Islamic University of Technology (IUT)

Department of Computer Science and Engineering (CSE)

Is Speech Emotion Recognition Language independent? A Comparative Analysis of Speech Emotion Recognition using English and Bangla Languages

Authors

Fardin Saad - 154419

&

Md. Al-Amin Shaheen - 154423

Supervisor

Prof. Dr. Md Kamrul Hasan

Professor, Department of CSE

A thesis submitted to the Department of CSE

in partial fulfillment of the requirements for the degree of B.Sc.

Engineering in CSE

Academic Year: 2018-19

November - 2019

Declaration of Authorship

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by Fardin Saad and Md. Al-Amin Shaheen under the supervision of Professor Dr. Md. Kamrul Hasan, Professor of the Department of Computer Science and Engineering (CSE), Islamic University of Technology (IUT), Dhaka, Bangladesh. It is also declared that neither of this thesis nor any part of this thesis has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Authors:

Fardin Saad

Student ID - 154419

Md. Al-Amin Shaheen

Student ID - 154423

Is Speech Emotion Recognition Language independent? A Comparative Analysis of Speech Emotion Recognition using English and Bangla Languages

Approved By:

Prof. Dr. Md Kamrul Hasan

Thesis Supervisor,

Professor

Department of Computer Science and Engineering

Islamic University of Technology (IUT)

Hasan Mahmud

Assistant Professor

Department of Computer Science and Engineering

Islamic University of Technology (IUT)

Acknowledgement

We would like to express our grateful appreciation for **Associate Professor Hasan Mahmud** and **Md Kamrul Hasan**, Professor of Department of Computer Science & Engineering, IUT for being our adviser and mentor. Their motivation, suggestions and insights for this research have been invaluable. Without their support and proper guidance this research would never have been possible. Their valuable opinion, time and input provided throughout the thesis work, from first phase of thesis topics introduction, subject selection, proposing algorithm, modification till the project implementation and finalization which helped us to do our thesis work in proper way. We are really grateful to them.

Abstract

Emotion recognition plays a major role in affective computing and adds value to machine intelligence. While the emotional state of a person can be expressed in different ways such as facial expressions, gestures, movements and postures, recognition of emotion from speech has gathered much interest over others. However, after years of research, recognizing the emotional state of individuals from their speech as accurately as possible still remains a challenging task. This motivates an attempt to study the factors that influence identification of Speech Emotion Recognition (SER) such as gender, culture, dialects, education, social status and age. The aim of this study is to investigate whether a SER system can identify the emotional state of a person regardless of the language used. To investigate the influence of languages in SER, we explored how spoken expressions of six selected emotions (happiness, anger, sadness, neutral, fear & disgust) varied in two languages of interest: English and Bangla. In addition, the perceptual outcomes were studied in relation to identifying the advantage of speech emotion expression produced by native speakers and also by bilingual speakers.

Contents

List of Figures

1 Introduction

A speech signal is naturally occurring signal and hence is random in nature.[2] The signal expresses different ideas, communication and hence has lot of information. There are number of automatic speech detection system and music synthesizer commercially available. However despite significant progress in this area there still remain many things which are not well understood. Detection of emotions from speech is such an area. The speech signal information may be expressed or perceived in the intonation, volume and speed of the voice and in the emotional state of people.

An emotional speech describes a particular prosody in speech. The prosodic rules of a language evolve with the culture of a community over ages.[4] In addition, speakers also have their own speaker dependent style, i.e. a characteristic articulation rate, intonation habit and loudness characteristic. Hence, emotion expressed and inferred in a speech, depends upon the speaker's community culture and language, gender, age, education, social status, health, physical engagements, etc. When a speaker is in a 'quiet room' with no task obligations.

Rapid development in affective computing and advancement in man-machine interaction allows researchers to explore more on the study of emotion recognition from various sources such as face analysis, skin temperature, galvanic resistance and gesture recognition. However, speech is the primary medium of communication[5] and therefore emotion recognition from speech has received great interest over others. While acknowledging that emotion can add intelligence to machines, reported studies revealed that speech emotion recognition still remains a challenging task, specially, when attention was given mainly to the paralinguistic information of speech.

Emotion detection of speech in human machine interaction is very important. Framework for emotion detection is essential, that includes various modules performing actions like speech to text conversion, feature extraction, feature selection and classification of those features to identify the emotions[2] The features used

for emotion detection of speech are prosody features, spectral features and voice quality features. The classifications of features involve the training of various emotional models to perform the classification appropriately. The features selected to be classified must be salient to detect the emotions correctly. And these features should have to convey the measurable level of emotional modulation.

Speech Emotion Recognition (SER) aims to automatically identify the current emotional state of a person from his or her speech [6]. In speech, discrete emotion expressions are associated with characteristic variations in the acoustic structure of the speech signals and the relative perturbation of specific acoustic cues over the course of an utterance [7]. During speech emotion analysis, these vocal cues are extracted from speech as a marker for the emotional state by assuming that there are objectively measureable cues that can be used for emotion recognition [8]. This led the researchers to study indepth on finding the most discriminative features that contribute to the performance of SER systems. In addition to that, there are some studies in literature which tried to find the factors that influence emotion identification from expressive speech such as gender and age. In analyzing the influences of language in SER, Pell [7] has highlighted the importance of acoustic data such as fundamental frequency and speaking rate for indicating vocal emotion in languages. However, how languages influence the SER is open for exploration.

In this study we investigate whether a Speech Emotion Recognition system can identify the emotional state of an individual regardless of the language used and the advantage of expressing emotion by native speakers and bilingual speakers. We adopt a Support Vector Machine (SVM) classifier to recognize six discrete emotions using acoustical features and explore how spoken expressions of the selected emotions varied in the two languages of interests.

2 Problem Description

Speech Emotion Recognition(SER) is the distinctive technique of recognizing emotions from speech by creating a dataset, extracting features from the dataset, selecting the most salient features amongst them, categorizing them into emotions and lastly classifying them using a trained model. Besides this, we are attempting to investigate whether a Speech Emotion Recognition System is Language Independent.

2.1 Speech Emotion Recognition

Emotion Recognition from Speech is usually done using corpus of agent client spoken dialogues from call centre like for medical emergency, security, prosody generation, etc. In linguistics, prosody is concerned with those elements of speech that are not individual phonetic segments but are properties of syllables and larger units of speech, including linguistic functions such as intonation, tone, stress, and rhythm. Such elements are known as suprasegmentals.

Another term of Corpus is the Database or Dataset. This database is used for training, testing and development of feature vector. A good database is important for desired result. Various databases are available created by speech processing community. The databases can be divided into training data set and testing data set. The famous databases are The Danish Emotional Speech Database (DES), and The Berlin Emotional Speech Database (BES), as well as The Speech under Simulated and Actual Stress (SUSAS) Database. For English, there is the 2002 Emotional Prosody Speech and Transcripts acted database. The databases that are used in Speech Emotion Recognition are classified into 3 types [2].

Type 1 is acted emotional speech with human labeling. Simulated or acted speech is expressed in a professionally deliberated manner. They are obtained by asking an actor to speak with a predefined emotion, e.g. DES, EMO-DB.

Type 2 is authentic emotional speech with human labeling. Natural speech is simply spontaneous speech where all emotions are real. These databases come

from real-life applications for example call-centers.

Type 3 is elicited emotional speech in which the emotions are induced with self-report instead of labeling, where emotions are provoked and self-report is used for labeling control. The elicited speech is neither neutral nor simulated.

Toronto Emotional Speech Set (TESS) is a dataset prepared for emotion recognition using two experiments. It is a type 1 Speech Emotion Recognition. This will be used as a standard for developing our Dataset.

After forming the speech corpus, feature extraction and selection dictates how efficient our SER system will be. Features are extracted using various assorted ways. It can be extracted through softwares like MATLAB, PRAAT etc. The most significant and cogent features are selected via multifarious Feature Selection Strategies. The most notable features used for a SER system are pitch, intensity, formants related features [1] and Mel Frequency Cepstral Coefficient (MFCC) features. The entire speech corpus together with its most salient features is categorized into discrete emotions such as anger, happiness, sadness, netral etc. during the dataset creation. Classifiers such as K-Nearest Neighbour(KNN), Support Vector Machine(SVM), Neural Networks, Gaussian Mixture Models(GMM), Hidden Markov Models(HMM) [2] are used for performing Speech Emotion Recognition. For our purpose we adopted SVM for measuring the recognition rates as well as to investigate the language independency in SER.

2.2 Language Independency in Emotion Recognition

In order to investigate the language independency in Speech Emotion Recognition we will be working with two languages: Bangla and English. These languages will be spoken by their respective native speakers. Additionally they will be spoken by bilingual speakers whose native language will be Bangla. Generally people generate assorted cues and prosody in speech. These cues are of great value in evaluating the emotional state of a person. These cues justify an individual's emotion. These cues in machine language terminology is known as features. We will also be correlating these features for both of the language under consideration.

Besides this, we will be scrutinizing the emotion recognition rates for both the language individually and unitedly. We will be considering the perspective native and bilingual speakers while evaluating the emotion recognition rates of Bangla and English.

3 Background Study

Speech Emotion Recognition requires dataset creation, feature extraction, feature selection and classifiers to perform SER. For dataset creation and feature extraction we adopted the software Praat which was developed in the Carnegie Melon University for recording audio samples and extracting features such as pitch, intensity, formants and spectogram related features. In this software we record an audio sample and process it in order to extract features such as Pitch Mean, Pitch Median, Standard Deviation, Jitter, Shimmer, Intensity and many more. For recording the audio samples we followed the Toronto Emotional Speech Set(TESS) as reference. They recorded 2800 samples by the assistance of two female speakers for 7 emotions(anger, happiness, fear, sadness, disgust, pleasant surprise and neutral).

For Speech Emotion Recognition the most significant area to work on are the features. Inefficient and insufficient choices of the features can cause overlaps and misclassification. There are two types of features that are widely used in Speech Emotion Recognition. They are Prosodic Features and Spectral Features. Prosodic features are pitch or energy related features and Spectral features can be categorized to Mel Frequency Cepstral Coefficient features. Prosodic features carry a large amount of information considering a user's emotion. The selected contours rely rather on broad classes of sounds while spectral characteristics in general seem to depend too strongly on phonemes and therefore on the phonetic content of an utterance [9]. This is a drawback thinking of the premise of independency of the spoken content or even the language.

The classifiers which are mostly adopted in Speech Emotion Recognition are K-Nearest Neighbour(KNN), Support Vector Machine(SVM), Neural Networks, Gaussian Mixture Models(GMM), Hidden Markov Models(HMM). In recent studies KNN, SVM and Neural Networks have been deployed for performing Speech Emotion Recognition.

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick.

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. KNN captures the idea of similarity (sometimes called distance, proximity, or closeness) by calculating the distance between points on a graph. However, the straight-line distance (also called the Euclidean distance) is a popular and familiar choice. The algorithm is simple and easy to implement. There's no need to build a model, tune several parameters, or make additional assumptions and the algorithm is versatile as it can be used for classification, regression, and search.

Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labeling or clustering raw input. The patterns they recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text or time series, must be translated. Neural networks help us cluster and classify. We can think of them as a clustering and classification layer on top of the data we store and manage. They help to group unlabeled data according to similarities among the example inputs, and they classify data when they have a labeled dataset to train on. Neural networks can also extract features that are fed to other algorithms for clustering and classification; so we can think of deep neural networks as components of larger machine-learning applications involving algorithms for reinforcement learning, classification and regression.

4 Related Works

In the following is a brief description of the related works for Emotion Recognition in Speech.

4.1 Influences of Languages in Emotion Recognition

What is the research area?

In the paper Influences of Languages in Speech Emotion Recognition: A Comparative Study Using Malay, English and Mandarin languages the aim is to investigate whether a SER system can identify the emotional state of a person regardless of the language used. To check the influence of languages in SER, we examined how spoken expressions of four selected emotions (anger, sadness, happiness and neutral) varied in the three languages such as Malay, English and Mandarin. While supporting the fact that SER is language independent, the study reveals that there are language specific differences in emotion recognition in which English shows a higher recognition rate compared to Malay and Mandarin. This study also demonstrated that emotions expressed by native speakers have higher accuracy rates.

Summary

This paper investigates whether a SER system can identify the emotional state of an individual regardless of the language used and the advantage of expressing emotion by native speakers. They adopt a k-Nearest Neighbor (kNN) classifier to recognize four discrete emotions using acoustical features and explore how spoken expressions of the selected emotions varied in the three languages of interests. A standard machine learning approach, which is SER, was followed. It is divided into three main tasks; database preparation, feature extraction and classification. In general the SER process includes the following steps:

- Prepare data set
- Identify emotion classes
- Extract appropriate features from speech signal

- Develop emotion recognition models using the feature vectors
- Use models to recognize emotions

The proposed experiments were conducted using Berlin Emotional database (BES) which is a simulated speech database. It covers the emotion anger, boredom, disgust, fear, joy, neutral and sadness. All sentences are acted in all seven emotional states by professional actors (five males and five females).

For the purpose of this research the most commonly used acoustical features; MFCC, pitch, energy and zero crossing rate were used.

Two sets of experiments were carried out to evaluate the influence of language in SER. In the first experiment, a total of 10 participants (5 male and 5 female students) from Theater and Drama society, Nilai University were randomly selected and asked to generate 120 utterances. All of them are Malaysians and have completed primary and secondary education in Malaysia. 12 different utterances, in four different emotional states and in three different languages were recorded for each speaker. The second experiment was carried out to demonstrate the advantage of identifying speech emotion expression when produced by native speakers. 10 native Mandarin speakers (5 males and 5 females) and 10 native Malay speakers (5 males and 5 females) were asked to generate 80 utterances.

