the region of emotion recognition via speech. We surveyed a significant number of research papers based on three factors—database, feature extraction and classifiers.

A majority of the recent research also focuses on feature extraction and choice so as to select the finest features and progress the performance accuracy. It has been renowned from analyses of the data that to increase the system performance and detect the exact emotions, classifier selection is a thought-provoking task. Many classifiers have been picked for speech emotion recognition system but it is very tough to determine which achieves better-there is no perfect champion [20].

We recorded an emotional database with the "big four basic emotions (Happy, Sad, Angry and Neutral). The emotional utterances were performed by the student of Department of Computer Science and Engineering of Hajee Mohammad Danesh Science and Technology, Dinajpur which belong male and female both. The age limit was around 25. The collection process of the data was in natural atmosphere. Recorded data were collected with safe and diligently. After the collection of the data, we process them to make suitable to use them. Then we extract the required features as Pitch and Mel-Frequency Cepstral Coefficients for the classification process.

In our proposed work, finally we conclude about recognition of speech using K-NN (K Nearest Neighbor Algorithm) that belongs a basic classifier technique. Ability of Feature extraction method in constructing features that have used to speech recognition as well as ability of classification method in educating process are two basic to fruitful emotional speech recognition process. Although it is difficult to exactly compare recognition accuracies from this study to other studies because of the different data sets used, the methods employed here are particularly promising. The recognition accuracies obtained using K-NN for Bengali speakers are quite satisfactory. Previous work in the track of Bengali speech emotion detection is very less paralleled to other tongues like English, German, Japanese etc. There are available datasets of other language, Bengali speech dataset are available but limited in actual format of cross lingual from which the emotions should be matched. This project shows that there is significant benefit in doing so on Bengali language. Our methods are sensibly accurate at recognizing emotions in all speakers. Our project shows that features derived from natural emotions such as happiness, sad, angry and neutral have different properties.

REFERENCES

[1] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” *Pattern Recognition.*, 2011, vol. 44, no. 3, pp. 572–587.

[2] Paul Ekman, "An Argument for Basic Emotions", *Recognition and Emotion*University of california, San Frnacisco, U.S.A. ,1992,vol.6(3/4),pp.169-200.

[2] Alan S. Cowen, and Dacher Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients", Proceedings of the National Academy of Science of the United States of America (PNAS) , 2017, vol.114 (38), E7900-E7909.

[3] Lawrence EM, Rogers RG, Wadsworth T. "Happiness and longevity in the United States." Soc Sci Med, 2015, vol.145, pp.115-9.

[5] Wolkowitz OM, Epel ES, Reus VI, and Mellon SH, "Depression gets old fast: do stress and depression accelerate cell aging?" Depress Anxiety, 2010, vol.27(4), pp.327-38.

[6] Staicu ML, and Cuţov M., "Anger and health risk behaviors." J Med Life. , 2010, vol.3(4), pp.372-5.

[7] Gruber J, Moskowitz JT, " Positive Emotion, Integrating the Light Sides and Dark Sides.", OUP USA, 2014.

[8] M. El Ayadi, M. S. Kamel, and F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases,” *Pattern Recognition.* , 2011b, vol. 44, no. 3, pp. 572–587.

[9] P. P.Shrishrimal, R. R. Deshmukh, and V. B. Waghmare, “Indian Language Speech Database: A Review,” *Int. J. Comput. Appl.* , 2012, vol. 47, no. 5, pp. 17–21.

[10] S. Kim, P. G. Georgiou, S. Lee, and S. Narayanan, “Real-time emotion detection system using speech: Multi-modal fusion of different timescale features,” in Proceedings of the IEEE 9th International Workshop on Multimedia Signal Processing (MMSP '07), Crete, Greece, October 2007, pp. 48–51.

[11] S. Kuchibhotla, B. S. Yalamanchili, H. D. Vankayalapati, and K. R. Anne, “Speech Emotion Recognition Using Regularized,” vol. 7, no. Table I, pp. 363–369, 2014.

[12] R. Nikopoulou, I. Vernikos, E. Spyrou, and P. Mylonas, “Emotion Recognition from Speech,” , 2018, vol. 3, no. 2, pp. 104–105.