The classification rates were calculated by dividing the total number of emotions belonging to a class in testing with the total number correctly recognized emotion of that class.

SER for English shows highest recognition rate compared to Malay and Mandarin. This could be because the participants were from English speaking environment and those could express their feelings well in English.

The comparisons focused on three languages which are English, Malay and Mandarin. The data analysis from this study shows that discrete emotions can be identified from the three languages at accuracy levels of 78.5%, 71.0% and 72.5%. Although recognition of emotions was reliable, the accuracy rate varied in the

Language	Classification Rate				
	Happiness	*Sadness*	*Anger*	*Neutral*	*Average*
English	75	86	83	70	78.5
Malay	70	75	74	65	71.0
Mandarin	75	71	74	70	72.5
Average	73.3	77.3	77	68.3	74.0

Figure 1: Classification rate for randomly selected participants

Language	Classification Rate				
	Happiness	*Sadness*	*Anger*	*Neutral*	*Average*
Malay	73	78	77	71	74.75
Mandarin	78	80	77	69	76
Average	75.5	79	77	70	75.4

Figure 2: Classification rate for native speakers

three languages of interest when the emotions were uttered by randomly selected participants. This shows that there are language specific differences in emotion recognition in which English shows higher recognition rate compared to Malay and Mandarin. This study also demonstrated that emotions expressed by native speakers have higher accuracy rates.

4.2 Processes in Developing SER Systems

What is the research area?

In the paper Analysis of Speech Features for Emotion Detection: A review, a general overview of developing SER systems is given. In general, emotion detection system consist of speech normalization, feature extraction, feature selection, classification and then the emotion is detected.

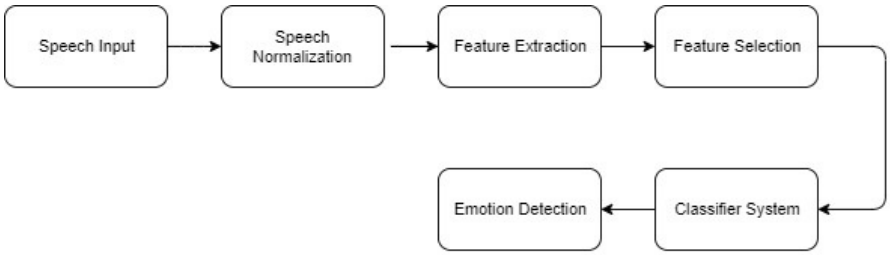


Figure 3: System for Emotion Detection of Speech Signal

Summary

The databases that are used in SER are classified into 3 types. First one is acted emotional speech with human labeling. Simulated or acted speech is expressed in a professionally deliberated manner. They are obtained by asking an actor to speak with a predefined emotion. Second one is authentic emotional speech with human labeling. Natural speech is simply spontaneous speech where all emotions are real. These databases come from real-life applications for example call-centers. Third one is elicited emotional speech in which the emotions are induced with self-report instead of labeling, where emotions are provoked and self-report is used for labeling control. The elicited speech is neither neutral nor simulated. After data collection we extract and select the features. These emotional speech features can be classified into different categories. One classification is long term features and short term features. The short term features are the short time period characteristics like formants, pitch and energy. And long term features are the statistical approach to digitized speech signal. Some of the frequently used long term features are mean and standard deviation. The larger the feature used the more improved will be the classification process. After extraction of speech features only those features which have relevant emotion information are selected. These features are then represented into n- dimensional feature vectors. The prosodic features like pitch, intensity, speaking rate and variance are important to identify the different types of emotions from speech. In Table 1 acoustic characteristics of various emotions of speech is given. The observations which are expressed in below table 1 are taken by using Praat software.

Characteristics / Emotion	Happy	Anger	Enquiry	Fear	Surprise
Pitch Mean	High	Very high	High	Very high	Very high
Pitch Range	High	High	High	High	High
Pitch Variance	High	Very high	High	Very high	Very high
Pitch Contour	Incline	Decline	Moderate	Incline	Incline
Speaking Rate	High	High	Medium	High	High

Figure 4: Accoustic Characteristics of Emotions

Various classifiers like GMM, HMM are used according to their specific usage based on selected features. Emotions are predicated using classifiers and selected feature vectors to predict emotion from training data set and the development data set. For the training data sets the emotion information are known whereas for testing data set the emotion information are unknown. When performing analysis of complex data one of the major problems comes from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power or a classification algorithm which over fits the training sample and generalizes poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still the data with sufficient accuracy. Typically, in speech recognition, we divide speech signals into frames and extract features from each frame. During feature extraction, speech signals are changed into a sequence of feature vectors. Then these vectors are transferred to the classification stage.

4.3 Analysis of Language Independent Features

What is the research area?

The paper Exploring Language-Independent Emotional Acoustic Features via Feature Selection proposes us a selection strategy to discover language-independent acoustic features that tend to be responsible for emotions regardless of languages, linguistics and other factors. In general, selected features are effective for a single corpus only and not generalized to other corpora in general. Unlike the previous

work, this paper aims at discovering those acoustic features that tend to be responsible for emotions regardless of linguistics and other factors and hence can be generalized to other corpora. In this paper, they name such features as language-independent emotional acoustic features.

Summary

In this paper, feature selection techniques are applied to a training corpus, and selected features are then tested on corpora of different languages and designs. A single feature selection method often biases to some certain aspects and fails to identify all language independent features. Thus, the use of selected features results in a lower recognition rate on a different corpus. As three feature selection techniques are used, three feature subsets are generated respectively. As a result, we combine three feature subsets by taking their union or intersection to form two combined subsets of selected features. Then we test combined feature subsets on all the corpora except the one used for feature selection and choose the combined feature set that gives better recognition rates on all the test corpora. There are often inconsistent emotional states in different emotional speech corpora. For this purpose, reconciling emotional states is required in each experiment. A reconciliation scheme may remain

a number of emotional states common to all the corpora and/or re-group other emotional states, which results in an emotional state alignment so that all corpora have the same emotional states. They collected a set of 318 potentially useful acoustic features. All 318 features referred to as the full feature subset. The joint use of those features in different ways forms two types of representations, utterance-based and segment-based representations. An utterance-based representation treats an utterance as a whole, and hence its representation is formed for the entire utterance. This representation mainly characterizes those global features captured by human listeners. The utterance-based representation used in our work includes all features in the full feature set. A segment-based representation blocks an utterance into several segments and a feature vector is extracted for each segment. This representation tends to capture critical local features un-

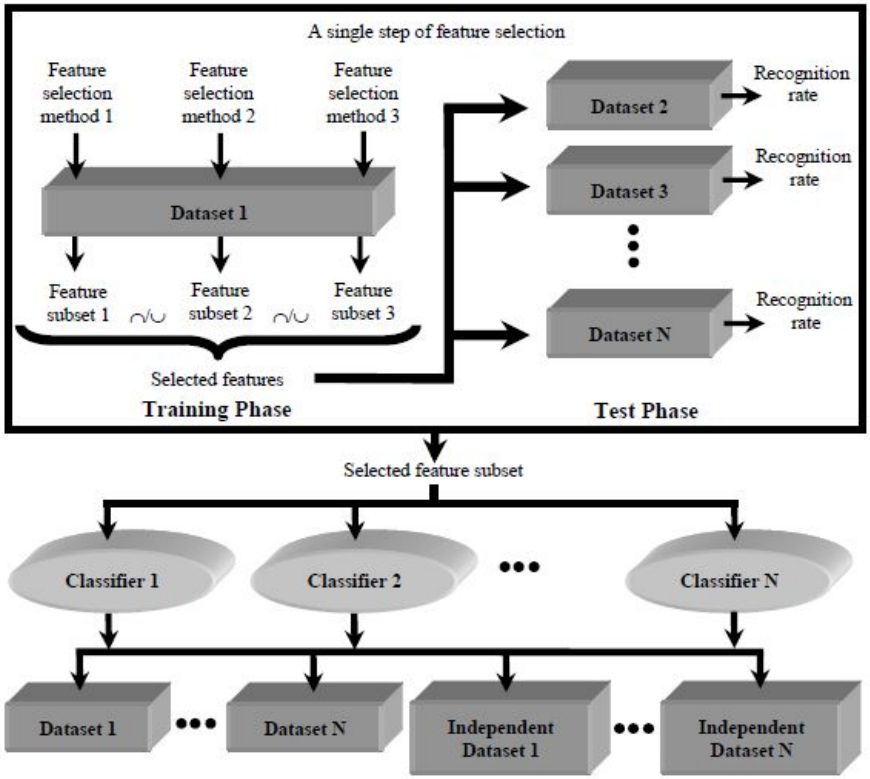


Figure 5: Schematic illustration of feature selection strategy

derlying an utterance especially as emotional information is unevenly distributed. In this paper, they used four speech corpora in four different languages especially designed for emotional speech studies. They used the Berlin emotional speech (BES), Danish emotional speech (DES), The Serbian emotional speech corpus named GEES and The BabyEars emotional speech corpus. In their experiments, first three corpora are employed for feature selection while the BabyEars corpus is simply used as a dataset independent of feature selection to monitor the stability of selected feature subsets. Happy, sad, anger and neutral were the emotions used as these are common in all the speech corpus. For feature selection, we employ three different techniques; i.e., sequential floating forward selection (SFFS), genetic algorithm (GA) and boosting based feature selection. Top 20 features listed are six low-pass intensity and 14 pitch related features, and most of them are also selected as language-independent features by our method. Finally, five and six formant related features are picked for utterance-based and segment-based

representations, respectively.

4.4 Developing a Bangla Emotional Speech Set

What is the research area?

In the paper, Recognition of Emotional Speech for Younger and Older Talkers: Behavioral findings from the Toronto Emotional Speech Set, one goal was to create a set of stimuli with well-controlled lexical and semantic properties based on an existing test of speech intelligibility, the Northwestern University Auditory Test-Number 6 (NU-6 [3]), so that the lists of stimuli in the set are balanced for properties such as word frequency as well as word and syllable length. Experiment 1 provides a description of the actors, recording process, and stimulus selection process used for the creation of the novel set of stimuli, the Toronto Emotional Speech Set (TESS). In Experiment 2 recognition rates for the emotions portrayed in these stimuli were determined for a group of healthy younger listeners.

Brief Overview

Two female actors, one younger and one older, were recruited from the community. Respectively, they were 26 and 64 years of age. The actors consented to create voice recordings which would be used as stimuli for research purposes, in educational presentations at scientific or professional conferences or in public education or community presentations. Both actors spoke English as a first language and had clinically normal hearing thresholds in the speech range (see Table 1 for demographic characteristics of the actors). The recording stimuli were the 200 items from the NU-6 test. Each item begins with the same carrier phrase and terminates in a monosyllabic noun (e.g., "Say the word bean"). The actors recorded each item to portray seven different emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). These seven emotions were chosen because they were recently used by two groups of researchers to create sets of German and Portuguese sentences. In this way we could extend work on affective prosody understanding for these seven emotions to the English language. A final set of

2800 set of samples were recorded. For each emotion 200 samples were recorded and there was 7 emotions in total. Thus 200 * 7 * 2 = 2800 samples were recorded for two speakers.

Fifty-six undergraduate students at the University of Toronto were tested in this experiment. All participants spoke English as a first language and had clinically normal hearing thresholds from 250 to 8000 Hz (see Table 2 for participant characteristics). Participants listened to stimuli spoken either by the younger or by the older talker. Each participant listened to an equal number of stimuli spoken in each of the seven emotions. The stimuli were presented through a loudspeaker in a sound-attenuating booth at an average presentation level of 70 dBA. In response to each stimulus they used a touch computer screen to indicate which emotion the talker was portraying.

Significant Contributions

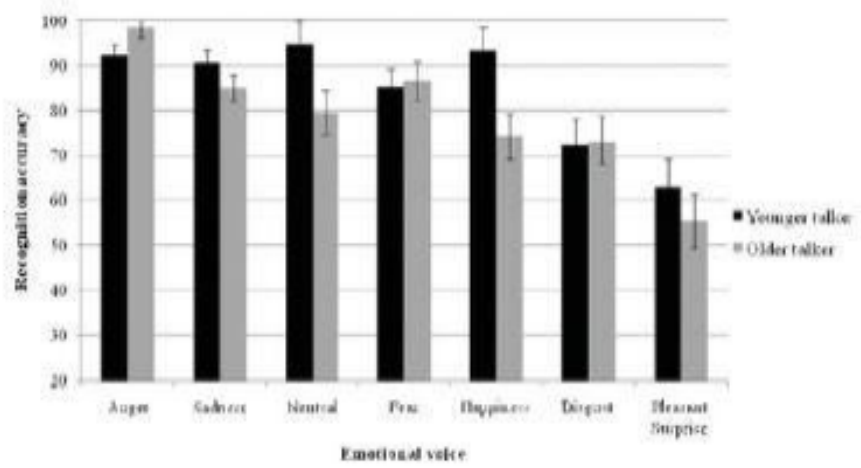
A standard dataset containing 2800 standard samples were recorded which came to be known as TESS which is the Toronto Emotional Speech Set. This speech set is used for recognizing emotions such as Anger, Happiness, Neutral, Sad, Surprise, Disgust and Fear. It used the North Western University Auditory Test no 6 for building the dataset. This NU – 6 contains the phonetic content of the standard list of words.

Results

The primary measure of interest was the percentage of correctly recognized emotions. The overall accuracy was 82%. Stimuli spoken to portray anger and sadness had the highest accuracy while stimuli spoken to portray disgust and pleasant surprise had the lowest accuracy.

Discussion, Future Scope, Limitation

The results from this experiment indicate that participants were able to recognize the emotions portrayed in the TESS stimuli with very good accuracy. The accuracy rate of 82% was almost six times greater than chance and higher than the 55-65% level described in recent reviews of studies in this field that used sentences with similar emotions [7, 8]. Furthermore, the lack of a main effect of talker



indicates that the two actors created highly recognizable portrayals of the seven different emotions. Although overall recognition of the emotions was high, there were significant differences in the accuracy with which some emotions were recognized. Consistent with previous findings angry and sad emotions had the highest recognition rates overall.

4.5 Selection of Salient Features for SER Systems

What is the research area?

In Hidden Markov Model-based speech emotion recognition we introduce speech emotion recognition by use of continuous hidden Markov models. Two methods are propagated and compared throughout the paper. Within the first method a global statistics framework of an utterance is classified by Gaussian mixture models using derived features of the raw pitch and energy contour of the speech signal. A second method introduces increased temporal complexity applying continuous hidden Markov models considering several states using low-level instantaneous features instead of global statistics. The paper addresses the design of working recognition engines and results achieved with respect to the alluded alternatives. A speech corpus consisting of acted and spontaneous emotion samples in German and English language is described in detail. Both engines have been tested and trained using this equivalent speech corpus. Results in recognition of seven dis-

crete emotions exceeded 86% recognition rate. As a basis of comparison the similar judgment of human deciders classifying the same corpus at 79.8% recognition rate was analyzed.

Extraction of the Raw Feature Contours

We chose the analysis of the contours of pitch and energy for their well-known capability to carry a large amount of information considering a user's emotion. The selected contours rely rather on broad classes of sounds while spectral characteristics in general seem to depend too strongly on phonemes and therefore on the phonetic content of an utterance. This is a drawback thinking of the premise of independency of the spoken content or even the language. In order to calculate the contours, frames of the speech signal are analyzed every 10ms using a Hamming window function. The values of energy are calculated by the logarithmic mean energy within a frame. The pitch contour is achieved by the use of the average magnitude difference function (AMDF).

Method 1: Global Statistic Using GMMs

Within the first method we derive 20 features of the underlying introduced raw contours. In general the introduced features have been chosen accepting speaker dependent recognition aiming at optimal results. The features concerning temporal aspects such as the rate of voiced sounds, are approximated with respect to zero levels in pitch contour due to the inharmonic nature of unvoiced sounds. The features are freed of their mean value and normalized to their standard deviation. They are classified by single state HMM's (GMM), which are able to approximate the probability distribution function of each derived feature by means of a mixture of Gaussian distributions. Up to four mixtures have been used. No further gain could be observed using more than these. Each emotion is modeled by one GMM in our approach. The maximum likelihood model will be considered as the recognized emotion at a time throughout the recognition process.

Method 2: Continuous HMMs

Within the second method we strive to increase the temporal complexity and use the warping capability of hidden Markov models by introduction of more states in

the models. Since global statistics are clearly invalid for this purpose, one has to carefully consider suited features. The continuous HMMs (CHMM's) were trained using Baum Welch re-estimation with a maximum of 10' iterations or an abruption criterion of a change in model parameters of e ¡ 10-4. Up to four Gaussian mixtures have been used to approximate the emission probability density functions according to the GMM solution. The HMM types were chosen as Left-Right-models, as in usual speech processing, which ideally models advances in time.

Evaluation Result

	Recognized emotion						
Labeled emotion	ang	dis	fea	sur	joy	ntl	sad
ang	**91.1**	1.2	0.6	6.3	0.4	0.1	0.3
dis	5.4	**76.8**	6.7	0.1	6.8	3.2	1.0
fea	0.2	6.4	**82.8**	0.6	3.0	0.3	6.7
sur	2.4	2.2	3.0	**87.2**	4.6	0.1	0.5
joy	3.0	0.7	0.8	0.0	**93.2**	0.2	2.1
ntl	0.2	3.4	0.4	0.5	2.7	**89.6**	3.2
sad	0.2	0.1	5.8	3.8	0.4	2.2	**86.6**

	Recognized emotion						
Labeled emotion	ang	dis	fea	sur	joy	ntl	sad
ang	**68.5**	12.7	2.6	1.8	2.7	8.4	3.3
dis	12.8	**84.7**	2.1	0.3	0.0	0.1	0.0
fea	1.8	0.1	**95.4**	0.2	2.0	0.4	0.1
sur	6.3	6.7	6.3	**73.5**	6.1	0.9	0.2
joy	10.1	11.8	7.9	1.2	**68.0**	0.5	0.5
ntl	10.4	0.9	1.0	0.1	1.9	**79.6**	6.1
sad	5.9	10.1	2.8	2.1	2.2	1.8	**75.1**

Here from the confusion matrices we can evaluate the results properly. The first matrix represents the confusion matrix of the first method i.e global statistics using GMMs. We can see that the accuracy of anger is 91.1%, disgust is 76.8%, fear is 82.8%, surprise is 73.5%, joy is 93.2%, neutral is 89.6% and sad is 86.6%. For the second matrix we can see that it is the confusion matrix of the continuous HMM method. The accuracy of anger is 68.5%, disgust is 84.7%, fear is 95.4%, surprise is 73.5%, joy is 68.0%, neutral is 79.6% and sad is 75.1%.

Challenges, limitations and future scope

We believe that this contribution shows important results considering emotion

recognition with hidden Markov models. The two introduced methods proved both capable of a rather reasonable model for the automatic recognition of human emotions in speech. The confusion matrices clearly show that some emotions are often confused with certain others. Furthermore some emotions seem to be recognized more easily. This may be due to the fact that the most test patterns were acted emotions and test-persons have difficulties with feigning certain emotions. Though the same training material and test sets were used, the two proclaimed solutions differ greatly in their behavior. Neither the confusion of emotions nor the performance of recognition of single emotions itself shows significant correlations in the result. While the global phrase statistics outperformed the instantaneous features, still both propagated solutions build a reasonable model. One reason for the better performance can be seen in the loss of information of durations of voiced sounds by eliminating these in the contours as described. The results of both engines reach the abilities of a human decider as described above. In our future work we aim at a hybrid approach combining neural networks and hidden Markov models for the automatic recognition.

5 Proposed Approach

5.1 Architectural Block

Speech Emotion Recognition contains feature extraction from a established or created corpus. In this study we created a dataset containing of English and Bangla language spoken by native and bilingual speakers. Afterwards we have selected the most salient features which are widely used in SER systems [1] [2] [3] [9].

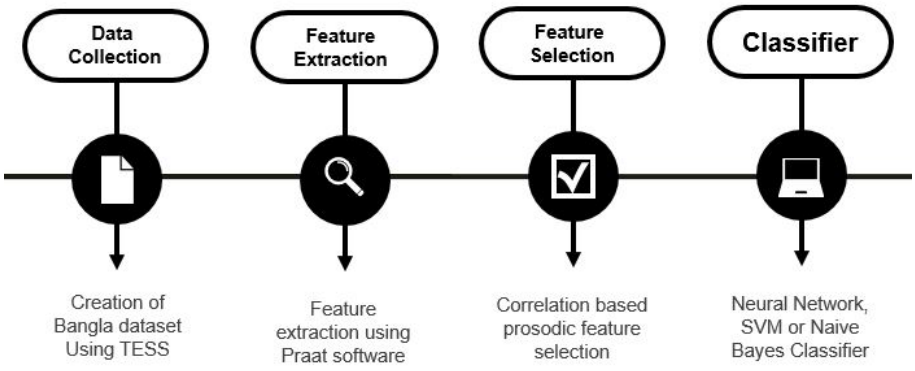


Figure 6: Architectural Block for a Speech Emotion Recognition System

We used four distinct features(pitch mean, pitch median, standard deviation & intensity) to identify six emotional states(anger, happiness, neutral, sad, fear & disgust) [4][9][10]. We used the PRAAT software for feature extraction. Afterwards we adopted KNN and SVM for Speech Emotion Recognition [1] [2] [3] of which SVM gave better recognition rate than KNN for our speech corpora. Henceforth we adopted SVM for this study.

In order to investigate whether the SER systems are independent of language we conducted two experiments. Firstly, we performed the SER on Bangla, English(spoken by a Bangla speaker), and English_TESS(spoken by a Canadian speaker) individually. Secondly, we performed the SER on Bangla-English and

Bangla-English_TESS unitedly, to scrutinize whether the SER systems are independent of language.

5.2 Dataset/Speech Corpora Creation

The success of a Speech Recognition System largely depends on the dataset that is being used. In this study we created a Bangla and English dataset from 11 speakers(7 male, 4 female) who are currently studying in Islamic University of Technology. Since there is no established Bangla Speech Corpora we followed the technique adopted to create the Toronoto Emotional Speech Set(TESS) [10].

We identified six emotional states such as Anger, Happiness, Neutral, Sad, Fear & Disgust for our speech corpora. Each emotion consists of 50 audio samples. Therefore for the two datasets we have a total of 50 * 6 * 2 = 600 audio samples; 300 for each dataset(English and Bangla). These were created by such speakers who were fluent in both English and Bangla. On the other hand the same samples from the Toronto Emotional Speech Set(TESS) were retrieved and another dataset named English_TESS was manifested. As already established these dataset was created with the help of a Canadian Actor. Therfore, we have established three datasets which we will be calling as English, Bangla and English_TESS. The first two were developed by bangla speaker and the latter was formed by a native English speaker.

Each data sample were recorded within 1-2 sec. Each sample contained the phrase "Say the Word ..." which was followed by a word. For instance, a phrase used in the dataset was "Say the word read". For developing the Bangla dataset the phrase was translated to Bangla and it read as "Poro shobdo ti bolo" which translates to "Say the word read" in English. Thus using this procedure 300 audio samples for Bangla as well as English were created.

5.3 Feature Extraction

Feature Extraction as well as recording audio samples were done via a software known as PRAAT. Firstly, in the software we go to "New" section and record an audio sample. Then we go to the "View & Edit" section to process the audio file.

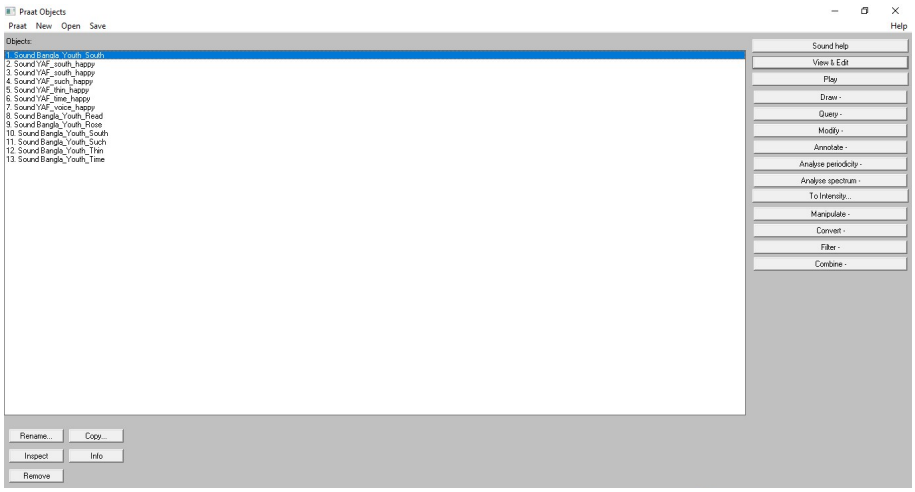


Figure 7: Interface of Praat Software

We select the duration of the audio sample and retrieve the features.

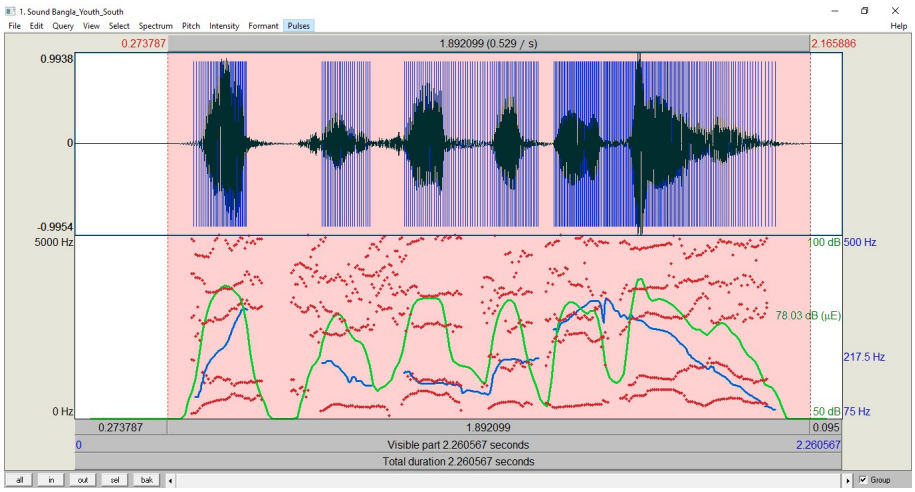


Figure 8: Processing of Audio Sample

After this we obtain the features by moving to the "Pulse" section and selecting the voice report option.

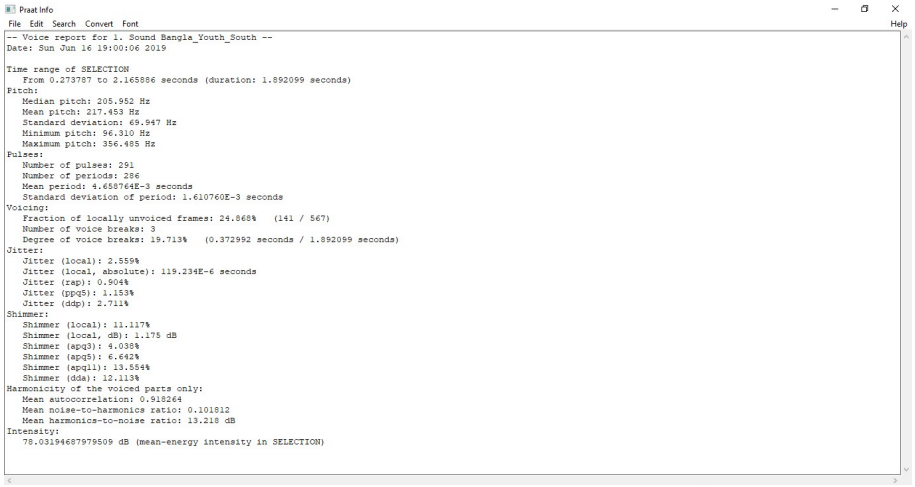


Figure 9: Features Extracted from Praat

We get assorted features like pitch, shimmer, jitter, intensity, voicing, harmonicity, pulses etc.