[13] T. R Sahoo, S. Patra, "Silence Removal and Endpoint Detection ofSpeech Signal for Text Independent SpeakerIdentification", I.J. Image, Graphics and Signal Processing, 2014, vol. 6, pp. 27-35.

[14] O. K Hamid, "Frame Blocking and Windowing Speech Signal", Journal of Information, Communication, and Intelligence Systems (JICIS), , December 2018, Vol. 4, Issue. 5.

[15] M. Sezgin, B. Gunsel, and G. K Kurt, "Perceptual audio features for emotion detection", EURASIP Journal on Audio, Speech, and Music Processing, , 2012, vol.1, pp.16.

[16] B.V. Pathak and A.R. Panat, "Comparison between Different Feature Extraction Techniques to Identify the Emotion ‘Anger’ in Speech", Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, CCSIT, Part II, LNICST 85, ,

2012, pp. 637–643.

[17] O. Chia Ai, M. Hariharan, S. Yaacob, and L. Sin Chee, "Classification of speech dysfluencies with MFCC and LPCC features" Expert Systems with Applications, , 2012, vol. 39, pp. 2157–2165.

[18] Y. A. Ibrahim and T. S. Ibiyemi, "A STUDY ON EFFICIENT AUTOMATIC SPEECH RECOGNITION SYSTEM TECHNIQUES AND ALGORITHMS", Anale. Seria Informatică., , 2018, Vol. XVI fasc. 2.

[19] M.A.Anusuya, S.K.Katti, "Classification Techniques used in SpeechRecognition Applications: A Review",IJCTA. Comp. Tech. Appl., , 2011, Vol 2 (4), pp. 910-954.

[20] Monorama Swain, Aurobinda Routray, and Prithviraj Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review", International Journal of Speech Technology, , 2018, vol.21, pp.93–120.

[21] Batliner, A., Huber, R., Niemann, H., Nöth, E., Spilker, J., & Fischer, K. "The recognition of emotion." In Verbmobil: Foundations of speech-to-speech translation, 2000, pp. 122–130.

[22] Gangamohan, P., Kadiri, S. R., & Yegnanarayana, B. . "Analysis of emotional speech at sub segmental level." Interspeech, Lyon, France, 2013, pp. 1916–1920.

[23] B. Das, S. Mandal and P. Mitra, "Bengali speech corpus for continuous automatic speech recognition system," Proc. Conf. Speech Database and Assessments (Oriental COCOSDA), Taiwan, 2011, pp.51-55.

[24] B. Das, S. Mandal and P. Mitra, “Shruti-II: A vernacular speech recognition system in Bengali and an application for visually impaired community,” in Students' Technology Symposium (TechSym), IEEE, 2010, TECHSYM.2010.5469156.

[25] Anuja Bombatkar, Gayatri Bhoyar, Khushbu Morjani, Shalaka Gautam,Vikas Gupta, "Emotion recognition using Speech Processing Using k-nearest neighbor algorithm", International Journal of Engineering Research and Applications(IJERA),2014, ISSN:2248-9622.

[26] Munish Bhatia, "Speaker Accent Recognition by MFCC Using K-Nearest Neighbour Algorithm: A Different Approach", International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), January 2015, Vol. 4, Issue 1.

[27] Kacur, J., Vargic, R., Mulinka, P., "Speaker identification byk-nearest neighbors: application of PCA and LDA prior tok-NN." In: Proceedings of the International Conference on Systems, Signals and Image Processing (IWSSIP), Bucharest, Romania, 2011, pp. 1–4.

[26] Bengali.Ai. Machine Learning Repository of Bengali.Ai, Copyright 2016, 2017, 2018 Google, Inc. Retrived January 15, 2020 from https://bengali.ai/datasets/?_sf_s=bangla%20speech.

[27] Crowdsourced high-quality Bengali [bn-in/bn-bd] multi-speaker speech dataset, Google Research, Retrived January 15, 2020 from https://research.google/tools/datasets/bengali-tts/.

[28] Crowdsourced Bengali Bangladesh [bn-bd] ASR dataset, Google Research, Retrived January 15, 2020 from https://research.google/tools/datasets/bengali-asr/.

[29] Machine learning mastery Retrived January 16, 2020 from http://machinelearningmastery.com/a-tour-of-machine-learningalgorithms/.