5.4 Feature Selection

Features dictate the performance for a Speech Emotion Recognition system. The most salient features can improve the efficiency of a model exponentially. Therefore, we selected 4 features from the extracted features from Praat. The 4 features were Pitch Median, Pitch Mean, Standard Deviation & Intensity [2].

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	THESIS DATA (Fear Set)												
2	Words	Median pitch (*Hz*)			Mean pitch (*Hz*)			Standard deviation (*Hz*)			Intensity (*dB*)		
3		English	English(*own*)	Bengali	English	English(*own*)	Bengali	English	English(*own*)	Bengali	English	English(*own*)	Bengali
4	Read	334.907	226.834	164.867	316.455	237.423	181.873	60.005	52.539	78.483	69.549	76.337	60.449
5	Rose	312.059	233.339	161.224	313.612	234.605	166.345	61.643	29.812	50.270	69.925	79.161	75.499
6	South	323.130	240.315	155.043	326.859	254.779	176.876	73.762	75.240	65.325	68.328	78.880	63.132
7	Such	312.231	244.665	169.523	321.646	238.215	175.216	53.469	48.196	36.630	69.093	76.302	65.022
8	Thin	317.177	234.278	169.746	319.256	242.016	167.939	55.873	48.422	26.645	69.506	77.024	65.584
9	Time	344.896	252.005	183.077	333.593	259.985	190.356	62.888	58.203	69.570	68.642	76.549	65.347
10	Voice	304.630	239.615	183.086	311.549	244.026	186.768	54.662	64.406	54.266	67.942	76.667	66.188
11	White	319.881	229.803	172.668	315.238	233.118	171.976	50.150	53.756	21.208	67.936	78.946	63.818
12	Yes	324.925	241.082	176.431	311.541	239.280	174.451	45.859	52.803	27.704	68.929	77.658	64.928
13	Youth	334.907	252.696	182.132	316.455	244.006	184.086	60.005	62.279	51.417	69.549	80.366	64.413
14	Hate	306.422	181.261	144.140	296.450	177.251	139.781	55.474	35.902	14.545	70.268	83.262	80.215
15	Hole	311.155	153.421	179.064	303.108	152.116	181.448	50.175	17.532	18.181	68.376	80.765	83.033
16	King	327.966	140.770	158.426	313.459	146.307	159.058	48.904	24.093	11.307	70.418	80.122	83.131
17	Late	305.940	162.968	195.493	313.238	165.306	190.757	55.375	19.349	16.359	69.725	84.357	83.966
18	Lid	326.576	197.329	191.793	326.455	193.890	193.627	55.296	25.818	24.165	70.222	83.909	82.905
19	Make	340.470	183.013	179.957	332.258	177.937	178.514	54.724	22.322	24.705	68.308	84.278	84.681
20	Nice	290.666	169.045	161.406	293.849	160.111	157.214	54.205	24.222	25.139	68.984	82.861	81.485
21	Page	311.317	213.016	198.212	308.749	204.867	191.344	58.252	27.651	31.338	69.058	83.363	82.183
22	Rat	311.310	200.303	167.084	305.660	196.189	175.718	45.278	19.587	32.851	69.641	83.897	82.464
23	Rot	308.669	155.551	195.889	299.445	150.224	201.286	50.418	18.040	32.173	72.109	84.833	81.714
24	Sheep	293.789	197.939	131.434	297.789	198.099	131.373	49.016	15.306	18.546	71.148	85.359	81.847
25	Sour	319.716	187.788	144.995	319.167	187.901	148.323	55.782	21.381	27.104	70.372	84.958	80.023
26	Tire	325.146	172.996	146.356	325.179	166.049	145.770	52.164	24.163	21.515	68.361	82.981	79.415

Figure 10: Salient Features

Pitch is known as a prosodic feature which can be characterized by the fundamental frequency [11] [12]. It is the lowest frequency component, and contains

speaker-specific information. It has been used in many studies on vocal emotion recognition, such as [13] [17]

Intensity refers to the loudness of the speech signal s. It is measured at the position of the syllable peak,which is most commonly a vowel [14]. Examples of articles that utilized intensity as an important feature for vocal emotion recognition are [15] [18].

Standard deviation of the speech signal s, std, is also used as one of the acoustic features [16].

5.5 Classifier

The actual emotion recognition in SER is performed by a classifier, which had been trained on the data set. In this study we used Support Vector Machine (SVM) classifier, which is one of the traditional classifiers and often used in SER. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick.

6 Experiments

We have developed three datasets with the help of two types of speakers. With these datasets we have conducted three experiments. The speakers(Bangla & English) and their role in forming the datasets are discussed below.

- Bangla speakers form two datasets. One dataset is formed in Bangla language and the other is formed in English language.
- Let's call this datasets as "Bangla" and "English" datasets which were both originated from the Bangla speakers.
- We used another dataset, famously known as TESS or the Toronto Emotional Speech Set which was developed by Canadian speakers.
- Let's call this dataset as "English_TESS".

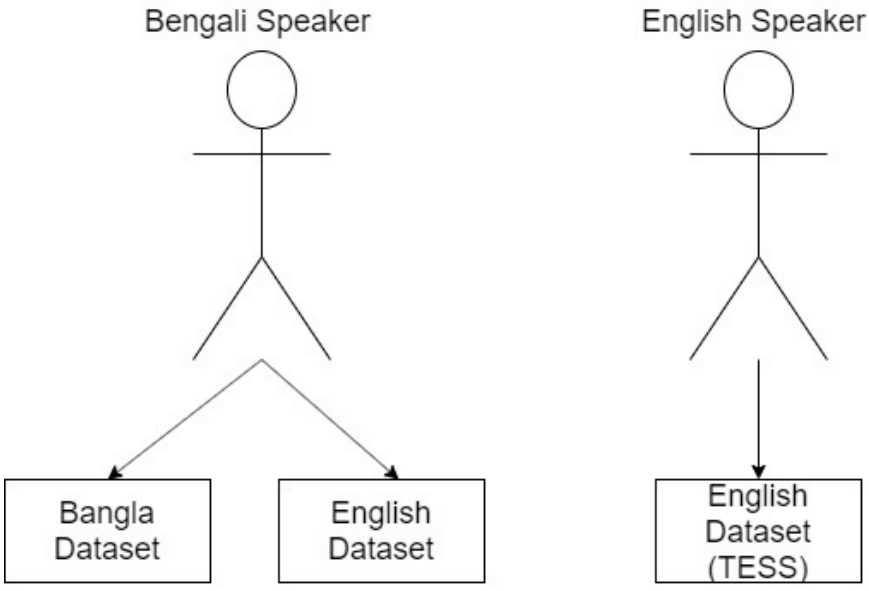


Figure 11: Speakers forming Datasets

The three experiments conducted on the datasets are explained in the following.

1. The first experiment was conducted on the individual datasets where training and testing were done by that same dataset.

2. The second experiment was conducted on the integrated datasets(Bangla-English & Bangla-English_TESS) where training and testing were done by that same dataset.

3. The third experiment was conducted on Bangla and English_TESS datasets. They were used for training while the other two datasets were used for testing. For instance if Bangla was used for training then English and English_TESS were used for testing the dataset.

6.1 Experiment 1: Individual Speech Corpora

In this experiment the Speech Emotion Recognition was performed on the individual datasets(English, Bangla & English_TESS). The English and Bangla dataset were spoken by speakers who were fluent in English and Bangla and the English_TESS was developed by a native Canadian speaker.

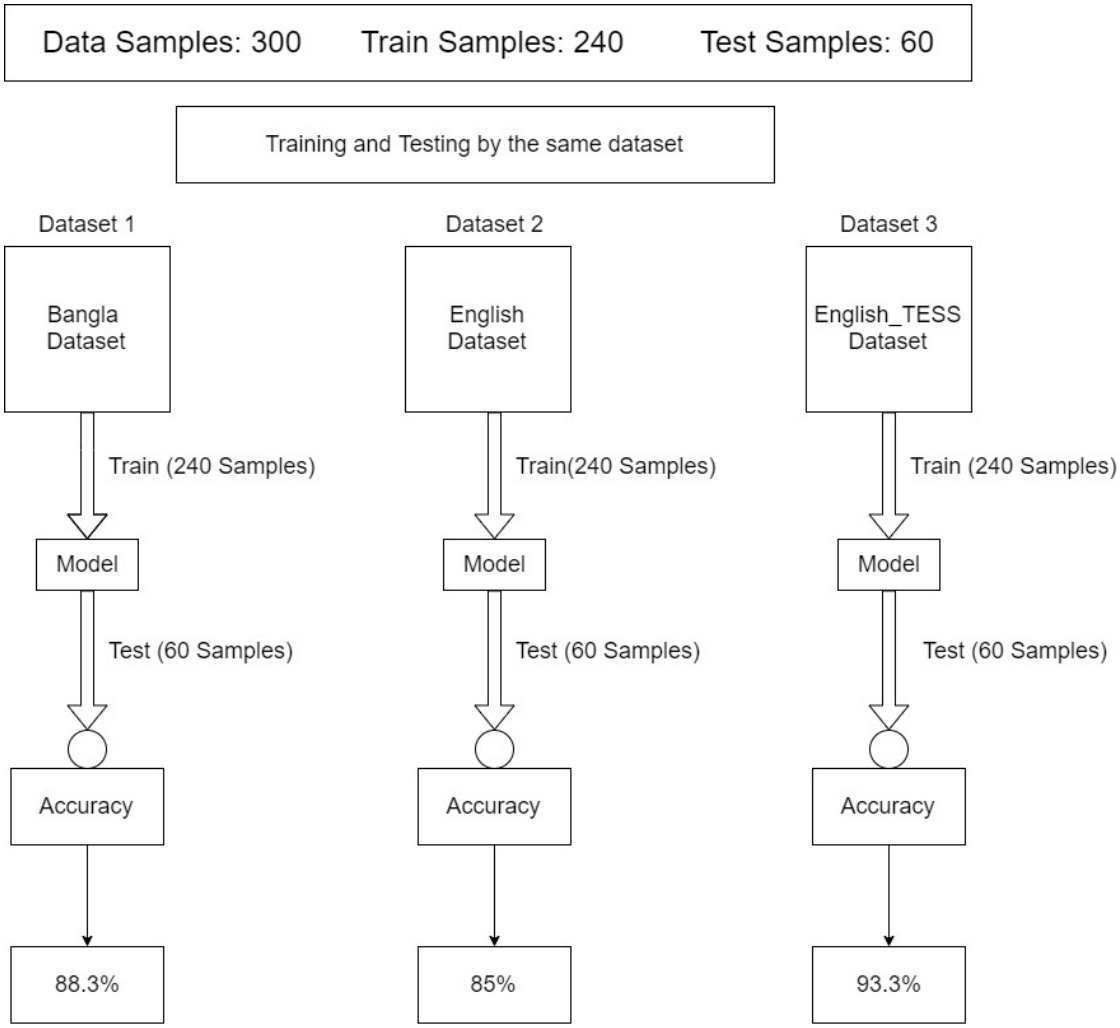


Figure 12: Experiment1

SVM was adopted for training the dataset and a model was built based on that which was later used in the testing phase. These datasets each have a total of 300 audio samples. For training purpose, 240 out of the 300 audio samples were taken; 40 samples each for the 6 emotions(40 * 6 = 240). For the testing purpose the remaining 60 audio samples were taken; 10 samples for each of the 6 emotions.

6.2 Individual Speech Corpora Experimental Evaluation

Here we represent the results of emotion recognition on individual datasets. Training and testing were done on the individual datasets.

6.2.1 Bangla Dataset Confusion Matrix

For the Bangla dataset which was developed by Bangladeshi speakers had a emotion recognition rate of 88.3% where the emotional state, anger was detected perfectly. 9 out of the 10 samples were detected for happy, neutral , sad and disgust were recognized accurately. Fear emotion was detected moderately in comparison to the other emotional states.

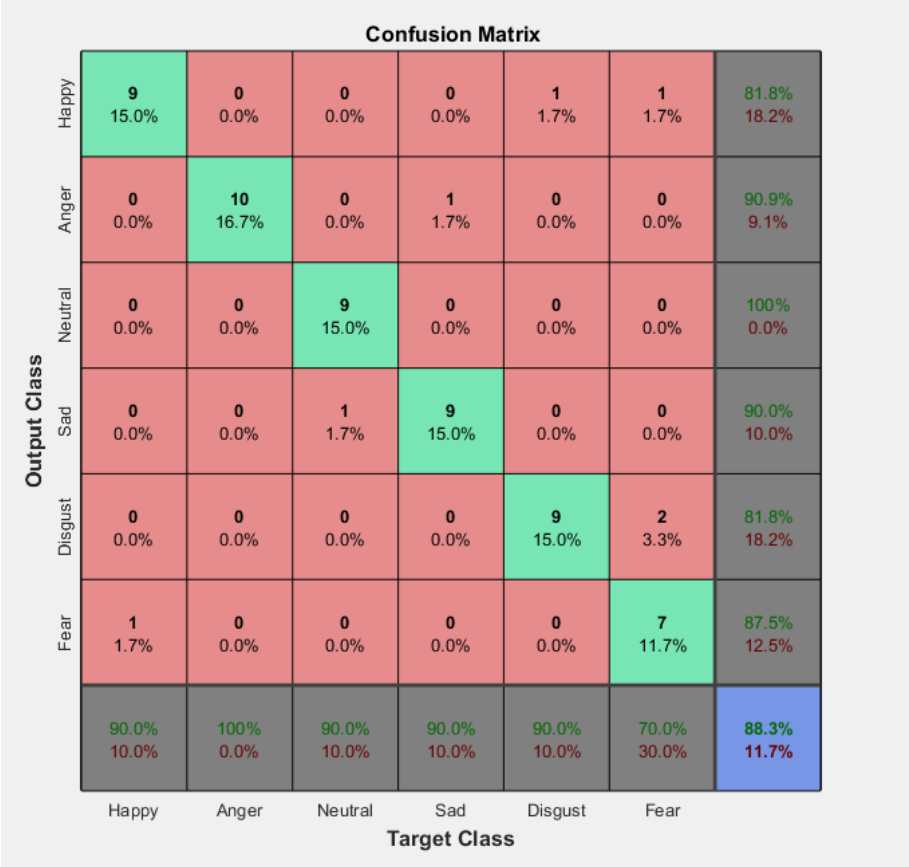


Figure 13: Bangla Testing Confusion Matrix

6.2.2 English Dataset Confusion Matrix

The English Dataset had a emotion recognition rate of 85% where the emotional state neutral and sad were detected perfectly. 8 out of the 10 samples were detected for disgust, fear and anger and the emotion was detected moderately in comparison to the other emotional states.

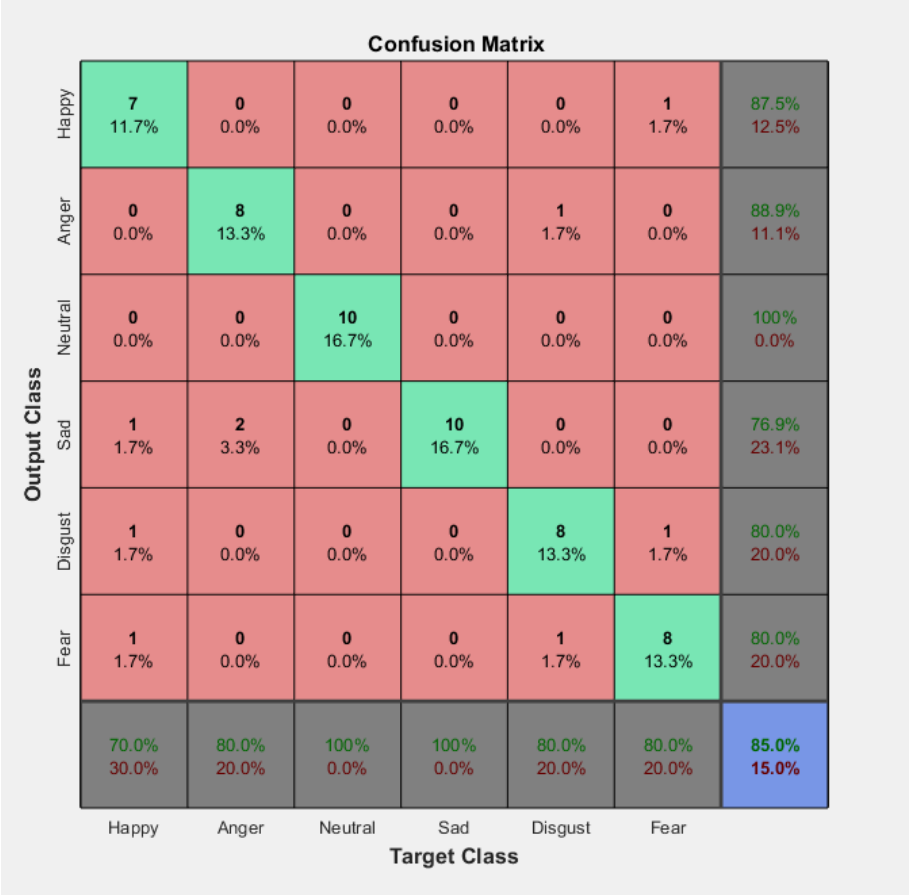


Figure 14: English Testing Confusion Matrix

6.2.3 English_TESS Confusion Matrix

This dataset had a staggering emotion recognition rate of 93.3% where the emotional state neutral and fear detected perfectly. 9 out of the 10 samples were detected for happy, anger, sad and disgust was recognized accurately. Almost all of the emotional states were accurately detected.

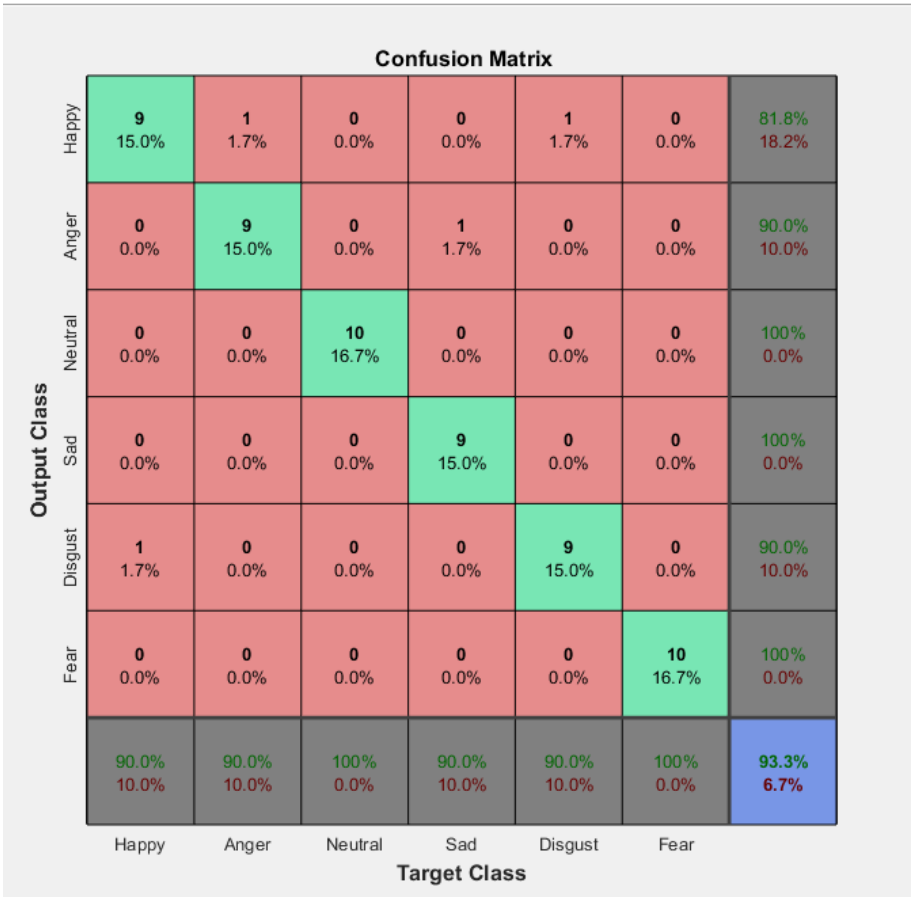


Figure 15: English_TESS Testing Confusion Matrix

6.3 Experiment 2: Integrated Speech Corpora

In this experiment SER was performed on integrated datasets of Bangla-English and Bangla-English_TESS. The datasets contained 600 audio samples of which 480 were applied in training whilst 120 samples were used for testing. Each testing dataset consisted of 60 audio samples of Bangla and 60 audio samples of English.

For the Bangla and English dataset a total of 85% emotion recognition was obtained whereas on the other hand for the Bangla and English_TESS dataset a total of 83.3% accuracy was achieved.

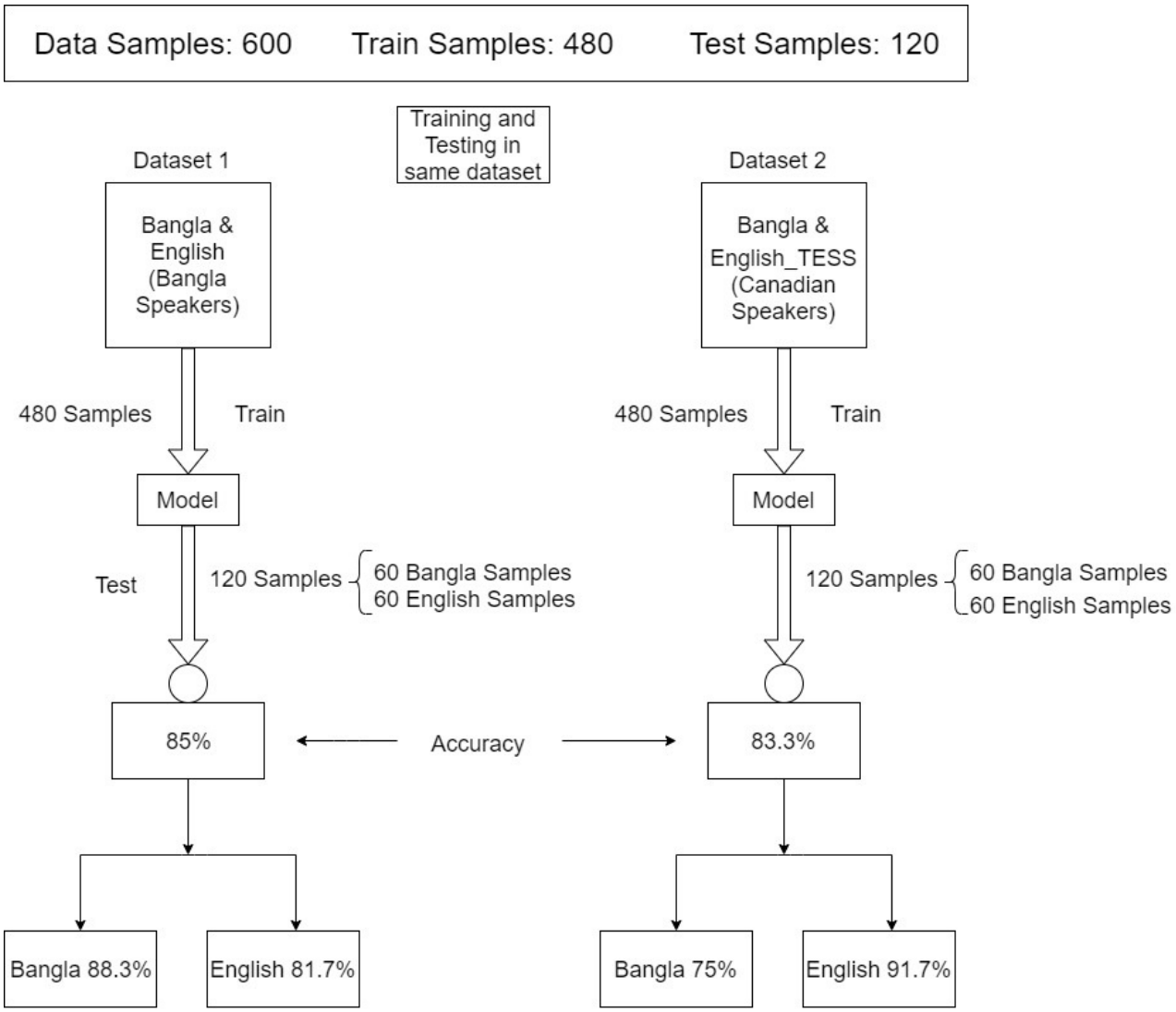


Figure 16: Experiment2

6.4 Integrated Speech Corpora Experimental Evaluation

Here we represent the results of training and testing on the integrated datasets.

6.4.1 Bangla-English Dataset Confusion Matrix

A total of 85% emotion recognition rate was achieved for this dataset. The emotions Happy, Anger, Sad and Neutral was detected accurately. Amongst them Neutral was detected most accurately. The emotions fear and disgust scored a very poor emotion recognition rate

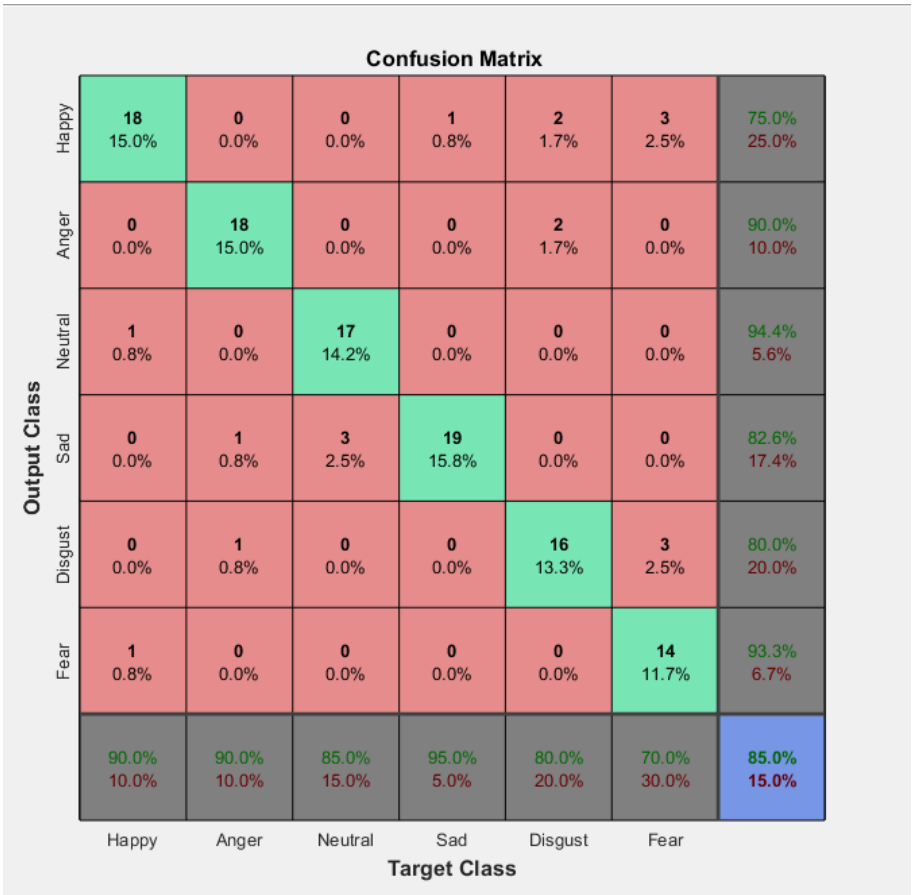


Figure 17: Bangla-English Testing Confusion Matrix

The emotion recognition performance for only the Bangla test samples were exemplary for the Bangla-English dataset. The Bangla audio samples had a emotion recognition rate of 88.3%.

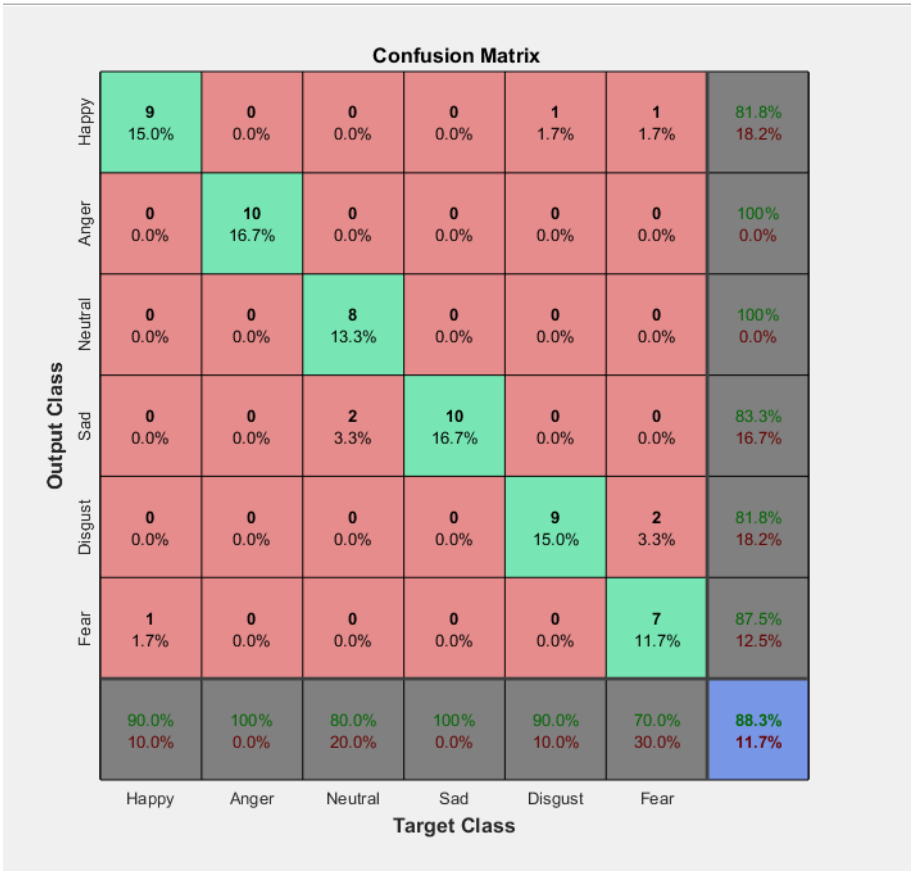


Figure 18: Bangla Accuracy in Bangla-English dataset

Amongst the six emotions anger and sadness were detected accurately and happy, neutral and disgust were detected almost precisely. The emotional state fear was however moderately detected.

The emotion recognition performance for only the English test samples for the Bangla-English dataset was 81.7%. Although the performance for speech emotion recognition was impressive, it lagged behind the performance rate of the Bangla audio samples.

Amongst the six emotions happy, neutral and sadness were detected relatively accurately but the emotional states disgust and fear were detected with an average rate. Out of the 10 samples each for fear and disgust only 7 each were detected correctly.

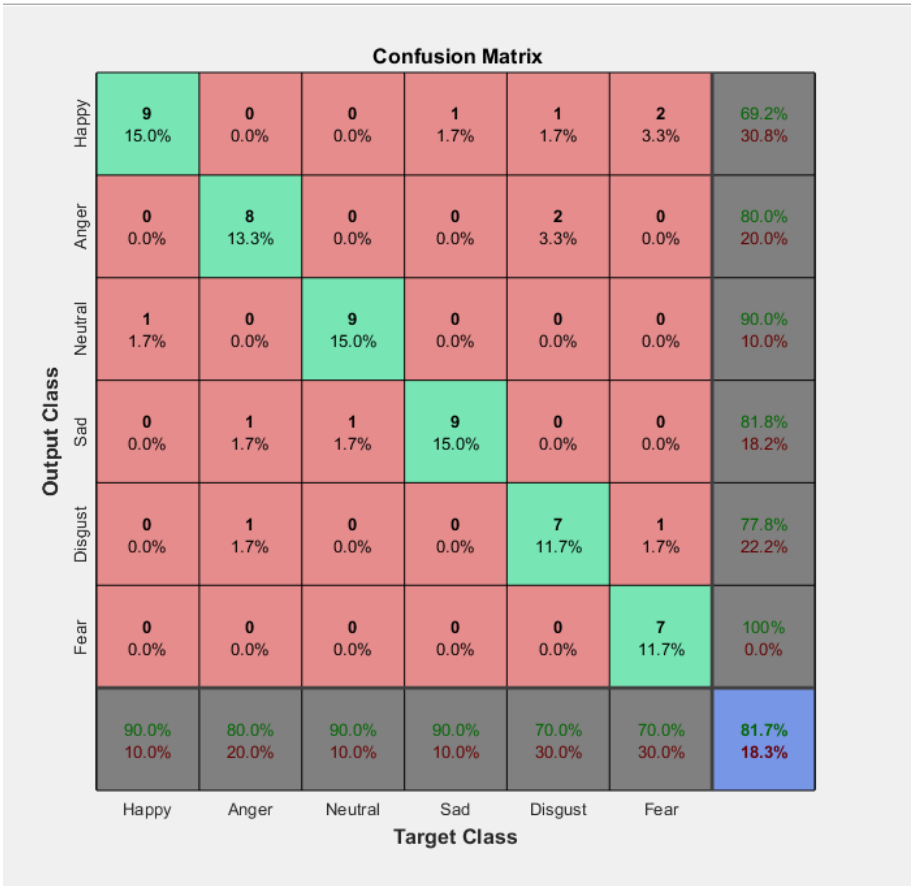


Figure 19: English Accuracy in Bangla-English dataset

6.4.2 Bangla-English_TESS Dataset Confusion Matrix

The emotion recognition rate achieved in the Bangla-English_TESS dataset was 83.3% which is relatively lower than the Bangla-English dataset. Amongst the six emotions anger and neutral was mostly detected appropriately and all the other emotions were detected moderately.

For anger and neutral emotions out of 20 audio test samples for each of them, 18 test samples were precisely detected. Emotions such as happy, sad, disgust and fear were detected averagely.

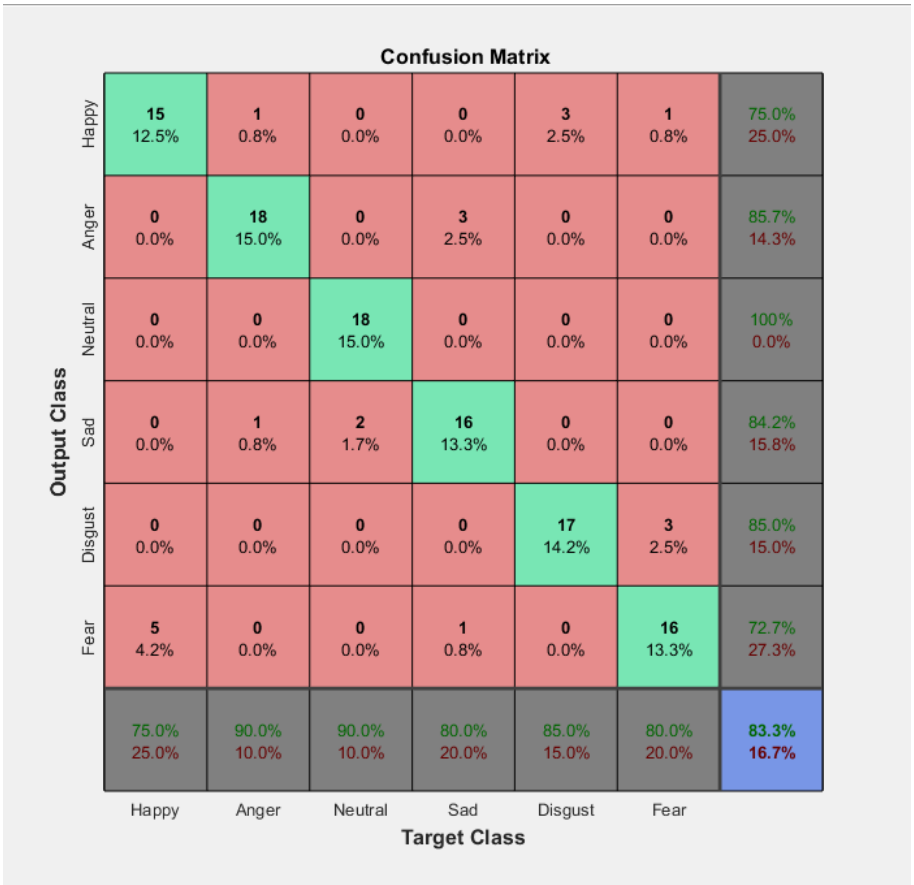


Figure 20: Bangla-English_TESS Testing Confusion Matrix

The emotion recognition rate for Bangla audio samples were significantly lower than the Bangla recognition rate in the Bangla-English dataset. It recorded a low performance score of 75%.

The emotional state anger was detected perfectly in this dataset for the Bangla audio samples. Neutral and sad emotions were detected accurately as well. But the emotions happy,anger and fear were detected very miserly. Out of the 10 audio samples for happy only 5 samples were accurately recognized. Again for fear only 6 out of the 10 audio samples were detected correctly.

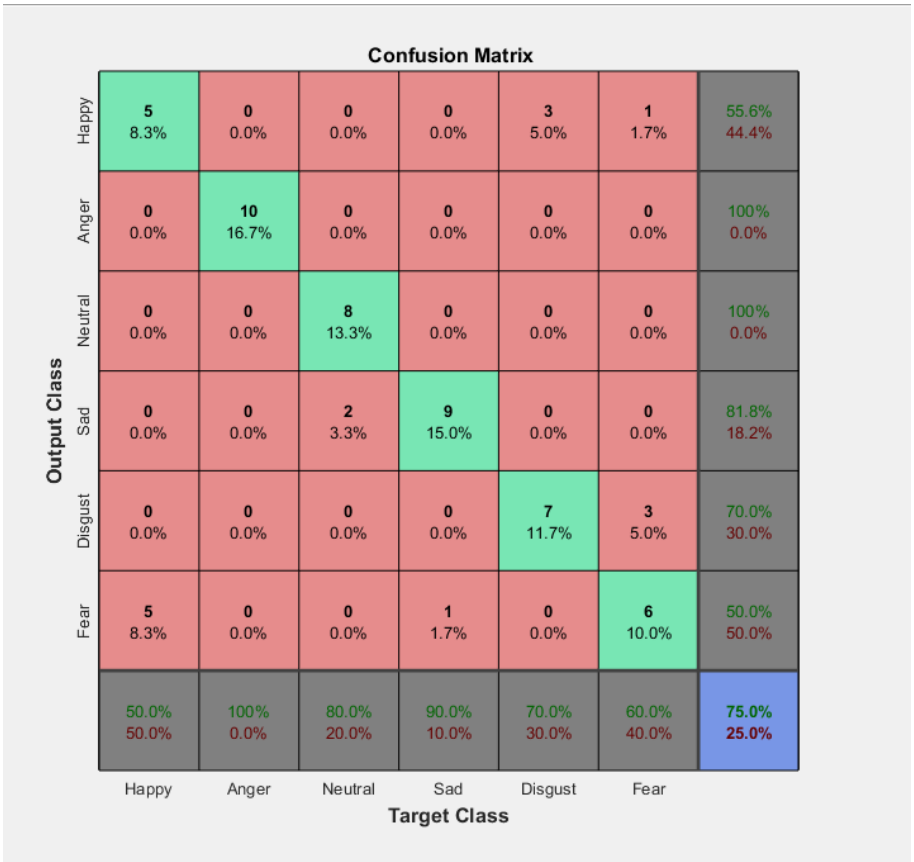


Figure 21: Bangla Accuracy in Bangla-English_TESS dataset

The emotion recognition rate for English_TESS audio samples was phenomenal. The recognition rate for the English_TESS dataset was overwhelmingly higher than the English recognition rate in the Bangla-English dataset. The recognition rate for the English_TESS audio samples were a paramount 91.7%.

The emotions such as happy, neutral, fear and disgust were detected perfectly. Out of the 10 audio samples for the mentioned emotions 10 audio English_TESS test samples were recognized accurately. Therefore the accuracy for English_TESS was so high reaching.

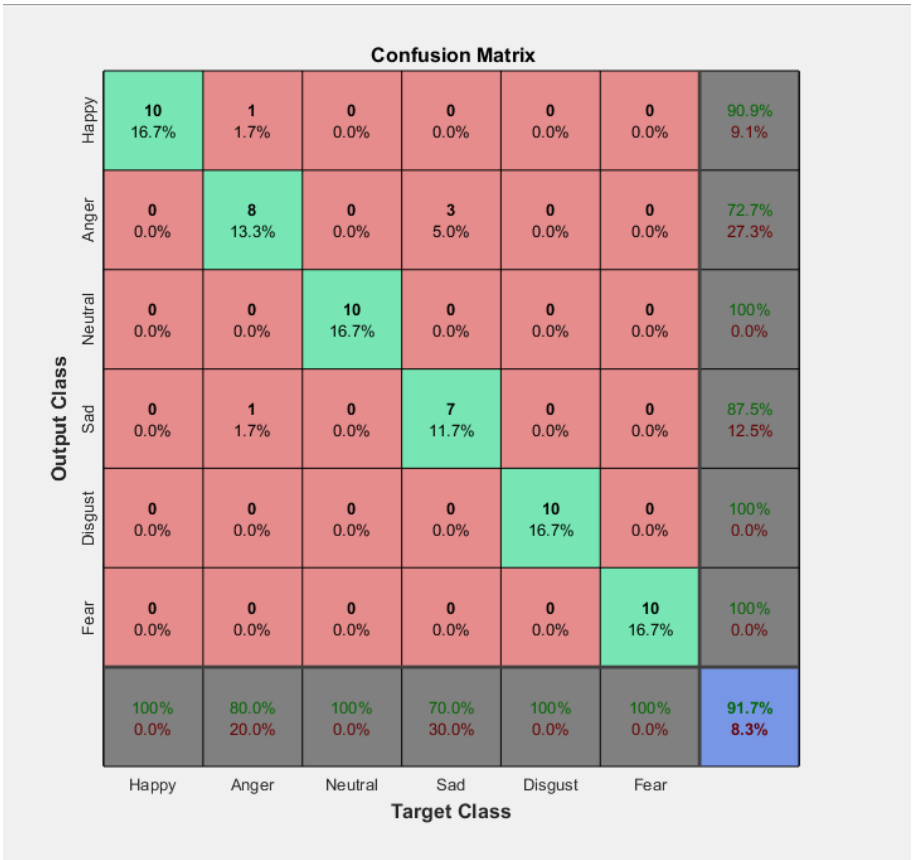


Figure 22: English_TESS Accuracy in Bangla-English_TESS dataset

6.5 Experiment 3: Distinct Speech Corpora for Training & Testing

In this experiment Bangla dataset was used for training and it was tested by English and English_TESS datasets. Similarly English_TESS dataset was used for training and it was tested using Bangla and English datasets. A total of 420 data audio samples were used to construct the datasets.

Each dataset was trained using 300 audio samples. For Dataset-1 300 Bangla audio samples were used for training and for Dataset-2 300 English_TESS audio samples were used for training.

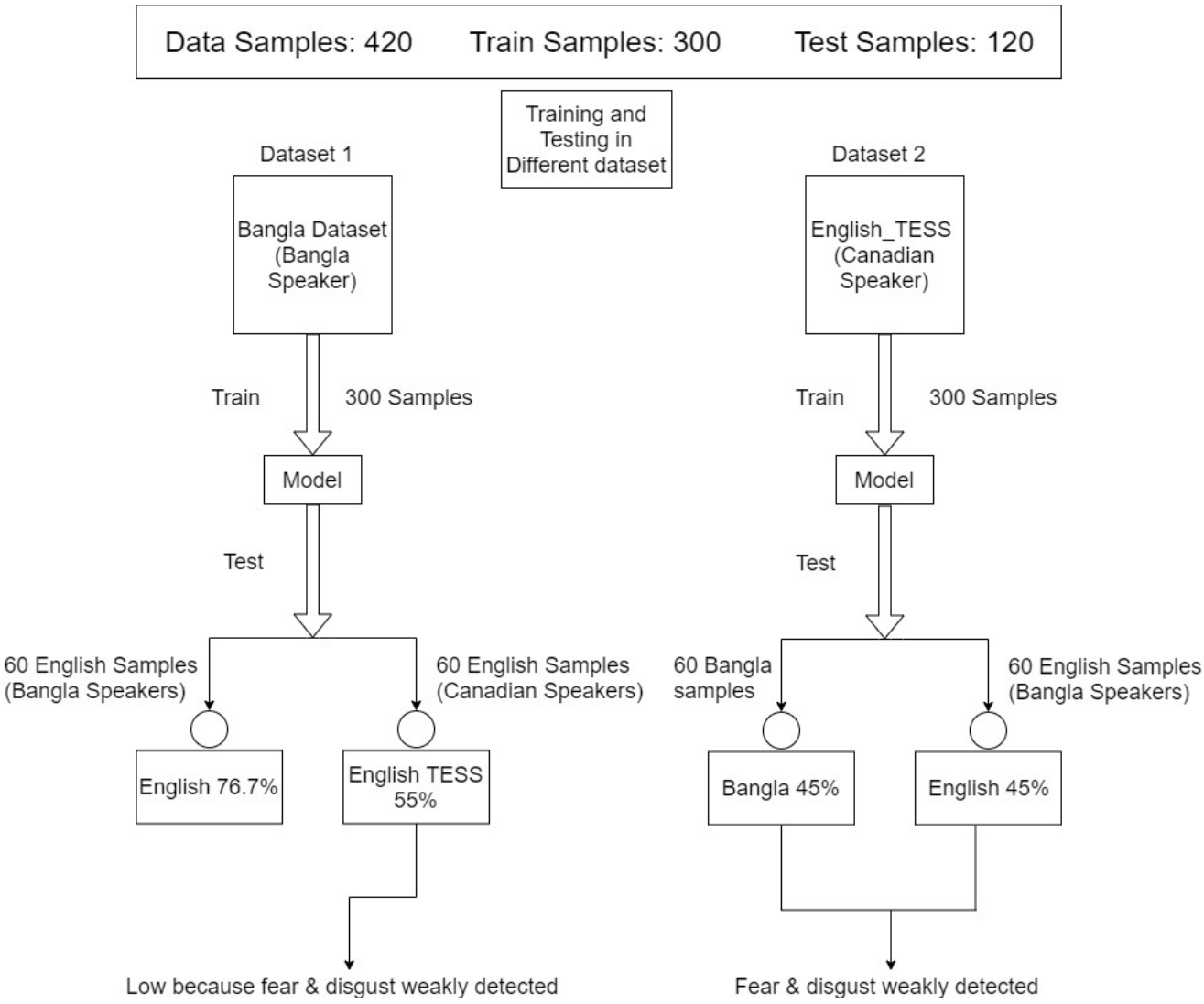


Figure 23: Experiment3

For testing in Dataset-1 60 English audio samples and 60 English_TESS audio samples were used. The English obtained a emotion recognition rate of 76.7% whereas for the English_TESS the emotion recognition rate was 55%. This low score maybe credited to the failure of recognizing the emotions fear and disgust in the dataset.

For testing in Dataset-2 60 English and Bangla audio samples were tested on the dataset trained by English_TESS. Both Bangla and English had a emotion recognition accuracy of 45%. This maybe owing to emotions fear and disgust not being detected accurately.

6.6 Experiment 3: Distinct Speech Corpora for Training & Testing Evaluation

Here we represent the results of testing and training on distinct speech corpora.

6.6.1 Bangla trained & English_TESS, English Tested

The emotions anger, neutral, sad and fear were detected with good precision but the emotions happy and disgust were detected averagely.

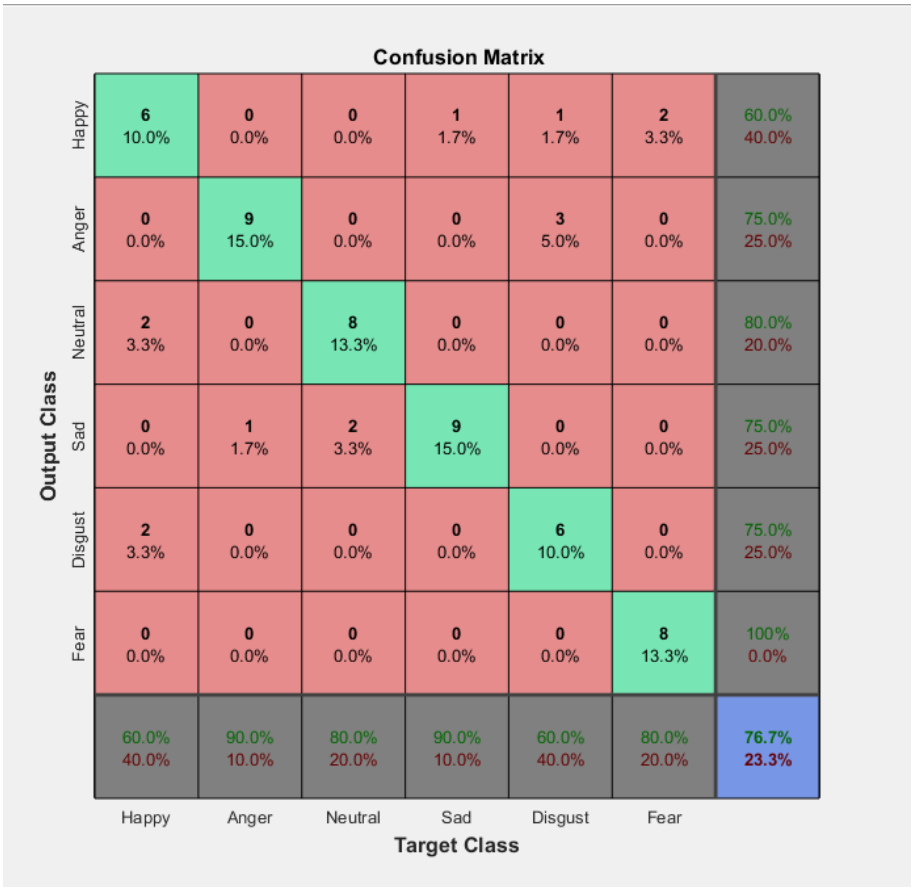


Figure 24: Bangla Trained & English Tested

For happy and disgust emotions out of 10 samples only 6 samples were accurately detected. However this result can be considered quite accurate since other emotions were detected accurately.

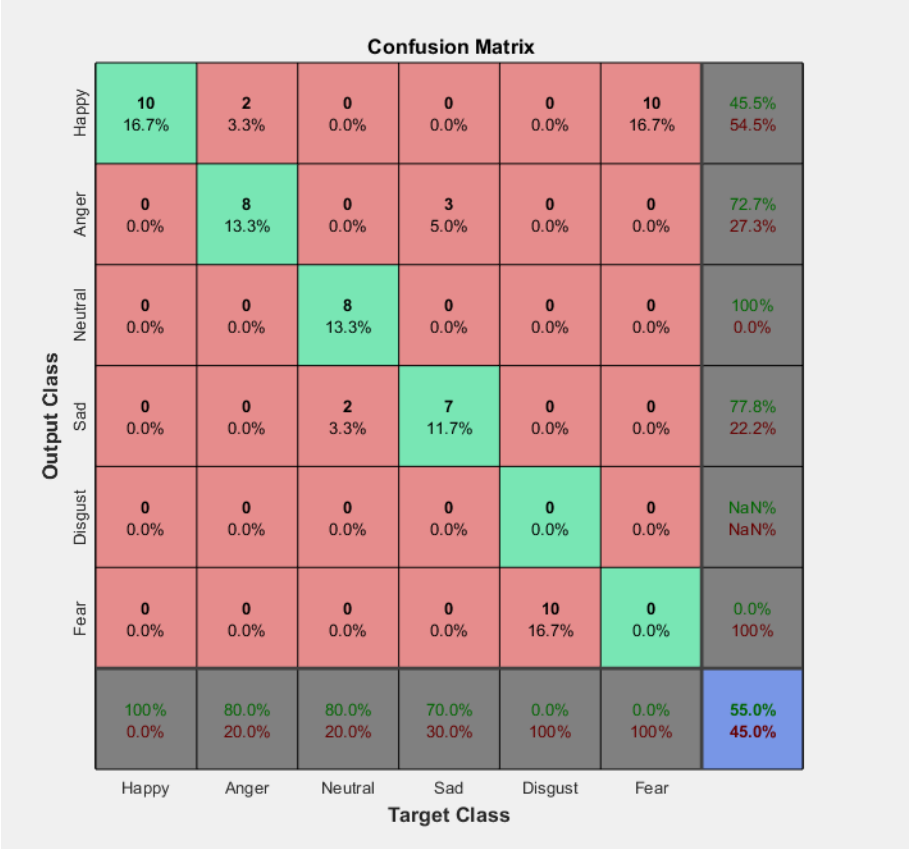


Figure 25: Bangla Trained & English_TESS Tested

When the Bangla dataset was tested by the English_TESS data samples the recognition rate was not satisfactory. The performance rate for emotion recognition stood out to be 55% which can considered to be quite low when comapared against the English dataset. This maybe justified because the emotions fear and disgust were not at all detected.

Out of the 10 audio samples for fear and disgust none was detected by the system. But other emotions such as happy, anger, neutral and sad were detected quite prominently by the model.

6.6.2 English_TESS trained & Bangla, English Tested

When the dataset trained with English_TESS was tested by Bangla the emotion recognition accuracy was 45%. This is a very poor result when compared against other recognition rates. The emotions anger and sad were detected with good

precision but the emotions happy and neutral were detected averagely.

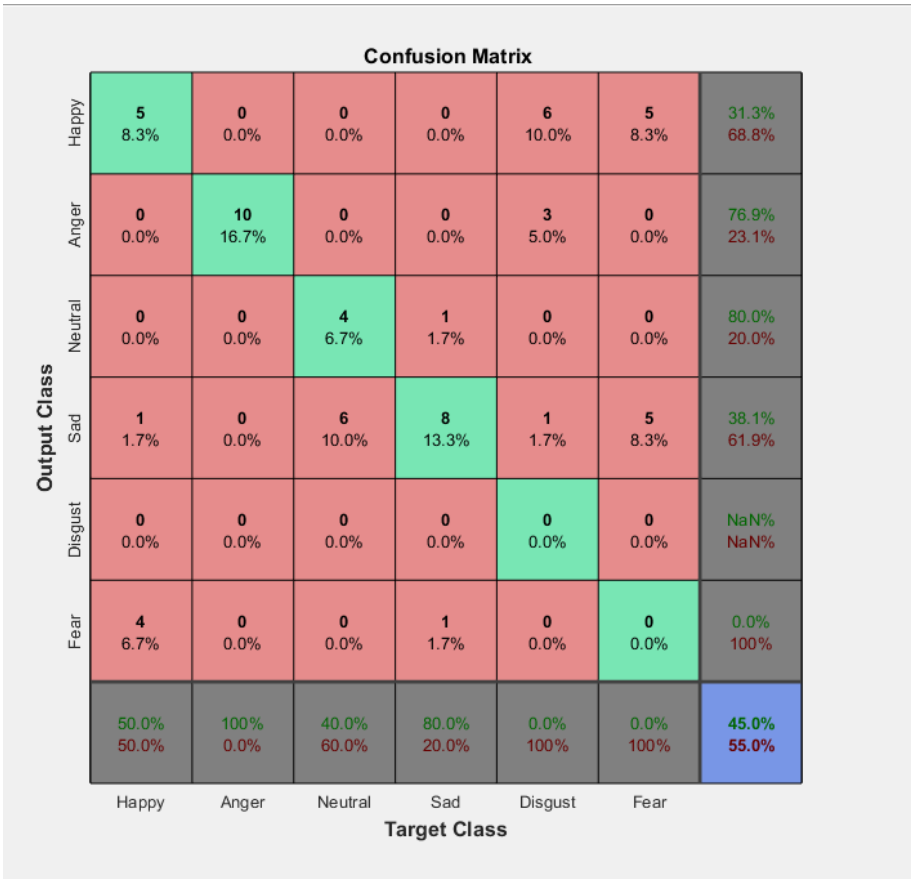


Figure 26: English_TESS trained & Bangla Tested

For happy emotion out of 10 samples only 5 samples were accurately detected. For neutral emotion out of 10 samples only 4 samples were accurately detected. The emotions fear and disgust were not detected at all which may be the reason for which the accuracy was so low.

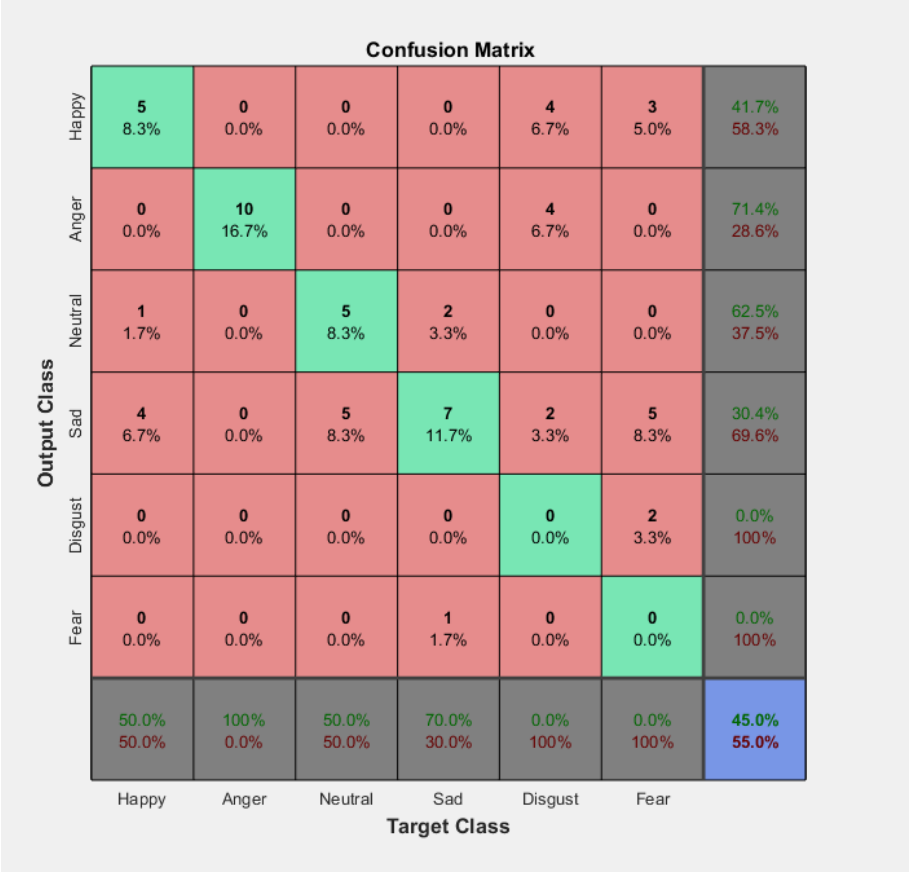


Figure 27: English_TESS trained & English Tested

When the dataset was tested by English the emotion recognition accuracy was 45% which is the same as the Bangla dataset's performance. The emotions anger and sad were detected with good precision but the emotions happy and neutral were detected averagely.

The reason of this poor recognition rate can again be credited to the failure of detecting emotions fear and disgust. This time as well the emotions fear and disgust were not detected at all and had 0% accuracy. Therefore for this experiment whatever irrespective of the language(Bangla & English) used the emotions fear and disgust were not detected at all.

7 Result Analysis

A total of 3 experiments were conducted to investigate whether a SER system can identify the emotional state of a person regardless of the language(Bangla & English).

- In the first experiment it can be observed that individually the datasets have an impressive emotion recognition rate. The recognition rates for the datasets Bangla, English and English_TESS were 88.3%, 85% and 93.3% respectively. Each of these datasets could predict the emotional states anger, happy, sad, neutral, fear and disgust accurately. The emotion recognition of English dataset maybe low since it was developed by Bangladeshi speakers whose second language is English.

- In the second experiment the accuracy of Bangla and English(developed by Bangladeshi speakers) on the integrated Bangla-English dataset were 88.3% and 81.7% respectively. The recognition rate of English is lower than Bangla since native Bangla speakers developed this dataset. Thus they may have not express their feelings well in English. Again, for the recognition rates of Bangla and English_TESS on the dataset Bangla-English_TESS were 75% and 91.7%. This dataset was developed by the native speakers of the language. However, the English_TESS recognition rate was higher than Bangla which suggests that there maybe language specific differences in emotion recognition.

- In the third experiment when English_TESS was tested on a model trained by the Bangla dataset and vice versa the emotion recognition rate was significantly low. This decrease in accuracy may be credited to the failure of recognizing the emotions fear and disgust. The prosodic cues generated during fear and disgust may be different in Bangla and English. There is a possibility of being influenced by culture and background. Thus it may be deduced that there maybe language specific differences in emotion recognition.

Training \ Testing	Bangla	English	English_TESS	Bangla & English	Bangla & English_TESS
Bangla	88.30%	76.70%	55%	N/A	N/A
English	N/A	85%	N/A	N/A	N/A
English_TESS	45%	45%	93.30%	N/A	N/A
Bangla & English	88.30%	81.70%	N/A	85%	N/A
Bangla & English_TESS	75%	N/A	91.30%	N/A	83.30%

Figure 28: Accuracy across different experiments

8 Conclusion and Future Works

Our analysis focused on two languages which are English and Bangla. The data analysis from this study demonstrated three experiments to detect discrete emotions. Although recognition of emotions was reliable, the accuracy rate varied in our three speech corpora of interest when the recognition rates of emotions where scrutinized under assorted circumstances.

This shows that there are language specific differences in emotion recognition in which the emotions fear and disgust appeared to give contradictory results. This study also demonstrated that emotions expressed by native speakers have higher accuracy rates. Thus there are language specific differences in emotion recognition and emotions expressed by native speakers have a higher accuracy rate.

Hence, we hope that the findings of this study will help to lift up the global accuracy of Speech Emotion Recognition. Our future research focus is to enhance these findings by including more languages and more emotional status. Another direction of the future research will be to determine whether there are any speech features that contribute exclusively to emotion recognition of a selected language.

References

[1] Rajoo, Rajesvary, and Ching Chee Aun. "Influences of languages in speech emotion recognition: A comparative study using Malay, English and Mandarin languages." 2016 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE). IEEE, 2016.

[2] Sudhakar, Rode Snehal, and Manjare Chandraprabha Anil. "Analysis of speech features for emotion detection: a review." 2015 International Conference on Computing Communication Control and Automation. IEEE, 2015.

[3] Noroozi, Fatemeh, Marina Marjanovic, Angelina Njegus, Sergio Escalera, and Gholamreza Anbarjafari. "A Study of Language and Classifier-independent Feature Analysis for Vocal Emotion Recognition." arXiv preprint arXiv:1811.08935 (2018).

[4] Kandali, Aditya Bihar, Aurobinda Routray, and Tapan Kumar Basu. "Emotion recognition from Assamese speeches using MFCC features and GMM classifier." TENCON 2008-2008 IEEE region 10 conference. IEEE, 2008.

[5] M. ElAyadi, M.S. Kamel and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases", Pattern Recognition, vol. 44, no.3, pp. 572–587, March 2011..

[6] R.W. Picard, "Affective computing. Technical Report 321", MIT Media Laboratory Perceptual Computing Section, Cambridge, MA,USA, November 1995.

[7] M.D. Pell, S. Paulmann, C. Dara, A. Alasseri, S. A. Kotz, "Factors in the recognition of vocally expressed emotions: A comparison of four languages". Journal of Phonetics, vol. 37, no. 4, pp. 417-435, October 2009.

[8] M. Gjoreski, H. Gjoreski, A. Kulakov, "Machine Learning Approach for Emotion Recognition in Speech", Informatica, vol. 38, pp. 377–384, December 2014.

[9] Schuller, Björn, Gerhard Rigoll, and Manfred Lang. "Hidden Markov model-based speech emotion recognition." 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).. Vol. 2. IEEE, 2003.

[10] Dupuis, Kate, and M. Kathleen Pichora-Fuller. "Recognition of emotional speech for younger and older talkers: Behavioural findings from the Toronto Emotional Speech Set." Canadian Acoustics 39.3 (2011): 182-183.

[11] W. Hess, Pitch determination of speech signals: algorithms and devices. Springer Science & Business Media, 2012, vol. 3.

[12] S.-H. Lee, T.-Y. Hsiao, and G.-S. Lee, "Audio–vocal responses of vocal fundamental frequency and formant during sustained vowel vocalizations in different noises," Hearing research, vol. 324, pp. 1– 6, 2015.

[13] E. Globerson, N. Amir, O. Golan, L. Kishon-Rabin, and M. Lavidor, "Psychoacoustic abilities as predictors of vocal emotion recognition," Attention, Perception, & Psychophysics, vol. 75, no. 8, pp. 1799–1810, 2013.

[14] J. Harrington, Phonetic analysis of speech corpora. John Wiley & Sons, 2010.

[15] G. Chronaki, J. A. Hadwin, M. Garner, P. Maurage, and E. J. Sonuga-Barke, "The development of emotion recognition from facial expressions and non-linguistic vocalizations during childhood," British Journal of Developmental Psychology, vol. 33, no. 2, pp. 218–236, 2015.

[16] R. Allgood and P. Heaton, "Developmental change and crossdomain links in vocal and musical emotion recognition performance in childhood," British Journal of Developmental Psychology, vol. 33, no. 3, pp. 398–403, 2015.

[17] Y. Pan, P. Shen, and L. Shen, "Speech emotion recognition using support vector machine," International Journal of Smart Home, vol. 6, no. 2, pp. 101–108, 2012.

[18] P. Laukka, D. Neiberg, and H. A. Elfenbein, "Evidence for cultural dialects in vocal emotion expression: Acoustic classification within and across five nations." Emotion, vol. 14, no. 3, p. 445, 2014.

[19] Amer, Mohamed R., et al. "Emotion detection in speech using deep networks." 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2014.

[20] M. R. Amer, B. Siddiquie, S. Khan, A. Divakaran, and H. Sawhney, "Multimodal fusion using dynamic hybrid models," in WACV, 2014.

[21] Yelin Kim, Honglak Lee, and Emily Mower Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in ICASSP, 2013.

[22] Bhatti, Muhammad Waqas, Yongjin Wang, and Ling Guan. "A neural network approach for human emotion recognition in speech." 2004 IEEE International Symposium on Circuits and Systems (IEEE Cat. No. 04CH37512). Vol. 2. IEEE, 2004.

[23] Shaukat, Arslan, and Ke Chen. "Exploring language-independent emotional acoustic features via feature selection." arXiv preprint arXiv:1009.0117 (2010).

[24] S.G. Koolagudi, S. Devliyal, B. Chawla, A. Barthwal, "Recognition of Emotions from Speech using Excitation Source Features". Procedia Engineering, vol. 38, pp. 3409 – 3417, 2012.

[25] M. Gjoreski, H. Gjoreski, A. Kulakov, "Machine Learning Approach for Emotion Recognition in Speech", Informatica, vol. 38, pp. 377–384, December 2014.

[26] P. Shen, Z. Changjun, X. Chen, "Automatic Speech Emotion Recognition Using Support Vector Machine", International Conference on Electronic & Mechanical Engineering and Information Technology, vol. 2, pp. 621 – 625, August 2011.